Reporte Final: Hacia un sistema de información geográfica de fincas bananeras en Urabá para monitoreo, control y optimización de la producción

Julián Ramírez, Andy Jarvis, Jairo Guerrero

Centro Internacional de Agricultura Tropical, CIAT, Cali, Colombia
E-mail: j.r.villegas@cgiar.org

Parte 1. Zonificación agroecológica sobre las fincas bananeras del Urabá

Resumen

El presente documento describe los materiales usados, métodos aplicados y resultados obtenidos por el Centro Internacional de Agricultura Tropical (CIAT) respecto a la zonificación agroecológica de las fincas del Urabá bajo producción bananera afiliadas a la compañía CI UNIBAN. El análisis consistió de 4 pasos básicos: (1) colección de datos de entrada, (2) uniformización y pre-procesamiento de datos de entrada, (3) selección de variables de importancia, (4) agrupamiento de zonas homólogas, (5) conclusiones. Se emplearon datos de análisis de suelos (296 perfiles de suelos), promedios de largo plazo de variables climáticas importantes (a través de la base de datos WorldClim), datos satelitales de índice de vegetación (NDVI), y el modelo de elevación digital de alta resolución de la zona. Se aplicaron funciones *Thin Plate Spline* para uniformizar la información espacial a la resolución del DEM (~90 metros), y usando un análisis de componentes principales (ACP) se determinaron las variables más importantes. Con las variables más importantes según el ACP, se realizó un análisis de conglomerados (*clustering*) de Ward sobre toda la región. Se agruparon los puntos analizados en 20 zonas homólogas, de acuerdo al criterio de Ward de maximización de la varianza y se establecieron estas como "zonas agroecológicas" estables, se describen dichas zonas en términos de todas las variables y se analizan algunos aspectos relevantes respecto a ciertas tendencias en los datos y el uso y actualización de los resultados.

Contenido

---

1. Introducción

Dada la cantidad de información que día tras día se colecta en campo, dadas las exigencias por parte del mercado, y asimismo la presión tanto biótica (enfermedades, plagas) como abiótica (factores ambientales), se hace necesario analizar las variables de producción y encontrar maneras de monitorear los efectos que tiene el ambiente en la respuesta de los diferentes genotipos cultivados, y en adición a esto, generar un despliegue de información que sea accesible para el productor y que permita a UNIBAN y al productor controlar, decidir y aplicar los correctivos necesarios, así como también prever futuras situaciones o riesgos de pérdida de producción y/o productividad.

Como primer paso en la consecución de lo anterior, es importante conocer la región de estudio, y aún más importante que eso, lograr optimizar los recursos y para esto, es fundamental realizar análisis de similitud ecológica al interior de la región, que permita agrupar áreas productoras de acuerdo a ciertos patrones edáficos, climáticos y fenológicos observados a través de la región. En el presente informe se presentan los resultados de la zonificación agroecológica, en la que se usaron variables de suelos, topográficas, de clima y de crecimiento y fenología del cultivo. Usando análisis estadísticos (análisis de componentes principales, ACP) se seleccionaron las variables de mayor importancia y se aplicó el método de análisis de conglomerados de Ward (1963) para agrupar zonas homólogas de acuerdo al criterio de máxima varianza entre grupos. Se describen los resultados más importantes y tendencias, y se presentan futuras aplicaciones de los datos.

2. Colección de datos de entrada

a. Datos de análisis de suelos

Se recibió información correspondiente a los estudios de suelos realizados sobre todas las fincas bananeras bajo estudio. Un total de 296 perfiles fueron objeto de estudio para la zonificación agroecológica (Figura 1).

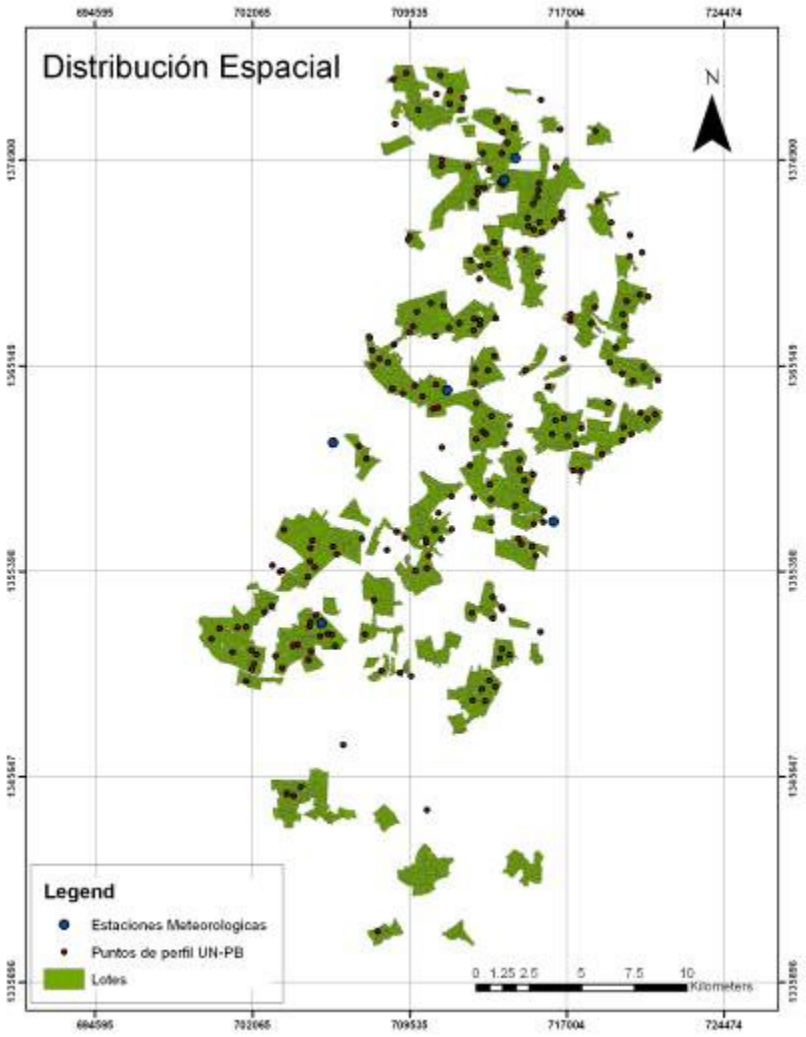

Figura 1 Distribución espacial de perfiles de suelo (puntos rojos) y estaciones meteorológicas (puntos azules) a través de las fincas bananeras (polígonos verdes)

De cada uno de estos perfiles, se contó con información de variables químicas y físicas de suelos (Tabla 1). Los datos se encuentran organizados en el archivo "perfilesSuelos.xls", que se encuentra en el folder ".\input-data\corregido\". Muchas de estas variables fueron medidas a diferentes profundidades (diferentes horizontes) y algunas de ellas presentan

una alta variabilidad entre un perfil y el otro, en adición a esto, también se provee la taxonomía del suelo de cada punto de muestreo.

Tabla 1 Estadística descriptiva de propiedades de suelos de la primera capa, sobre el total de perfiles de los cuales se tuvieron datos

Variable	Promedio	Desviación estándar	CV (%)	Máximo	Mínimo
Arena (%)	17.38	13.40	77.10	65.40	0.10
Limo (%)	44.29	10.87	24.53	72.90	14.00
Arcilla (%)	38.34	13.64	35.59	74.70	5.90
Densidad real (g/cm3)	2.46	0.27	11.13	2.80	0.00
Densidad aparente (g/cm3)	1.12	0.20	17.96	1.60	0.00
Humedad gravimetrica saturacion (%)	53.13	11.70	22.03	104.51	29.57
Humedad gravimetrica (0.3 atm)	41.52	9.77	23.54	87.51	17.68
Humedad gravimetrica (15 atm)	25.56	7.06	27.63	48.74	2.40
Humedad gravimetrica aprovechable (%)	16.04	5.38	33.55	44.62	5.97
Humedad volumetrica saturacion (%)	58.07	12.37	21.30	104.51	0.00
Humedad volumetrica (0.3 atm)	45.30	9.49	20.94	74.15	0.00
Humedad volumetrica (15 atm)	27.94	7.53	26.95	46.39	0.00
Humedad volumetrica aprovechable (%)	17.45	5.24	30.03	35.09	0.00
Estado de agregacion	91.02	12.03	13.22	99.97	37.16
Diametro medio ponderado (mm)	4.14	1.65	39.90	6.66	0.04
Infiltracion basica (cm/h)	5.01	6.65	132.73	30.30	0.00
Cond. Hidr. Pozo invertido (cm/h)	12.91	17.79	137.83	119.69	0.07
Cond. hidr. pozo barrenado (cm/h)	4.71	12.39	263.31	111.51	0.00
pH	5.62	0.79	14.10	7.80	3.60
Nivel freatico (cm)	153.97	39.13	25.41	260.00	70.00

Hay un determinado rango de variabilidad entre los diferentes perfiles, lo que sugiere que existen gradientes a través de la región, y que probablemente hay un nivel de similitud al cual ciertas regiones comparten características, lo que hace posible una clasificación agroecológica con miras a la optimización de los recursos y de la producción. Algunas de las variables presentan una variabilidad relativamente baja (e.g. densidad real, pH, densidad aparente), mientras otras presentan una muy alta variabilidad a través de los perfiles (e.g. infiltración básica, conductividad hidráulica), aunque la gran mayoría se mantienen en un rango moderado de variabilidad.

b. Promedios climáticos de largo plazo

Para la obtención de promedios de variables climáticas de largo plazo se pensó inicialmente en usar datos de estaciones meteorológicas, como mediciones de campo de alta precisión que brindaran superficies adecuadas para el análisis de zonificación. Sin embargo, los datos de las 7 estaciones meteorológicas entregados por UNIBAN CI tenían algunos problemas, entre ellos:

- Varios meses sin datos en la información recibida.
- Datos de precipitación diaria variando entre -5000 y 4500 mm/día, lo que es físicamente imposible y deberá ser verificado. Estos datos ocurren durante períodos seguidos en algunos anos.
- Datos de evapotranspiración diaria (mm/dia) que llegan a ser negativos. Esto ocurre en varios meses, y de manera relativamente frecuente (32% de los datos)

Entre otros problemas que pueden observarse dando una mirada general a la base de datos. Se recomienda que CI UNIBAN realice un cuidadoso análisis de estos datos, y verifique su validez, y en caso de ser necesario, corrija y revise las estaciones, re-calibre, y garantice que la información es confiable.

Como solución a lo anterior, se usó la base de datos de WorldClim (Hijmans et al. 2005, disponible en http://www.worldclim.org), una base de datos de precipitación total, temperatura máxima, mínima y media mensuales, a 30 arco-segundos de resolución espacial (aproximadamente 1km en el Ecuador), y con una cobertura global. La base de datos WorldClim fue desarrollada a partir de datos de estaciones meteorológicas de diferentes fuentes, y es hasta la fecha la base de datos de mayor resolución existente para análisis climático. Los datos en WorldClim representan promedios de largo plazo (período comprendido entre 1950 y 2000) y por tanto son útiles para representar el clima de grandes y pequeñas regiones en diferentes áreas geográficas. De la base de datos de WorldClim se extrajeron 19 variables climáticas que representan tendencias anuales (promedio histórico) (Tabla 2).

Tabla 2 Estadística descriptiva de variables climáticas colectadas sobre la zona analizada

Variable	Promedio	Desviación estándar	CV (%)	Máximo	Mínimo
Elevación (m)	17.1	9.1	53.1	56.0	-2.0
P1. Temp. media anual (°C)	26.8	0.1	0.4	27.0	26.6
P2. Rango medio diurno de Temp. (°C)	8.3	0.2	1.9	8.7	7.9
P3. Isotermalidad (P2/P7)	87.1	1.6	1.8	90.2	83.9
P4. Estacionalidad de temperatura (desv. est.) (°C)	2.5	0.3	10.6	3.1	2.0
P5. Temp. máxima del mes mas caliente (°C)	31.7	0.3	0.9	32.4	31.1
P6. Temp. mínima del mes mas frío (°C)	22.2	0.1	0.3	22.4	22.0
P7. Rango anual de temperatura (°C)	9.5	0.3	3.3	10.3	8.8
P8. Temp. media del trimestre mas humedo (°C)	26.7	0.1	0.3	26.9	26.4
P9. Temp. media del trimestre mas seco (°C)	26.9	0.1	0.4	27.2	26.7
P10. Temp. media del trimestre mas caliente (°C)	27.1	0.2	0.6	27.4	26.8
P11. Temp. media del trimestre mas frio (°C)	26.5	0.1	0.3	26.7	26.3
P12. Precipitación total anual (mm)	2694.7	188.2	7.0	3414.5	2485.4
P13. Precipitación del mes mas húmedo (mm)	315.2	28.8	9.1	431.5	281.9
P14. Precipitación del mes mas seco (mm)	77.3	8.3	10.7	91.1	65.9
P15. Estacionalidad de precipitación (CV) (%)	36.5	3.0	8.3	46.2	32.9
P16. Precipitación del trimestre mas húmedo (mm)	872.7	90.8	10.4	1219.6	768.2
P17. Precipitación del mes mas seco (mm)	262.4	16.7	6.4	280.1	217.1
P18. Precipitación del trimestre mas caliente (mm)	435.8	9.8	2.3	453.1	411.7

P19. Precipitación del trimestre mas frío (mm)	797.6	38.4	4.8	940.6	742.4

De igual manera que para los suelos, se observan diferencias a través de la región. En algunos casos (i.e. elevación), estas diferencias son significativas, en tanto que en otros casos (i.e. temperaturas medias, mínimas y máximas), estas diferencias son casi insignificantes, de tan solo un décimo de grado centígrado. Las variables de precipitación, en general, presentan una moderada varianza a través de la región. De nuevo, estos resultados, junto con los resultados de suelos, sugieren que hay variaciones significativas que permitirían un agrupamiento de los datos a un nivel de similitud determinado. Todos estos datos se entregan en la carpeta ".\zonificacion\variables-originales\clima".

c. Datos de índice de vegetación (NDVI)

Los datos de NDVI se extrajeron a partir de imágenes del satélite TERRA MODIS, producto MOD13Q1. La resolución espacial de estos datos es de aproximadamente 7.5 arco-segundos (aproximadamente 250 metros en el Ecuador), y la frecuencia de los datos es de 16 días, esto significa que cada 16 días hay un nuevo dato de NDVI resultado de la medición del satélite, y que es publicado por la NASA, descargado y post-procesado por el CIAT. El período para el cual se extrajeron los datos de NDVI fue el período 1999-2000.

Dado que el objetivo de tener datos de NDVI para un análisis de zonificación es lograr capturar patrones temporales (de mediano y largo plazo) en los datos espaciales, no se usaron los datos cada 16 días, sino que se calcularon variables que representan tendencias centrales y de dispersión en los datos (Tabla 3). Con estas se realizaron todos los análisis posteriores.

Tabla 3 Estadística descriptiva de las variables de desarrollo fenológico (NDVI) usadas en el análisis, sobre la región de estudio de Urabá

Variable	Promedio	Desviación estándar	CV (%)	Máximo	Mínimo
NDVI desviación estándar período 2000-2009	0.0141	0.0034	24.0	0.0422	0.0065
NDVI promedio histórico período 2000-2009	0.8240	0.0386	4.7	0.8935	0.4343
NDVI máximo histórico período 2000-2009	0.8535	0.0370	4.3	0.9159	0.4765
NDVI mínimo histórico período 2000-2009	0.7917	0.0418	5.3	0.8750	0.3978
NDVI rango medio anual período 2000-2009	0.0272	0.0064	23.5	0.0979	0.0112
NDVI rango total histórico período 2000-2009	0.0617	0.0156	25.3	0.1809	0.0272

Igualmente se observa que existe un nivel de dis-similaridad espacial a través de la región, que se presenta de manera mucho más significativa en las variables que representan variabilidad temporal (dispersión) que en las que representan tendencia central o extremos (promedio, máximo y mínimo), no obstante, en todos los casos, el rango de variación es de al menos 50% sobre el valor máximo, lo que sugiere,

nuevamente, diferencias y similitudes entre áreas de la región bananera bajo estudio. Estos datos se entregan en la carpeta ".\zonificacion\variables-originales\ndvi"

d. Modelo de elevación digital (DEM)

Los datos de elevación usados para el presente estudio se obtuvieron a través del Centro Internacional de Agricultura Tropical (CIAT), y su base de datos SRTM. La base de datos SRTM (Jarvis et al. 2008) fue construida a partir de mediciones del satélite SRTM de la NASA, y corregida a través de interpolación espacial, usando diferentes algoritmos. La base de datos contiene valores de elevación cada 3 arco-segundos (aproximadamente 90 metros en el Ecuador) y es hasta la fecha la base de datos de elevación más precisa y comprehensiva a nivel mundial (Figura 2).

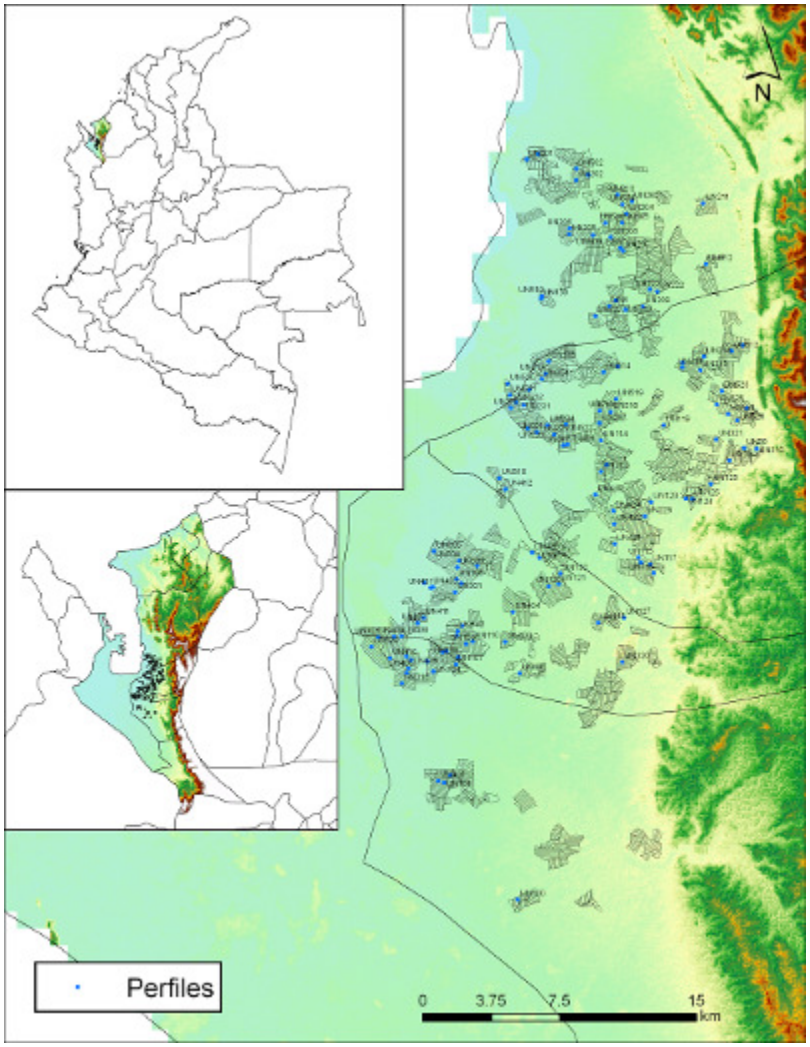

Figura 2 Variaciones altitudinales a través de la región bananera de Urabá, fincas UNIBAN CI, datos de la base de datos SRTM

En la Tabla 2, junto con las variables climáticas, se puede observar la variación en la elevación, cuyo rango va desde -2 hasta 56 metros sobre el nivel del mar, y presenta un coeficiente de variación por encima del 50%, lo que indica una alta variabilidad, y por tanto una alta potencialidad de desagregar la región en pequeñas zonas para manejo (zonas agroecológicas).

3. Uniformización y pre-procesamiento de datos de entrada

Debido a las diferencias que existen entre las resoluciones espaciales e incluso de los tipos de datos climáticos, de elevación, de NDVI y de suelos, es necesario uniformizar los datos tanto en formato como en resolución, para lograr un óptimo resultado con cualquier procedimiento estadístico que se realice. Se decidió que la resolución espacial del estudio sería 3 arco-segundos (~90m), la cobertura sería de toda el área donde hay fincas, y que el formato de trabajo para todos los datos sería el raster (grillas). Para lograr tener las superficies climáticas y de NDVI en resolución menor, se usó el algoritmo TOPOGRID de ArcGIS, mientras que para generar superficies de las variables de suelos se usó el método de interpolación espacial *Thin Plate Spline* (TPS) (Hutchinson, 1984; Hutchinson & de Hoog, 1985). Finalmente, se definió una máscara de análisis, para evitar sub-estimación y sobre-estimación de los datos.

a. Aplicación de función *Thin Plate Spline* a datos de suelos

Usando los valores correspondientes a cada uno los puntos de análisis de suelos, y con el fin de obtener una superficie que describiese el gradiente espacial de cada una de las variables anteriores, se usó el método de interpolación *Thin Plate Spline* (TPS) (Hutchinson, 1984; Hutchinson & de Hoog, 1985) que ha sido empleado por diversos autores para ajustar la distribución de variables con muy alta varianza, o cuyas mediciones de campo tienen ruido. El método, como se aplicó a los datos de los estudios de suelos de UNIBAN y PROBAN, usa cada una de las variables de suelos como variables dependientes, y ajusta su distribución a una función multivariada usando la altitud (en metros, derivada de SRTM, Jarvis et al. 2008), la latitud (unidades dependiendo del sistema de coordenadas) y la longitud (unidades dependiendo del sistema de coordenadas). Una vez se encuentra la función de ajuste, dicha función se proyecta sobre toda el área bajo análisis, incluyendo aquellas áreas en las que no se han realizado muestreos de suelos.

Para cada una de las variables de suelos bajo análisis, se ajustó una función TPS, y se calculó una superficie de distribución de la variable a través de la geografía de la región. Este proceso se realizó usando el software R (http://www.r-project.org), con las librerías *fields*, *spam*, *rgdal*, *sp* y *raster*. Tanto R, como las librerías en cuestión pueden descargarse de la página http://www.r-project.org. Para más información sobre

instalación, notas y librerías de R, referirse al manual de este software. Se provee un script genérico a través del cual todos estos ajustes fueron realizados ("./_scripts/TPS-Uniban.R"). Este script debe cargarse en R usando el comando *source*, así:

```
source("TPS-Uniban.R")
```

Esto creará una función llamada "crearSuperficie", a través de la que se realiza el proceso, se ejecuta así:

```
salida <- crearSuperficie(inputData, dem, outputSurf, plotSrf=F)
```

Donde:

**__inputData__ es un archivo csv con n filas y 3 columnas (valores,x,y) separado por comas.
**__dem__ es el modelo de elevación digital del terreno como ASCII GRID
**__outputSurf__ es el nombre del raster de salida interpolado
**__plotSrf__ es un falor logico (TRUE/FALSE) para graficar/no graficar resultado (raster interpolado)

Este script de R no sólo ajustará los datos a una función TPS, sino que producirá dos gráficos adicionales en formato JPEG. El primero, llamado "01-SQDiff-Histograma.jpg" muestra el histograma de errores relativos (residuales) de interpolación. Estos residuales son calculados como:

$$Error = \frac{(V_{REAL} - V_{PREDICHO})^2}{(V_{REAL-MAXIMO})^2} * 100$$

El segundo archivo, llamado "02-XYChart-MeasvsPred.jpg", contiene el gráfico de dispersión (X-Y) entre el valor predicho y el valor real, y reporta también el coeficiente de correlación de Pearson (R). De la misma manera, el script genera una tabla de datos de evaluación (archivo .csv) de salida con los valores de entrada (coordenadas y valor real de la variable), así como la elevación (extraída del DEM), el valor predicho de la variable (ajustado por TPS), la diferencia de cuadrados, y la diferencia de cuadrados relativa al máximo. Con este archivo se pueden realizar cualquiera de los gráficos de evaluación que se necesiten.

La interpolación de los datos de suelos permitió determinar de una manera aproximada, los patrones de distribución espacial de las variables más importantes, y que probablemente influyen la producción de las diferentes fincas. Además de esto, permitió la obtención de superficies continuas a través de las diferentes fincas que a su vez permitieron la aplicación de un método estadístico concreto para realizar la zonificación (Figura 3, ejemplo con pH del suelo).

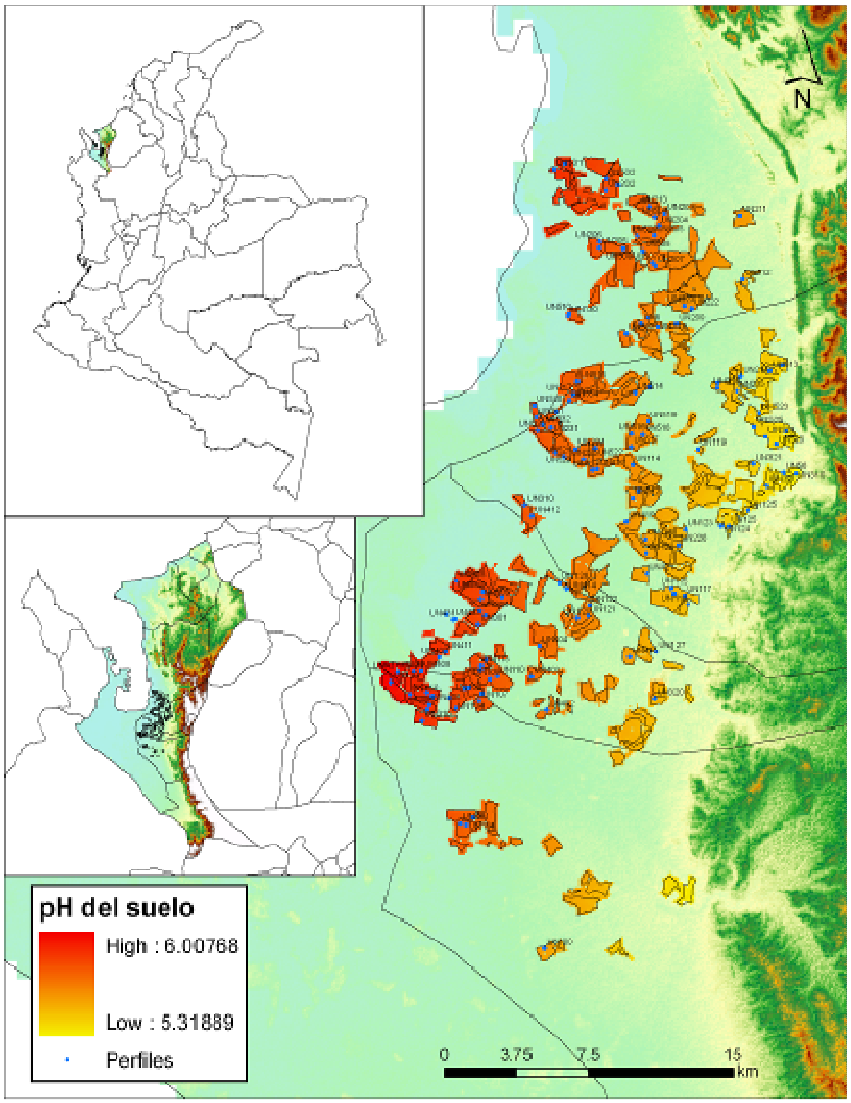

Figura 3 Distribución espacial del pH en el suelo

Estas superficies fueron ajustadas para toda la región inicialmente y luego fueron cortadas alrededor de las áreas de las fincas para evitar sesgos y valores anómalos por extrapolación.

Sobre cada una de estas interpolaciones, el script de R reporta el error que el proceso de ajuste de los puntos a la función TPS produce (Figura 4, ejemplo para pH). En la mayoría de los casos (>80%), los errores de interpolación estuvieron bastante bajos (<10% error), y sólo en unos pocos casos (<5%) los errores fueron estadísticamente significativos (>50% error). La distribución de errores de interpolación fue similar para las demás variables bajo análisis, indicando un buen desempeño del ajuste TPS en general. El script también realiza un gráfico de dispersión X-Y de valor predicho versus valor real, aunque

estadísticamente, como son datos ruidosos, tiene una mayor significancia la diferencia de cuadrados.

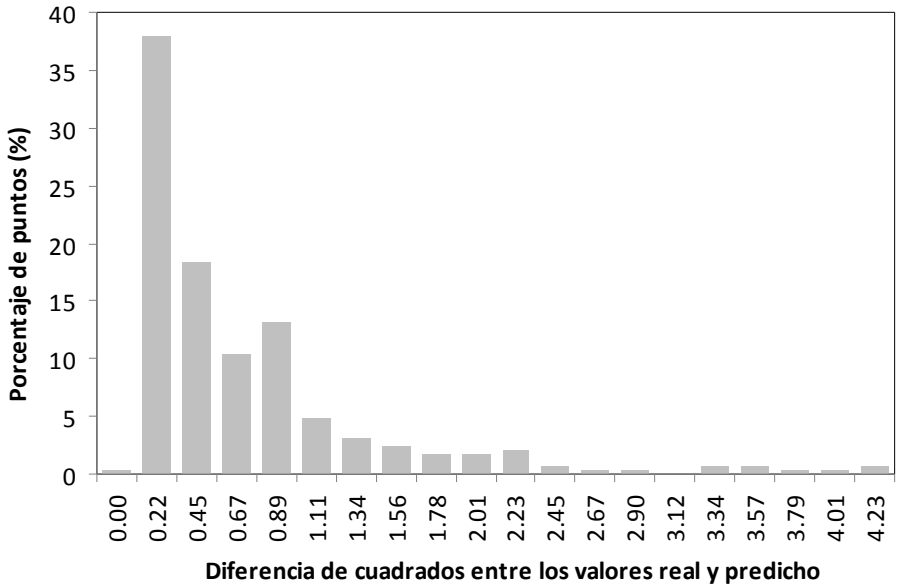

Figura 4 Distribución de frecuencias para la diferencia media de cuadrados (error de interpolación) para los datos de pH de la primera capa

Fácilmente observable es el hecho de que hay poca variabilidad climática a través de la zona, lo que se debe particularmente a que la zona presenta una topografía bastante uniforme (variando entre 0 y 57 metros sobre el nivel del mar). Las zonas más húmedas (de mayor precipitación y humedad relativa) se encuentran al sur, al igual que las áreas de más alta temperatura. No obstante lo anterior, la variación de la temperatura media sólo es de 0.4%, de la temperatura mínima es de 0.3% y de la temperatura máxima es de 1%.

Por su parte, las variables de suelos tienen una mayor variabilidad a través de la zona de estudio, con la composición textural siendo la de mayor variación. Menor variación se encontró en las demás variables (Tabla 1), indicando cierto grado de similaridad, pero aún una significativa variación entre ciertas unidades. En la carpeta ".\zonificacion\variables-originales\suelos" se encuentran todas estas variables en resolución de 3 arco-segundos, en dos sistemas de coordenadas, y en dos formatos diferentes de raster (ESRI Grid y ESRI ASCII)

b. Suavización de datos climáticos y de NDVI

Debido a que hay una diferencia en la resolución espacial de los datos de clima, con respecto al modelo de elevación digital, se realizó un proceso de "smoothing" de las superficies climáticas y de NDVI, con resolución objetivo de 3 arco-segundos (~90m).

Este proceso se realizó mediante interpolación espacial (algoritmo TOPOGRID, ESRI 2008). En ArcGIS, la función usada fue "Toolbox > Spatial Analyst Tools > Topo to Raster" con los parámetros por defecto y usando los centroides de las celdas de 250m (NDVI) y 1km (WorldClim) como puntos de interpolación respectivamente.

Cabe anotar que este procedimiento se realizó con el único fin de uniformizar la resolución, y debe tenerse en cuenta que en ningún caso está proporcionando mayor precisión al clima o NDVI (más allá de sus resoluciones originales), sino simplemente suavizando los valores entre los píxeles de tal manera que los procedimientos estadísticos no se vean sesgados por discontinuidad en los datos espaciales (saltos entre un píxel y otro, y repetición excesiva de valores).

c. Máscara de análisis

La máscara de análisis, por obvias razones es el límite de las fincas, pero dado que el tamaño de píxel brinda una limitada precisión en ciertas zonas (esquinas de ciertas fincas), esta máscara se expandió en aproximadamente el equivalente a 200 metros (3 pixeles) (Figura 5). Con esta máscara se cortaron todos los rasters anteriormente producidos para todos los subsecuentes análisis estadísticos. Estos datos están en el folder ".\zonificacion\variables-analisis-estadistico".

En adición, todos los rasters también se cortaron sobre los perímetros de las fincas. Estos datos se encuentran en el folder ".\zonificacion\variables-corte-fincas". Los datos para la máscara de análisis estadístico se crearon con el único fin de realizar los análisis estadísticos pertinentes, en tanto que los datos que están cortados por los linderos de las fincas, tienen como objetivo principal, ser usados en mapeo.

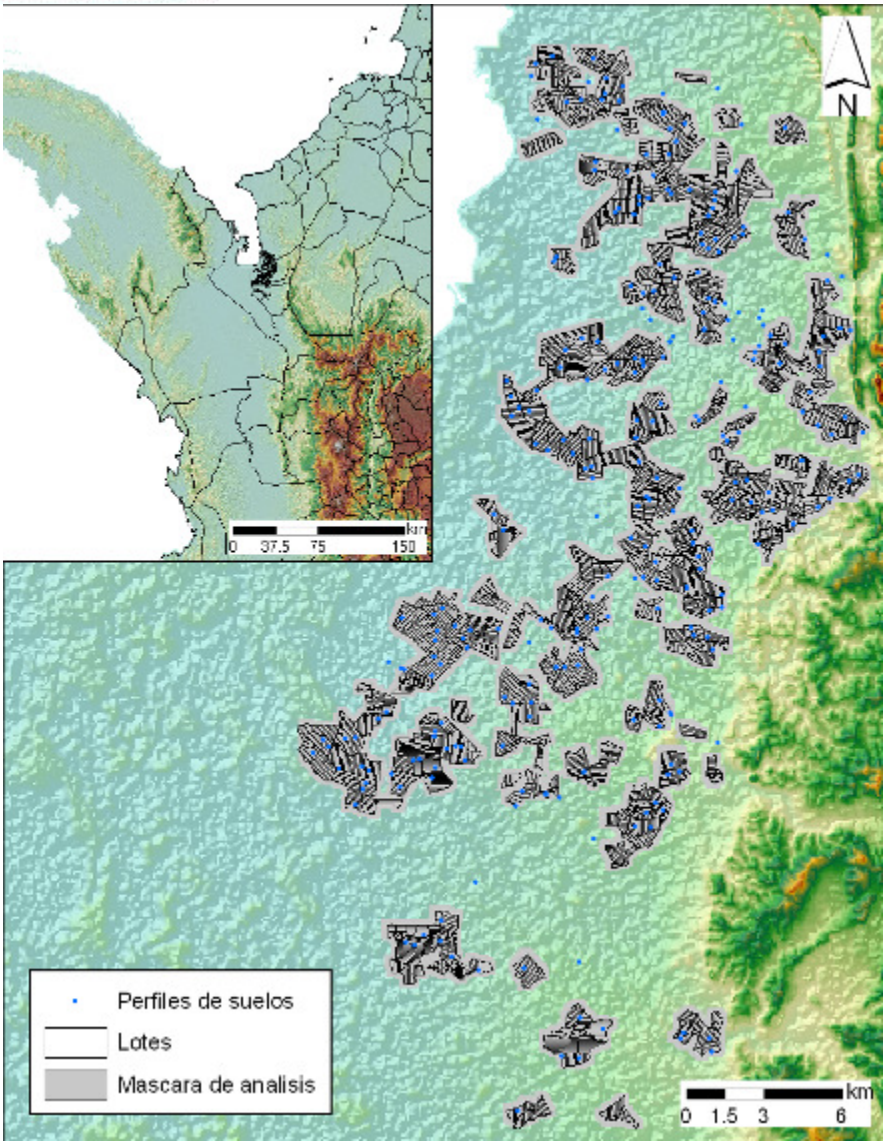

Figura 5 Mascara de análisis (área gris) y linderos de las fincas, superpuesta con modelo de elevación digital (DEM)

4. Selección de variables más importantes

Es sabido que cualquier procedimiento de estadística multivariada (regresiones, agrupamientos, entre otros) tiende a verse afectado por el número de variables de entrada usadas para la ejecución del procedimiento en cuestión. Este fenómeno es conocido como "sobre-estimación" u "*overfitting*". Cuando ocurre sobre-estimación en un modelo es probable que el desempeño del mismo sea considerablemente alto, pero que se convierta en algo tan específico que tienda a fallar una variedad de circunstancias para las que no está preparado. En otras palabras, se pierde algo de generalización en el modelo, y se tiende a volver demasiado específico a los datos.

Para evitar este fenómeno, existen métodos de selección de variables (procedimientos *stepwise, forward o backward regression*), análisis de correlación canónicos, análisis de discriminante canónico, análisis de correlación sencilla, análisis de componentes principales, entre otros. Para el efecto del presente proyecto, el uso del set completo de variables dado el pequeño espacio que ocupan las 179 fincas analizadas, podría causar sobre-estimación y dada esta condición, se realizaron dos pruebas estadísticas: análisis de correlación entre las variables y análisis de componentes principales.

a. Correlaciones entre variables usadas en el estudio

En ocasiones un grupo de variables que describen una determinada región o un determinado fenómeno no necesariamente contiene información no-relacionada entre ella. En otras palabras, a diferentes niveles, dentro de las variables, existe siempre cierto grado de duplicidad en la información. Esta duplicidad se da porque existen relaciones entre las variables. La forma más fácil de detectar estas relaciones es una matriz de correlaciones.

Se extrajeron los valores en las variables climáticas, de suelos, de NDVI y la elevación, en todos los puntos donde se tomó un perfil de suelos (usando el script "Correl-Uniban.R"), y usando Excel, se efectuó un análisis de correlaciones sobre los datos (Tabla 4).

Como se puede observar, hay muchas correlaciones estadísticamente significativas, lo que significa que hay un determinado grado de duplicidad en la información, y en particular en la información de suelos, que parece estar altamente relacionada con la elevación. Sin embargo, en algunos casos estas correlaciones no parecen ser fuertes (encima de 0.7). Por otro lado, hay variables que están totalmente representadas en otras, tal es el caso del porcentaje de limos, que tiene una correlación muy alta con el porcentaje de arcillas (R=1.00), o el rango de NDVI total, que está altamente relacionado con la desviación estándar histórica del NDVI (R=0.94). Muchos otros casos pueden observarse en la Tabla 4. El porcentaje de limos estuvo parcialmente representado en el porcentaje de arenas y de arcillas, y casi todas las variables presentaron muy alta correlación con la elevación, indicando que esta última es un factor preponderante en la distribución tanto del clima, como de las propiedades físico-químicas del suelo. Muy probablemente también sea un factor importante en cuanto a la producción en las diferentes fincas.

No obstante lo anterior, se requiere un criterio estadístico que converja con las correlaciones y que permita decidir cuál variable de un par de variables debe conservarse y cuál debe ser conservada. Por este motivo, se usó el análisis de componentes principales, que permite conocer cuáles variables tienen más significancia estadística, explicando mayores proporciones de varianza que las demás.

Tabla 4 Matriz de correlaciones (n=296) de todas las variables usadas en procedimientos estadísticos

	Alt	P1	P4	P7	P12	P15	NDVIx	NDVIm	NDVIn	NDVImr	NDVIr	NDVIsd	K-Barr	K-inve	K-labo	Dr	DMP	Est-Agr	Hg-cc	NF	Ar(%)	A(%)	L(%)	pH
Alt	1.00	-0.10	0.01	**0.28**	**0.22**	0.08	-0.10	-0.07	-0.02	**0.13**	**-0.14**	**-0.19**	**-0.68**	**0.56**	**-0.47**	**0.70**	**-0.42**	**-0.97**	**-0.82**	**0.57**	**0.81**	**0.94**	**0.81**	**-0.73**
P1	-0.10	1.00	**0.88**	**0.88**	**0.79**	**0.85**	-0.10	**-0.19**	**-0.24**	**0.17**	**0.36**	**0.39**	**0.57**	**-0.50**	**0.14**	-0.02	**-0.58**	**0.17**	**0.38**	**-0.64**	**-0.27**	**-0.37**	**-0.27**	0.06
P4	0.01	**0.88**	1.00	**0.88**	**0.86**	**0.90**	-0.07	**-0.14**	**-0.20**	**0.19**	**0.33**	**0.36**	**0.51**	**-0.35**	-0.02	0.11	**-0.70**	0.06	**0.23**	**-0.63**	-0.11	**-0.28**	-0.11	0.07
P7	**0.28**	**0.88**	**0.88**	1.00	**0.87**	**0.87**	**-0.16**	**-0.23**	**-0.28**	**0.28**	**0.32**	**0.32**	**0.29**	**-0.29**	-0.03	**0.27**	**-0.75**	**-0.20**	0.07	**-0.41**	0.04	-0.01	0.04	**-0.22**
P12	**0.22**	**0.79**	**0.86**	**0.87**	1.00	**0.97**	**-0.16**	**-0.23**	**-0.28**	**0.28**	**0.33**	**0.33**	**0.38**	**-0.25**	-0.10	**0.39**	**-0.89**	**-0.15**	0.07	**-0.45**	0.10	-0.05	0.10	**-0.13**
P15	0.08	**0.85**	**0.90**	**0.87**	**0.97**	1.00	**-0.15**	**-0.23**	**-0.29**	**0.26**	**0.38**	**0.40**	**0.52**	**-0.27**	**-0.12**	**0.26**	**-0.82**	-0.01	**0.15**	**-0.62**	-0.04	**-0.22**	-0.04	0.06
NDVIx	-0.10	-0.10	-0.07	**-0.16**	**-0.16**	**-0.15**	1.00	**0.97**	**0.90**	**-0.22**	-0.07	**-0.12**	0.05	0.04	-0.04	0.03	0.10	0.08	0.02	-0.02	0.11	-0.06	0.11	0.11
NDVIm	-0.07	**-0.19**	**-0.14**	**-0.23**	**-0.23**	**-0.23**	**0.97**	1.00	**0.97**	**-0.39**	**-0.29**	**-0.33**	-0.02	0.09	-0.07	0.08	**0.14**	0.04	-0.03	0.06	**0.18**	-0.01	**0.18**	0.09
NDVIn	-0.02	**-0.24**	**-0.20**	**-0.28**	**-0.28**	**-0.29**	**0.90**	**0.97**	1.00	**-0.51**	**-0.49**	**-0.52**	-0.11	**0.14**	-0.08	**0.13**	**0.16**	0.00	-0.08	**0.13**	**0.23**	0.05	**0.23**	0.05
NDVImr	**0.13**	**0.17**	**0.19**	**0.28**	**0.28**	**0.26**	**-0.22**	**-0.39**	**-0.51**	1.00	**0.74**	**0.73**	0.00	-0.10	0.06	0.00	**-0.23**	-0.10	0.01	-0.03	-0.07	0.06	-0.07	**-0.15**
NDVIr	**-0.14**	**0.36**	**0.33**	**0.32**	**0.33**	**0.38**	-0.07	**-0.29**	**-0.49**	**0.74**	1.00	**0.94**	**0.35**	**-0.24**	0.10	**-0.23**	**-0.17**	**0.17**	**0.22**	**-0.33**	**-0.33**	**-0.25**	**-0.33**	**0.12**
NDVIsd	**-0.19**	**0.39**	**0.36**	**0.32**	**0.33**	**0.40**	**-0.12**	**-0.33**	**-0.52**	**0.73**	**0.94**	1.00	**0.39**	**-0.25**	0.09	**-0.26**	**-0.16**	**0.21**	**0.25**	**-0.38**	**-0.36**	**-0.30**	**-0.36**	**0.18**
K-Barr	**-0.68**	**0.57**	**0.51**	**0.29**	**0.38**	**0.52**	0.05	-0.02	-0.11	0.00	**0.35**	**0.39**	1.00	**-0.59**	**0.29**	**-0.36**	-0.11	**0.73**	**0.71**	**-0.78**	**-0.64**	**-0.83**	**-0.64**	**0.58**
K-inve	**0.56**	**-0.50**	**-0.35**	**-0.29**	**-0.25**	**-0.27**	0.04	0.09	**0.14**	-0.10	**-0.24**	**-0.25**	**-0.59**	1.00	**-0.90**	**0.43**	-0.05	**-0.69**	**-0.93**	**0.27**	**0.70**	**0.63**	**0.70**	0.03
K-labo	**-0.47**	**0.14**	-0.02	-0.03	-0.10	**-0.12**	-0.04	-0.07	-0.08	0.06	0.10	0.09	**0.29**	**-0.90**	1.00	**-0.49**	**0.35**	**0.58**	**0.82**	**0.12**	**-0.64**	**-0.41**	**-0.64**	**-0.23**
Dr	**0.70**	-0.02	0.11	**0.27**	**0.39**	**0.26**	0.03	0.08	**0.13**	0.00	**-0.23**	**-0.26**	**-0.36**	**0.43**	**-0.49**	1.00	**-0.69**	**-0.70**	**-0.60**	**0.29**	**0.84**	**0.63**	**0.84**	**-0.45**
DMP	**-0.42**	**-0.58**	**-0.70**	**-0.75**	**-0.89**	**-0.82**	0.10	**0.14**	**0.16**	**-0.23**	**-0.17**	**-0.16**	-0.11	-0.05	**0.35**	**-0.69**	1.00	**0.39**	**0.23**	**0.27**	**-0.44**	**-0.19**	**-0.44**	**0.20**
Est-Agr	**-0.97**	**0.17**	0.06	**-0.20**	**-0.15**	-0.01	0.08	0.04	0.00	-0.10	**0.17**	**0.21**	**0.73**	**-0.69**	**0.58**	**-0.70**	**0.39**	1.00	**0.90**	**-0.57**	**-0.85**	**-0.96**	**-0.85**	**0.65**
Hg-cc	**-0.82**	**0.38**	**0.23**	0.07	0.07	**0.15**	0.02	-0.03	-0.08	0.01	**0.22**	**0.25**	**0.71**	**-0.93**	**0.82**	**-0.60**	**0.23**	**0.90**	1.00	**-0.44**	**-0.84**	**-0.85**	**-0.84**	**0.31**
NF	**0.57**	**-0.64**	**-0.63**	**-0.41**	**-0.45**	**-0.62**	-0.02	0.06	**0.13**	-0.03	**-0.33**	**-0.38**	**-0.78**	**0.27**	**0.12**	**0.29**	**0.27**	**-0.57**	**-0.44**	1.00	**0.47**	**0.78**	**0.47**	**-0.78**
Ar(%)	**0.81**	**-0.27**	-0.11	0.04	0.10	-0.04	0.11	**0.18**	**0.23**	-0.07	**-0.33**	**-0.36**	**-0.64**	**0.70**	**-0.64**	**0.84**	**-0.44**	**-0.85**	**-0.84**	**0.47**	1.00	**0.81**	**1.00**	**-0.46**
A(%)	**0.94**	**-0.37**	**-0.28**	-0.01	-0.05	**-0.22**	-0.06	-0.01	0.05	0.06	**-0.25**	**-0.30**	**-0.83**	**0.63**	**-0.41**	**0.63**	**-0.19**	**-0.96**	**-0.85**	**0.78**	**0.81**	1.00	**0.81**	**-0.76**
L(%)	**0.81**	**-0.27**	-0.11	0.04	0.10	-0.04	0.11	**0.18**	**0.23**	-0.07	**-0.33**	**-0.36**	**-0.64**	**0.70**	**-0.64**	**0.84**	**-0.44**	**-0.85**	**-0.84**	**0.47**	**1.00**	**0.81**	1.00	**-0.46**
pH	**-0.73**	0.06	0.07	**-0.22**	**-0.13**	0.06	0.11	0.09	0.05	**-0.15**	**0.12**	**0.18**	**0.58**	0.03	**-0.23**	**-0.45**	**0.20**	**0.65**	**0.31**	**-0.78**	**-0.46**	**-0.76**	**-0.46**	1.00

*Valores en negrilla son estadísticamente significativos a 0.05

b. Análisis de componentes principales

El análisis de componentes principales (Shaw, 2003) es una técnica de análisis multivariado que permite transformar un grupo de variables N, relacionadas entre ellas, en un grupo de variables N, completamente ortogonales entre ellas, y con diferente significancia en términos de relación con los datos (explicación de varianza). El análisis de componentes principales no sólo es útil para reducir la dimensionalidad de un grupo de variables que caracteriza una serie de datos, sino que también es útil para escoger las variables que mayor representatividad tienen respecto al set de datos, mediante la escogencia umbrales de significancia.

Usando cada uno de los puntos (pixeles) correspondientes a las fincas bajo análisis (máscara con 3 pixeles de buffer) con el DEM (Modelo de Elevación Digital) como base cartográfica, se obtuvieron un total de 29,570 diferentes puntos y se extrajeron los valores para cada una de las variables, luego se estandarizaron los valores de tal manera que se obtuviera una varianza igual a la unidad (1) y un promedio igual a cero (0). Esta estandarización se realizó con el objetivo de igualar las escalas de las variables.

Mediante un script de SAS "./zonificacion/analisis-estadistico/ UNIBAN-PCA.sas", utilizando los procedimientos "PROC STANDARD" y "PROC PRINCOMP", se realizó tanto la estandarización de los datos como el análisis de componentes principales. Usando todos los datos del área de estudio. El análisis de componentes principales (Figura 6) indicó que las cuatro primeras componentes explicaron más del 85% de la varianza presente en los datos, en tanto que las primeras dos componentes explicaron (juntas) más del 63% de la varianza.

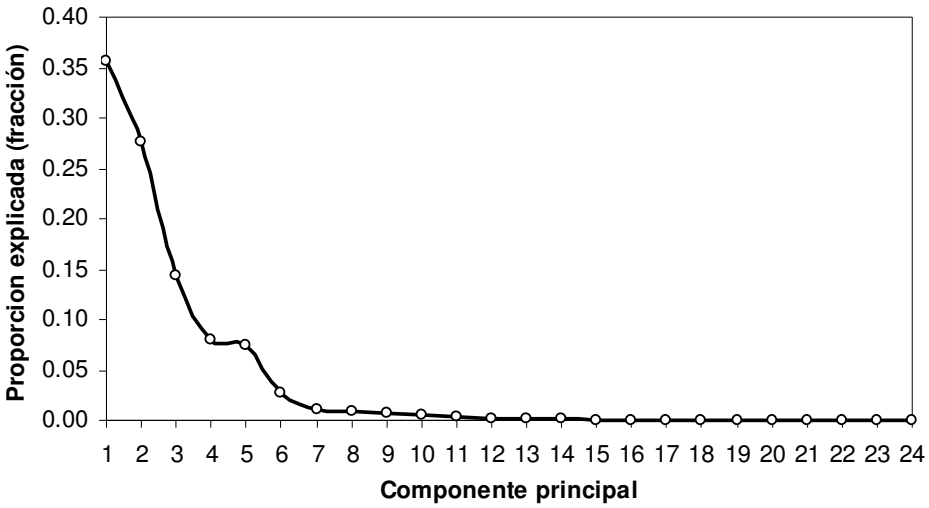

Figura 6 Proporción de varianza explicada por componente principal

Cabe anotar que algunas de las variables climáticas y de suelos no fueron incluidas en el análisis de componentes principales debido a que tenían inicialmente muy poca

variabilidad, o debido a que estaban representadas en otras vairables (e.g. contenidos de humedad volumétricos y grafimétricos, densidad aparente, temperaturas), o porque tienden a presentar alta variación temporal, esto es, ser inestables en el tiempo (densidad aparente, que cambia con respecto a la humedad, la porosidad, entre otros).

Mediante el análisis de componentes principales se encontró que 20 de las 24 variables explican de manera significativa la varianza de los datos. Estas variables presentaron pesos considerables (arriba del 7%) en las dos primeras componentes principales (Figura 7). Una cosa importante que cabe anotar, es que las variables quedaron agrupadas por tipo. En este sentido, las variables climáticas (cuadrante arriba izquierda) estuvieron relativamente cercanas en sus puntajes en ambas componentes (PC1 y PC2). De la misma manera ocurrió con el porcentaje de arenas, arcillas, y la profundidad del perfil (nivel freático), que son parámetros que tienden a estar significativamente relacionados. El porcentaje de limo en el suelo, por ejemplo, muestra poca importancia, probablemente debido al hecho que está altamente correlacionado con el porcentaje de arcillas.

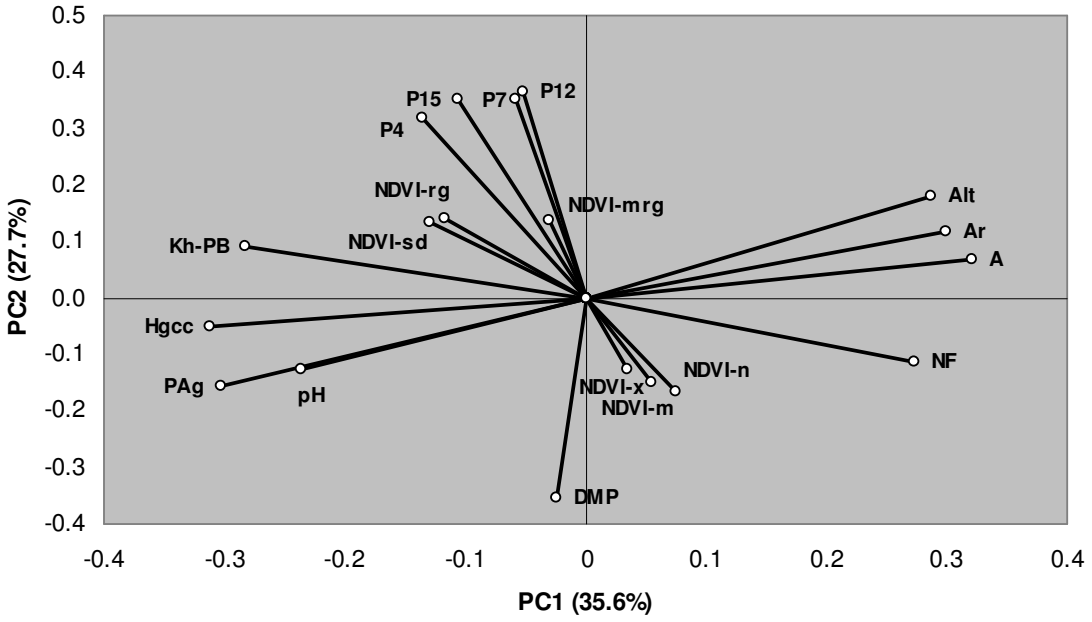

Figura 7 Resultados del análisis de componentes principales para las 20 variables que resultaron importantes para la región. IDs y Nombres de las variables están indicados en el archivo "nombreVariables.xls" columna ID variable.

Las variables de NDVI, de otro lado, están agrupadas en el cuadrante inferior derecho, y las variables relacionadas con el movimiento del agua en el suelo (conductividad hidráulica, humedad gravimétrica a capacidad de campo, estado de agregación) también aparecen diferenciadas de las demás variables hacia la derecha. El análisis de componentes principales demostró que aunque una variable esté relacionada con otra esto no significa que no aporte características adicionales a la explicación de un fenómeno.

5. Agrupamiento de zonas homólogas

Usando cada uno de los puntos (pixeles) correspondientes a las fincas bajo análisis con el DEM (Modelo de Elevación Digital) como base cartográfica, se obtuvieron un total de 29,570 diferentes puntos a ser agrupados. Para dicha agrupación se utilizó un método estadístico conocido como "análisis de conglomerados" o *clustering*. Un análisis de conglomerados permite agrupar con un criterio estadístico sólido una serie de puntos en un espacio multidimensional compuesto por un número de variables N. Seleccionando un número de conglomerados o *clusters* determinado, o estableciendo un nivel de similitud (umbral), se pueden crear grupos estables de puntos a partir de los cuales se realizan análisis posteriores, ya con un número menor de puntos (típicamente los centroides de cada conglomerado o *cluster*).

Para el caso de las fincas de CI UNIBAN, se usó el método de *clustering* jerárquico de Ward (1963), usando como criterio de separación la varianza entre los diferentes posibles grupos de datos. Las variables incluidas en el clustering fueron las seleccionadas de acuerdo al ACP. El análisis de conglomerados se realizó mediante el software SAS (PROC CLUSTER, PROC TREE) y el script con el que se realizó este análisis se encuentra en "./zonificacion/analisis-estadistico/UNIBAN-conglomerados.sas"

El número óptimo de conglomerados (grupos o *clusters*) para los puntos y variables usadas fue de 20, después de realizar diferentes pruebas con diferentes números de conglomerados. Esto resulta en que, de las 179 fincas que se analizaron, sólo se requieren 20 diferentes grupos agroecológicos para el manejo de la producción. No obstante, este número puede ser sometido a revisión por parte de expertos de UNIBAN o productores, que seguramente tienen un mayor conocimiento de la región. Se encontraron diferencias significativas entre las fincas del norte y del sur (Figuras 8, 9 y 10), aunque las diferencias más grandes se presentaron entre fincas de la zona este (cercanas a la cordillera) y fincas de la zona oeste (cercanas a la cosa).

Interesantemente, y aunque la variabilidad dentro de la región no es excesivamente alta para las variables estudiadas, se encontró un nivel de agrupamiento, que resultó en el agrupamiento de algunos sectores de fincas con alta similaridad agroecológica. Además, se encontró que las propiedades de los suelos, así como las variables climáticas, no tienen un comportamiento que responda a divisiones creadas por el hombre (divisiones de parcelas, o de fincas), sino que responden a gradientes que dependen de la topografía, de la génesis del suelo, y de patrones de nubosidad, y demás. Las zonas agroecológicas se entregan en dos formatos diferentes, shapefile y raster, ambos en el folder "./zonificacion/zonas-agroecologicas". Consultar el documento "./zonificacion/MetadatosGeneral.doc" para detalles técnicos de los archivos contenidos en estas carpetas.

Cada zona agroecológica (*cluster*) tiene unas características específicas que la diferencian de las demás zonas, tal como se observa en la tabla 5. Tal como se observa, las zonas de mayor elevación, corresponden también con aquellas zonas con suelos menos pesados, y menos profundos, aunque con una mayor densidad real, probablemente debido al mayor contenido de arenas. Las zonas de alta precipitación coinciden con suelos con tabla de agua mucho menos profunda, y viceversa. En cuanto al pH de la primera capa de suelo, los suelos más ácidos (pH más bajo) se encuentran en zonas más cercanas a la cordillera, donde probablemente, la influencia de la misma causa una ligera acidez en los suelos.

Tabla 5 Propiedades de suelos y principales variables ambientales características de cada zona

Zona	Alt (m)	Limo (%)	Arena (%)	Arcilla (%)	Dr	NDVI promedio	NDVI DE	pH	NF	Prec
1	17.7	36.1	18.5	45.4	2.49	0.8447	0.0116	5.62	159.4	2582.5
2	19.3	34.8	18.9	46.3	2.50	0.8628	0.0110	5.63	158.7	2596.4
3	16.7	37.6	18.3	44.1	2.47	0.8373	0.0157	5.58	161.0	2589.9
4	17.4	36.2	18.7	45.1	2.48	0.8548	0.0135	5.60	161.4	2568.8
5	17.6	37.7	17.4	44.9	2.50	0.8305	0.0133	5.63	152.2	2703.9
6	16.1	38.6	17.6	43.8	2.47	0.8191	0.0133	5.61	157.0	2611.1
7	21.1	35.7	17.6	46.6	2.52	0.8489	0.0117	5.63	147.4	2819.1
8	11.7	43.6	15.1	41.3	2.42	0.7786	0.0128	5.70	146.9	2703.3
9	22.4	36.0	17.6	46.3	2.57	0.8197	0.0153	5.56	145.6	3123.7
10	25.5	34.8	18.7	46.6	2.59	0.7777	0.0173	5.49	147.8	3276.5
11	15.0	40.8	16.4	42.8	2.45	0.8018	0.0148	5.64	150.6	2700.8
12	10.3	44.7	14.9	40.4	2.39	0.7766	0.0195	5.72	148.2	2643.8
13	10.3	45.9	13.5	40.6	2.43	0.8172	0.0226	5.79	137.6	2766.2
14	19.4	38.7	18.1	43.2	2.43	0.7451	0.0151	5.65	153.5	2676.0
15	21.6	37.1	17.2	45.7	2.57	0.8003	0.0163	5.56	144.6	3179.3
16	15.5	41.1	15.5	43.3	2.47	0.8297	0.0178	5.70	142.9	2792.4
17	30.8	30.9	23.5	45.6	2.50	0.6602	0.0189	5.45	170.7	2612.6
18	22.3	36.0	20.0	44.0	2.48	0.7170	0.0217	5.52	162.7	2714.7
19	30.2	30.8	23.4	45.8	2.50	0.5860	0.0144	5.47	170.8	2594.9
20	31.0	30.7	23.7	45.5	2.50	0.5160	0.0162	5.43	172.0	2620.8

6. Conclusiones

En resumen, se ha realizado una clasificación agroecológica basada en variables edáficas, topográficas y climáticas. La clasificación está sujeta a modificaciones, dada la posibilidad de involucrar variables adicionales, especialmente aquellas provenientes de mediciones de campo con estaciones meteorológicas, que hasta la fecha han sido difíciles de involucrar en el análisis, debido a que los datos están en algunos casos sucios, incompletos y muestran algunas inconsistencias.

Con esta segunda clasificación, sin embargo, es posible trabajar los productos complementarios del convenio CIAT-UNIBAN, dado que provee la base necesaria para una agrupación de fincas, con base tanto en los datos de estudio semi-detallado de suelos, datos confiables ambientales (clima, topografía), y en un criterio estadístico estable.

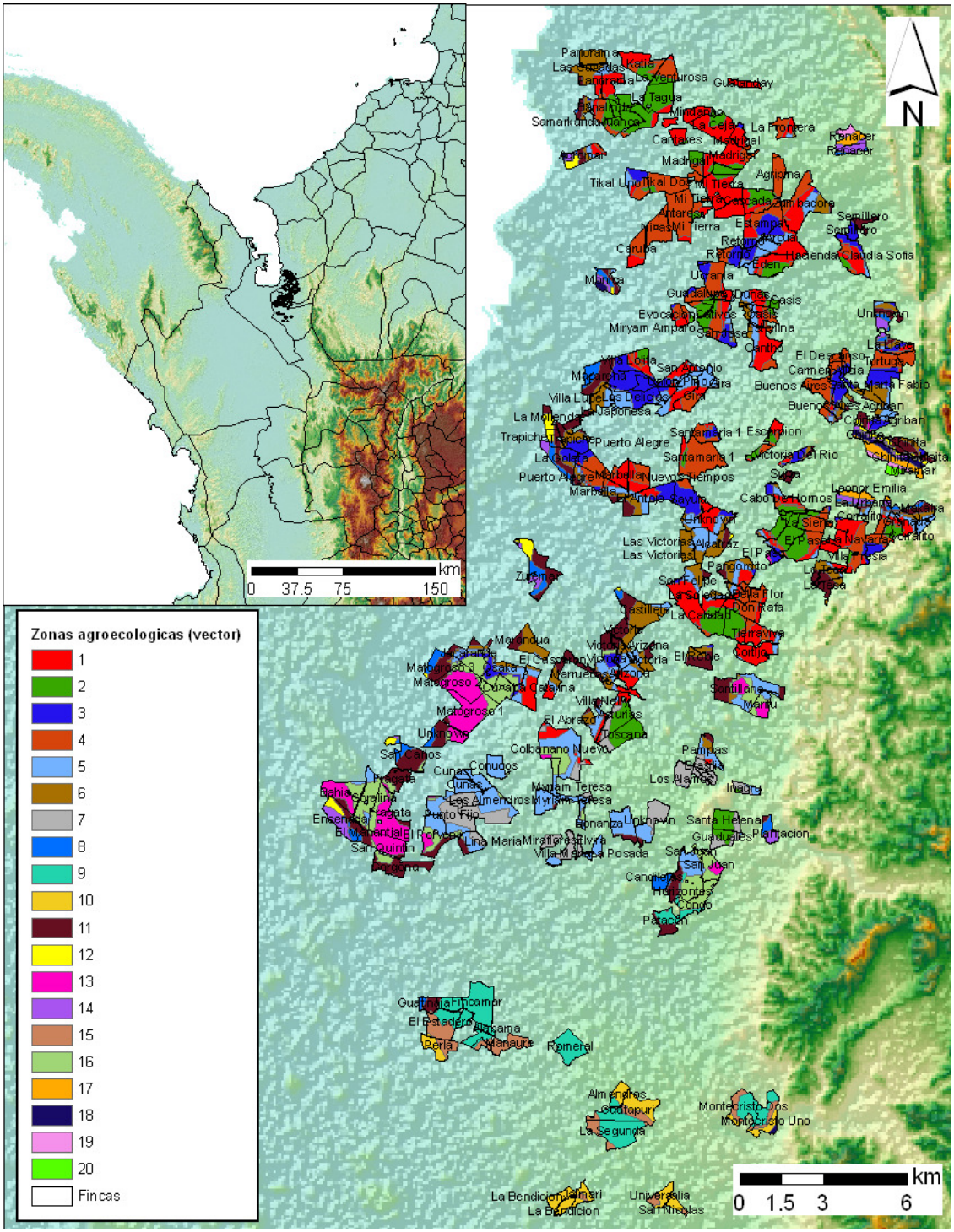

Figura 9 Versión 2 de la zonificación agroecológica

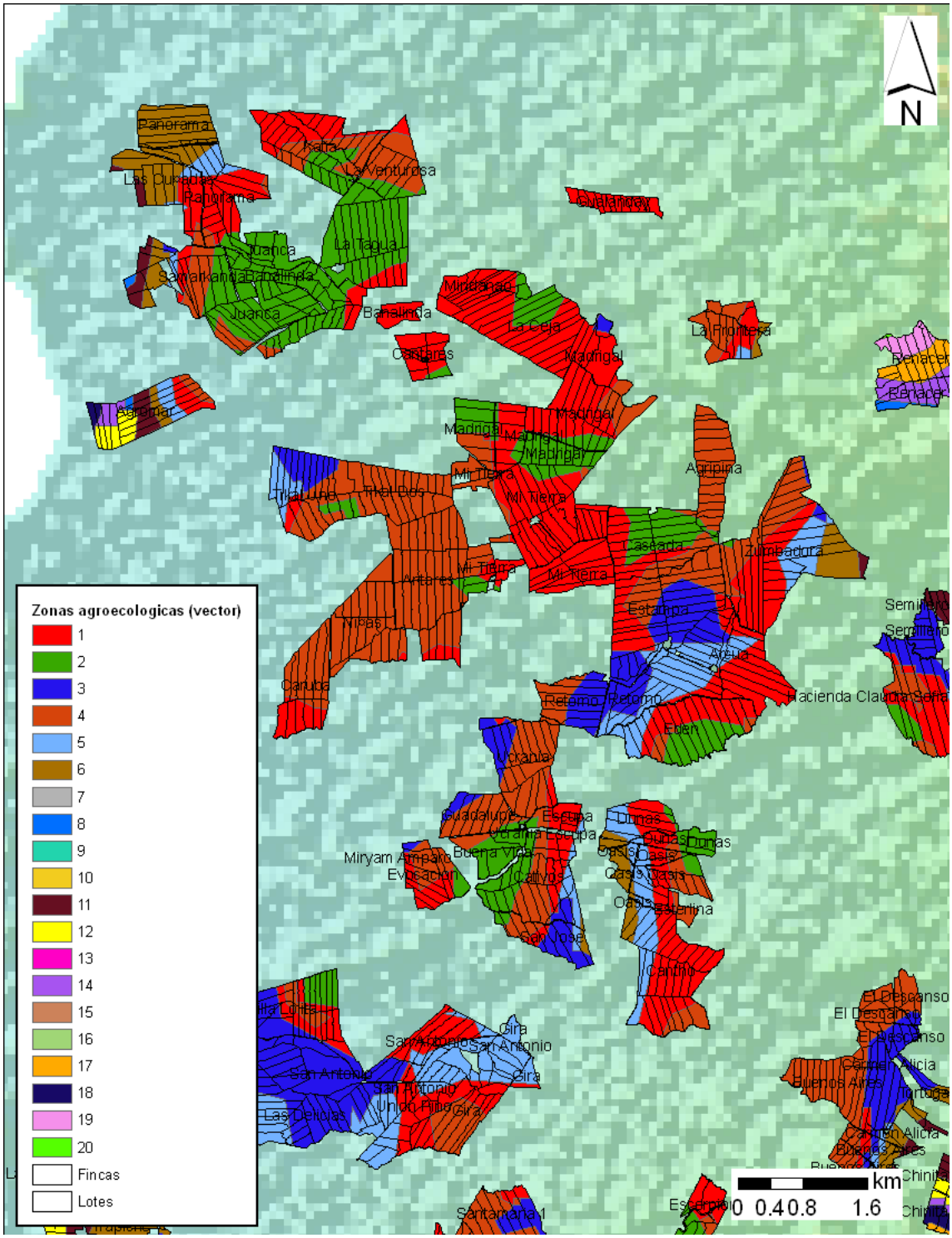

Figura 10 Versión 1 de la zonificación agroecológica, ampliada al norte

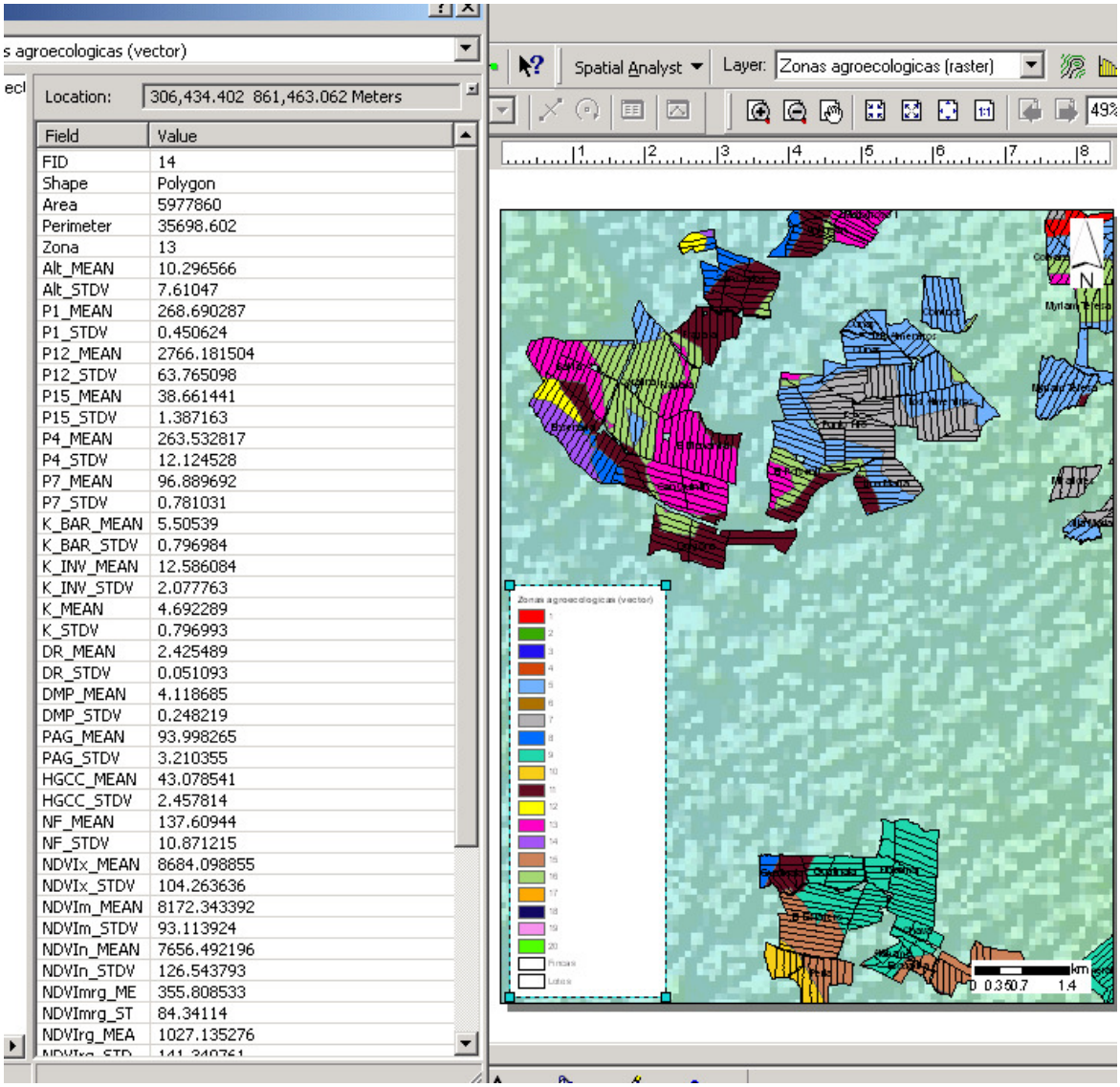

Figura 11 Versión 2 de la zonificación agroecológica, ampliada al sur, con consulta sobre tabla de atributos del shapefile

Referencias

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 25:1965-1978.

Hutchinson MF (1984) A summary of some surface fitting and contouring programs for noisy data.*CSIRO Division of Mathematics and Statistics, Consulting Report ACT* 84/6. Canberra, Australia.

Hutchinson MF, de Hoog FR (1985) Smoothing noisy data with spline functions. *Numerische Mathematik* 47: 99-106.

Jarvis A, Reuter HI, Nelson A, Guevara E (2008) Hole-filled seamless SRTM data V4, International Centre for Tropical Agriculture (CIAT), available from http://srtm.csi.cgiar.org.

Ward JH (1963) Hierarchical Grouping to optimize an objective function. Journal of American Statistical Association, 58(301), 236-244.

Shaw PJA (2003) Multivariate statistics for the Environmental Sciences. Hodder-Arnold.

Parte 2. Estudio de productividad sobre 14 fincas bananeras de Urabá

Resumen

El presente documento describe los materiales usados, métodos aplicados y resultados obtenidos por el Centro Internacional de Agricultura Tropical (CIAT) respecto al estudio de productividad de 14 fincas del Urabá bajo producción bananera afiliadas a la compañía CI UNIBAN. El análisis consistió de 4 pasos básicos: (1) colección de datos de entrada, (2) uniformización y pre-procesamiento de datos de entrada, (3) análisis de productividad y variabilidad histórico al nivel de finca, (4) análisis comparativo de productividad media al nivel de fincas, (5) análisis comparativo de productividad al nivel de zonas agroecológicas, (6) análisis comparativo de productividad al nivel de variedades (clones), y (7) análisis complementarios de datos de productividad. Se emplearon datos de campo colectados al nivel de lote y al nivel de finca (sobre 14 fincas en total) de peso de racimo, productividad, racimos embolsados, variedades (clones) utilizados, racimos cosechados (cortados), y los resultados de la zonificación agroecológica. Se aplicaron procedimientos estadísticos como el "análisis de mundo pequeño" para llenar vacíos en la información, un algoritmo de aprendizaje llamado *Feed-forward propagation neural network* con la función de Levenberg-Marquardt para optimización para realizar el ajuste de datos de productividad hasta el nivel de lote, y con esto se sumarizaron los datos en gráficos históricos y en curvas de "isoproductividad". Se presenta la metodología utilizada, y se realizan algunas recomendaciones para la información suministrada.

Contenido

---

1. Introducción

Dada la cantidad de información que día tras día se colecta en campo, dadas las exigencias por parte del mercado, y asimismo la presión tanto biótica (enfermedades, plagas) como abiótica (factores ambientales), se hace necesario analizar las variables de producción y encontrar maneras de monitorear los efectos que tiene el ambiente en la respuesta de los diferentes genotipos cultivados, y en adición a esto, generar un despliegue de información que sea accesible para el productor y que permita a UNIBAN y al productor controlar, decidir y aplicar los correctivos necesarios, así como también prever futuras situaciones o riesgos de pérdida de producción y/o productividad.

Como segundo paso en la consecución de lo anterior, es importante conocer el comportamiento de la producción en diferentes dimensiones y/o niveles productivos: lotes, fincas, zonas agroecológicas, variedades utilizadas, para así no sólo detectar problemas, sino oportunidades y aplicar los correctivos que sean necesarios de tal manera que la producción sea realizada de la manera más eficiente, optimizando insumos y obteniendo mejores márgenes de utilidad. Para ese fin, se hace necesario desarrollar métodos comparativos y automatizados que permitan analizar la producción de una manera interactiva y eficiente. En este documento presentamos resultados usando

diferentes variables y diferentes tipos de gráfico. Estos métodos pueden ser aplicados por UNIBAN CI para elaborar un sistema mucho más completo de despliegue de información.

2. Colección de datos de entrada

a. Datos de productividad (cajas/semana/ha)

Se recibió información correspondiente a productividad desde el año 2000 hasta el año 2009, para las fincas que conforman la región, mensualmente. Sin embargo, se observaron algunos problemas con la uniformidad de la información recibida (tabla 1).

Tabla 1 Disponibilidad y uniformidad de información de productividad. Los números dentro de las cajas indican el número de fincas con datos

Año/mes	Ene	Feb	Mar	Abr	May	Jun	Jul	Ago	Sep	Oct	Nov	Dic
2000	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	106
2001	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	122
2002	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	123
2003	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	131
2004	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	139
2005	139	141	140	140	140	140	140	141	141	142	142	141
2006	141	142	141	141	141	140	142	141	140	140	140	141
2007	140	139	139	139	139	139	137	138	143	143	136	136
2008	137	137	138	138	141	141	196	196	194	193	193	193
2009	193	195	196	196	197	197	197	197	197	197	197	NA

Como se puede observar, solo a partir de 2005 hay una uniformidad considerable en los datos disponibles, y a partir de la mitad del 2008 hay un aumento considerable en el número de fincas. Dado que el análisis es para solamente 14 fincas, la uniformidad de los datos es mucho mejor para estas. No obstante, los análisis se enfocarán en los años para los que existan suficientes datos. Inicialmente podría sugerirse que este período fuese desde 2005 hasta 2009, pero esto depende también de la disponibilidad de datos de peso de racimo, embolse, entre otros.

b. Datos de peso de racimo

Se recibieron datos de peso de racimo (en kilogramos) a nivel de lotes, por semana, para 14 fincas: Castilletes, La Ceja, Colbanano, Coralina, El Edén, Estadero, Inagrú, Leonor Emilia, Marandua, Pan Gordito, Paso Estoril, Porvenir, Puerto Alegre, Yerbabuena. La disponibilidad de los datos también fue un factor importante en el peso de racimo. En general, las 14 fincas mostraron un número de años con datos muy variable entre ellas (tabla 2). En solo algunos casos el total de medidas por año y por finca fue del 100% (El Edén año 2007, entre otros), mientras que en otros hubo una evidente carencia de

información, y el número de datos disponibles llegó sólo al 50% (Castilletes 1999, Pan Gordito 2008, entre otros). Los años para los que la mayoría de las fincas tuvo al menos 50% de los datos fueron 2007, 2008 y 2009, aunque algunas fincas sólo presentaron datos para uno de estos tres años.

Tabla 2 Disponibilidad de datos de peso de racimo por finca y por año. Casillas por año indican porcentaje de datos sobre un total de 52 semanas y el total de lotes

Finca	#Lotes	#Medidas potencial	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Castilletes	35	1820	49.8	64.1	76.0	85.7	96.8	97.5	99.7	94.3	94.0	89.1	63.6
La Ceja	13	676	NA	NA	NA	NA	NA	NA	NA	NA	99.9	84.6	80.8
Colbanano	46	2392	NA	NA	NA	NA	NA	NA	NA	NA	NA	56.9	NA
Coralina	11	572	NA	NA	NA	NA	NA	NA	NA	NA	99.8	82.7	NA
El Eden	16	832	NA	NA	NA	NA	NA	NA	NA	NA	100.0	83.9	80.8
Estadero	21	1092	NA	NA	NA	NA	NA	NA	NA	NA	100.0	100.0	25.1
Inagru	4	208	NA	NA	NA	NA	NA	NA	NA	NA	NA	100.0	86.5
Leonor Emilia	7	364	NA	NA	NA	NA	NA	NA	NA	NA	65.1	65.1	NA
Marandua	21	1092	NA	NA	NA	NA	NA	NA	NA	NA	71.8	NA	NA
Pan Gordito	14	728	NA	NA	NA	NA	NA	NA	NA	NA	100.0	55.6	NA
Paso Estoril	50	2600	NA	NA	NA	NA	NA	NA	NA	NA	NA	81.9	75.0
Porvenir	15	780	NA	NA	NA	NA	NA	NA	NA	NA	99.9	86.4	NA
Puerto Alegre	23	1196	NA	NA	NA	NA	NA	NA	NA	NA	NA	91.1	NA
Yerbabuena	13	676	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	61.5

Los análisis ulteriores se enfocarán, por tanto, en aquellos años con mejor disponibilidad de datos, dependiendo de las fincas.

c. Datos de corte y embolse de racimos

Se recibieron datos de embolse a nivel de lote y corte a nivel de fincas. Dado el enfoque del presente estudio, los datos de corte a nivel de finca no se usaron. Tiempo después fueron recibidos los datos de corte a nivel de lote, pero dado el corto tiempo para entrega de resultados, no se incorporaron estos datos en el análisis. Se sugiere, sin embargo, que los datos de corte se incorporen por parte de UNIBAN CI, y que se apliquen las a ellos las metodologías usadas en el presente documento.

Tabla 3 Disponibilidad de datos de embolse por finca y por año. Casillas por año indican porcentaje de datos sobre un total de 52 semanas y el total de lotes

Finca	#Lotes	#Medidas potencial	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Castilletes	35	1820	54.7	66.3	79.5	87.7	99.9	100.0	98.1	96.4	97.4	96.8	71.0
La Ceja	13	676	NA	NA	NA	NA	NA	NA	NA	NA	96.2	73.1	78.8
Colbanano	46	2392	NA	NA	NA	NA	NA	NA	NA	NA	NA	75.0	68.3
Coralina	11	572	NA	NA	NA	NA	NA	NA	NA	NA	80.8	79.5	NA
El Eden	16	832	NA	NA	NA	NA	NA	NA	NA	NA	96.2	78.8	78.8

Estadero	21	1092	NA	NA	NA	NA	NA	NA	NA	NA	98.1	100.0	91.6
Inagru	4	208	NA	NA	NA	NA	NA	NA	NA	NA	98.1	98.1	84.6
Leonor Emilia	7	364	NA	NA	NA	NA	NA	NA	NA	NA	98.1	98.1	98.1
Marandua	21	1092	NA	NA	NA	NA	NA	NA	NA	NA	NA	73.4	70.9
Pan Gordito	14	728	NA	NA	NA	NA	NA	NA	NA	NA	98.1	98.1	98.1
Paso Estoril	50	2600	NA	NA	NA	NA	NA	NA	NA	NA	74.1	84.0	82.7
Porvenir	15	780	NA	NA	NA	NA	NA	NA	NA	NA	98.1	84.1	NA
Puerto Alegre	23	1196	NA	NA	NA	NA	NA	NA	NA	NA	NA	83.4	NA
Yerbabuena	13	676	NA	NA	NA	NA	NA	NA	NA	NA	NA	97.9	73.1

Tal como para los datos de peso de racimo, la disponibilidad de datos fue dispareja, aunque en algunos casos hubo más años disponibles para el análisis. De nuevo, se plantea la realización de los análisis de acuerdo a los años disponibles por cada finca.

d. Datos de variedades (clones) utilizados

Se recibieron datos de las variedades utilizadas en cada lote de cada una de las 14 fincas bajo análisis. La mayoría de estos datos brindaron información suficiente para el análisis. Sin embargo, para las fincas Colbanano, Marandua, y algunos lotes de la finca Paso Estoril la información no fue recibida. Esto reduce sustancialmente la disponibilidad de los datos, y por tanto la cantidad de resultados. Se destaca la importancia de tener la información de variedades utilizadas en cada lote de cada finca, de una manera precisa, como base para la realización de cualquier tipo de análisis.

3. Uniformización y pre-procesamiento de datos de entrada

Debido a los vacíos en la información recibida, y a las diferencias en escala (productividad a nivel de fincas y demás variables a nivel de lotes), debieron aplicarse algunos procedimientos estadísticos con el fin de completar vacíos existentes en la información recibida y de realizar un *downscaling* en la información de productividad hasta el nivel de lote. Esto involucró dos procedimientos básicos: (a) análisis de mundo pequeño (SWA) para completar la información y (b) regresión de tipo aprendizaje *feed-forward backpropagation* para el *downscaling*.

a. Análisis de mundo pequeño (SWA) para llenado de vacíos en la información

Se usó el análisis de mundo pequeño (*Small World Analysis, SWA*, en inglés) (Watts & Strogatz, 1998) basándonos en la premisa que los datos de producción recibida reflejan una red de mundo pequeño. El análisis de mundo pequeño se fundamenta en regresiones lineales entre conjuntos de datos de diferentes lugares, o de diferentes épocas (nodos de la red de mundo pequeño), pero de la misma variable, así por ejemplo si para un tiempo *n* no existe un dato, pero para un tiempo *n+1* el dato sí existe, la relación entre el conjunto

de datos *n* y *n+1* permitirá llenar cualquier información faltante. Gráficamente y aplicándolo a los datos de producción de UNIBAN CI, si se tiene una matriz de *N* semanas del año por *M* lotes de la siguiente manera:

	S1	**S2**	**S3**	**…**	**Sn**
L1	X11	X12	X13	…	X1n
L2	X21	X22	X23	…	X2n
L3	X31	X32	X33	…	X3n
L4	X41	X42	X43	…	X4n
L5	X51	X52	X53	…	X5n
…	…	…	…	…	…
Lm	Xm1	Xm2	Xm3	…	Xmn

Y se encuentran relaciones lineales fuertes y estadísticamente significativas bien sea entre las filas (lotes) o las columnas (semanas), cualquier elemento faltante de la matriz que se encuentre en una columna y/o fila y no en el otro, puede ser completado.

Siendo así, lo primero que se debe realizar para la aplicación del SWA es una limpieza profunda de los datos para que las regresiones (relaciones) entre filas y/o columnas no se vean afectadas, posteriormente se realiza un inventario de la información actual, luego se detectan las relaciones entre columnas y/o filas. Se escoge, entonces, la columna/fila con mayor cantidad de datos y a partir de esta se empiezan a llenar los datos de las demás. Si alguna de las columnas/filas tiene datos que otras no tienen, se usa para también llenar datos, teniendo cuidado de no sobre-estimar usando datos ya llenados para llenar más. Como base, por tanto, deben usarse siempre datos reales.

Se usaron correlaciones estadísticamente significativas ($p<0.001$) y fuertes ($R^2 > 0.65$) en todos los casos para las variables peso de racimo, embolse y productividad tanto a nivel de finca como a nivel de lote cuando los datos estuvieron disponibles. Aunque los resultados de la aplicación del método son meras estimaciones y podrían generar un nivel de incertidumbre en los datos, el método es útil cuando el objetivo, más que obtener un dato "real" de campo, es comparar conjuntos de datos (épocas, fincas, zonas agroecológicas, entre otros).

Tabla 4 Resultados de la aplicación del análisis de mundo pequeño sobre los datos de UNIBAN CI para 3 variables y 2 niveles de producción (fincas y lotes)

Variable	Unidades	Nivel	Total	Faltante	Porcentaje faltante (i)	Completado SWA	Porcentaje completado	Porcentaje faltante (f)
Embolse	Rac/Ha	Lotes	58671	8103	13.8	6782	83.7	2.3
Embolse	Rac/Ha	Fincas	2862	356	12.4	306	86.0	1.7
Peso racimo	kg/rac	Lotes	49396	9572	19.4	9360	97.8	0.4
Peso racimo	kg/rac	Fincas	2226	294	13.2	289	98.3	0.2
Productividad	cajas/ha/semana	Fincas	1260	551	43.7	542	98.4	0.7

La disponibilidad inicial de la información fue considerablemente variable (tabla 4), y en algunos casos (i.e. peso de racimo, productividad), el porcentaje faltante de los datos fue bastante alto (19.4% y 43.7% respectivamente). A través del AMP, se logró completar esta información hasta diferentes niveles. En algunos casos debieron usarse varias semanas base (embolse, productividad), pero siempre teniendo en cuenta que no debe usarse información ya estimada para volver a estimar. En algunos casos las correlaciones estuvieron debajo de 0.65, pero arriba de 0.6, pero esa reducción se debió fundamentalmente a la aparición de datos extremos que no representaban la tendencia general en los datos, y por consiguiente fueron usadas estas regresiones también. Se completó información entre 80 y 98%, lo que brindó una mejor base de datos para el análisis de productividad posterior.

b. Regresión *feed-forward backpropagation neural network* con optimización Levenberg-Marquardt para datos de productividad

Debido a la diferencia en escalas entre la información de productividad (fincas) y la información de peso de racimo y embolse (lotes), y dado que las unidades de manejo actual son los lotes, y que idealmente la unidad de análisis para el presente estudio debería ser el lote-zona agroecológica, es crítico tener la información de productividad descrita en (1a) también a nivel de lote, para permitir un análisis apropiado de producción. A través de una regresión lineal entre el peso de racimo y la productividad, se encontró una tendencia en los datos (figura 1), con lo que se concluyó que existe una relación entre las dos variables.

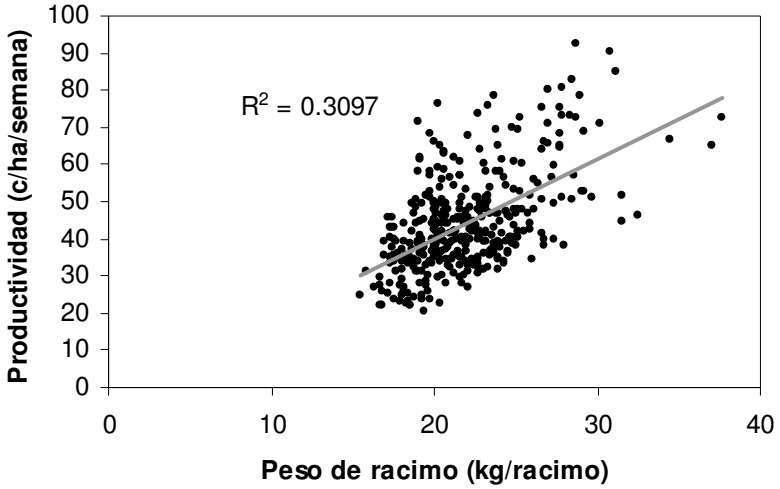

Figura 1 Relación entre productividad y peso de racimo. El coeficiente de determinación R^2 es el resultado de un ajuste lineal.

Después de explorar algunos algoritmos, se decidió usar un algoritmo de aprendizaje llamado *feed-forward backpropagation neural network*, con el peso de racimo como variable independiente y la productividad como variable dependiente. Los datos de ajuste

(entrenamiento) del algoritmo fueron los de nivel de fincas, y los datos de proyección fueron los de nivel de lotes. El algoritmo se ejecutó usando el paquete "AMORE" de R, disponible en http://www.r-project.org/, con ajuste de mínimos cuadrados de Levenberg-Marquardt (Marquardt, 1963) 1000 iteraciones, 3 capas, y 400x378x1 pesos. El ajuste rindió resultados bastante precisos tanto a nivel de fincas (figura 2) como de lotes, aunque en algunas ocasiones el patrón de aprendizaje no coincidió con los datos de proyección y por tanto no se logró un resultado (menos de 1% de los casos usando el total de datos a nivel de lotes)

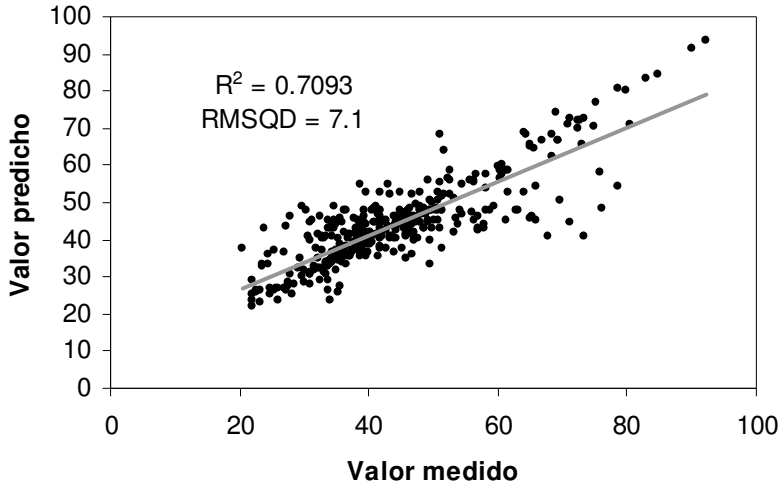

Figura 2 Resultados del proceso de *downscaling* sobre la información de productividad a nivel de fincas.

Aunque el algoritmo tiende a sub-estimar el valor medido, el resultado de la aplicación de este algoritmo permite derivar información de productividad a nivel de lote a partir de información a nivel de finca. Para evitar esta estimación, es recomendable que se colecten los datos de campo necesarios para tener datos de productividad, en otras palabras, que se determine el número de cajas exportadas a nivel de lote directamente con información de campo. El actual resultado brinda una estimación, pero también agrega algo de incertidumbre pues la regresión usada logra explicar el 85% de la varianza de los datos, pero no el 100%.

c. Matrices de datos

Se generaron matrices de N lotes por M semanas para las variables peso de racimo, embolse, y productividad. Los datos de peso de racimo y productividad se encuentran en el archivo "./isoproductividad/datos-entrada/corregido/Productividad-PesoRacimo-Fincas-Semanal.xls", mientras que los de embolse se encuentran en "./isoproductividad/datos-entrada/corregido/Embolse-Lotes-Semanal.xls". A partir de estos datos se realizaron todos los análisis posteriores.

4. Análisis de productividad histórico y comparativo

El análisis de productividad histórico y comparativo se fundamenta en gráficos que muestran el comportamiento de las variables importantes (peso de racimo, productividad) a través del tiempo, y como promedios de los años disponibles. Esta información esta disponible en "./isoproductividad/resultados/Resultados.xls".

a. Por finca y lote

Con el objetivo de observar el comportamiento de la producción en las fincas bananeras, se realizaron gráficos históricos que muestran la variabilidad dentro de la finca y los cambios a través del tiempo. Estos gráficos pueden usarse para comparar la producción de una finca y otra (figuras 3a y 3b). En este caso, se observa que la finca con mejor productividad es Paso Estoril, y la de productividad mas baja es Inagru, aunque en algunos periodos las diferencias son relativamente pequeñas, especialmente a finales del año 2009.

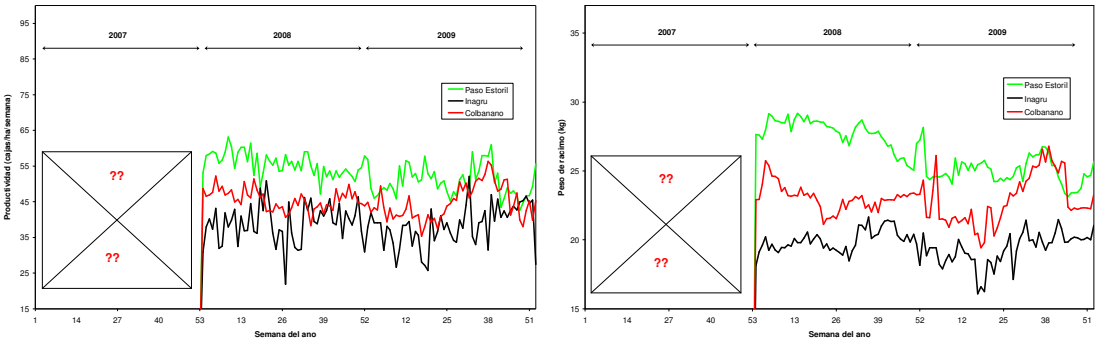

Figura 3 (a) Productividad histórica y (b) peso de racimo histórico para 3 fincas (Paso Estoril, Inagru, y Colbanano)

Este tipo de gráficos ayudan a visualizar los cambios a través del tiempo para tomar decisiones respecto a la producción (renovación de lotes, cambios en manejo, entre otros). De otro lado, ayudan a conocer el comportamiento histórico de las diferentes fincas para detectar altibajos en producción y tomar correctivos rápidamente. Permiten, además, observar tendencias en las fincas, en este caso, por ejemplo, la producción de la finca Paso Estoril decae a través del tiempo, mientras que las otras dos fincas muestran incluso un ligero aumento en 2009 respecto de los dos años anteriores. Otro tipo de gráfico que es de suma utilidad es observar la variabilidad dentro de la finca (figura 4)

Las líneas rojas (figura 4) muestran el rango completo de variabilidad al interior de la finca (entre los diferentes lotes), mientras que la línea gruesa indica la media de todos los lotes. Claramente se observa que hay una gran variabilidad en cierto periodo (finales de 2007, 2008 y 2009), mientras que durante casi todo 2010, hay poca variabilidad. El objetivo de un mejor control de la producción es que los lotes tengan no sólo una

producción óptima, sino que haya pocas diferencias en producción entre ellos. Una producción uniforme indica que las tecnologías están siendo usadas apropiadamente y que se está obteniendo el máximo de producción de todos los lotes.

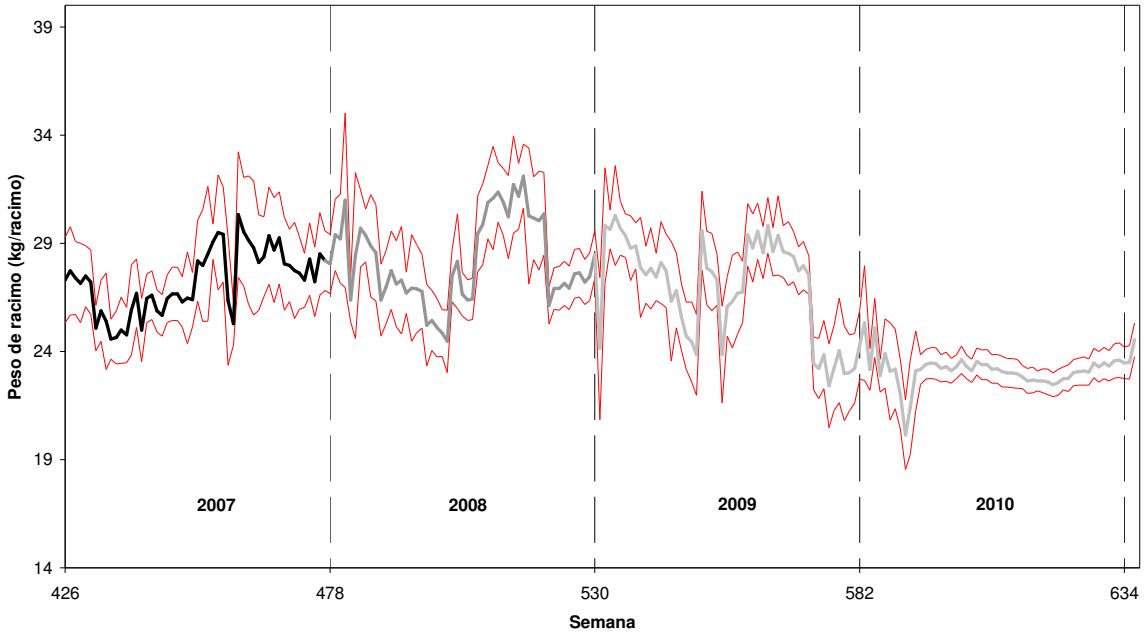

Figura 4 Peso de racimo histórico para la finca Coralina. La línea gruesa indica el promedio de todos los lotes, mientras que las líneas delgadas rojas corresponden a un intervalo de confianza del 95% alrededor de la media de todos los lotes.

De esta manera, muy sencillamente, con solamente tener los datos organizados adecuadamente, es fácil derivar varios gráficos que permitan analizar la producción de una manera eficiente y tomar conclusiones rápidas sobre la misma. Adicionalmente también pueden graficarse diferentes lotes (figura 5)

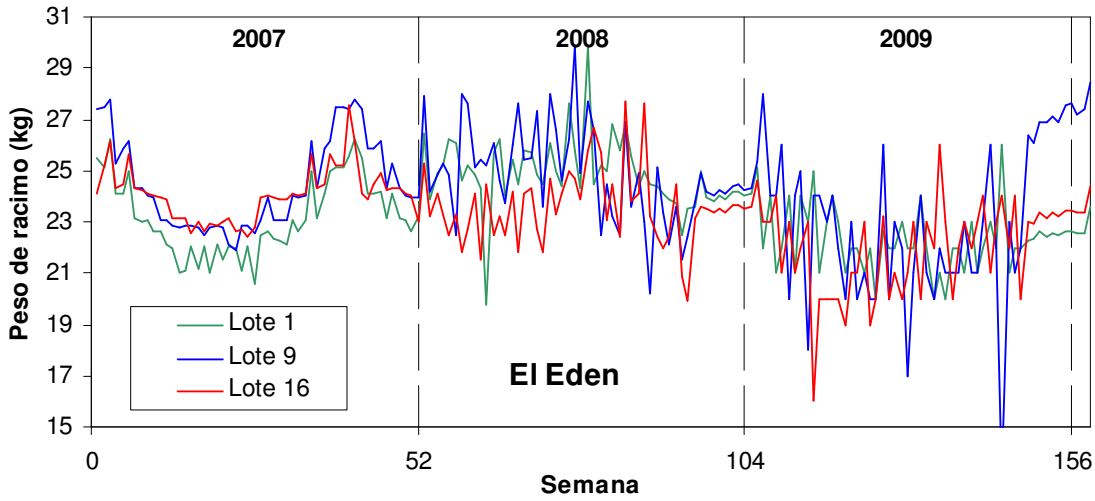

Figura 5 Peso de racimo histórico para la finca tres lotes de la finca El Eden

De la misma manera que con los gráficos anteriores, este gráfico ayuda a comparar el comportamiento de los diferentes lotes, y detectar problemas en producción en uno o en otro lote. Es de utilidad cuando la producción de una finca determinada decrece, y se requiere detectar si es una tendencia general en todos sus lotes, o sólo en algunos de ellos.

b. Comparación de productividad media histórica de fincas

Usando los promedios y desviaciones estándar sobre los años disponibles, pueden realizarse gráficos que permitan comparar la productividad media de las diferentes fincas y con eso observar las ventajas y desventajas competitivas de dichas unidades. En este caso, se observa, en primer lugar, que las fincas Paso Estoril, Coralina y Puerto Alegre, presentan una productividad media alta y un peso de racimo medio alto en comparación a las demás fincas, mientras que la finca Yerbabuena, presenta considerables deficiencias en ambos índices de producción, lo que indica algunos problemas de desempeño. Las líneas de variabilidad, sin embargo, muestran que en algunas ocasiones (algunos lotes, algunos años), la producción de las fincas de alta producción llegó a decrecer hasta más del 50%. Un aspecto importante a mencionar es que a medida que los valores de los índices de producción analizados son mayores, las varianzas son mayores.

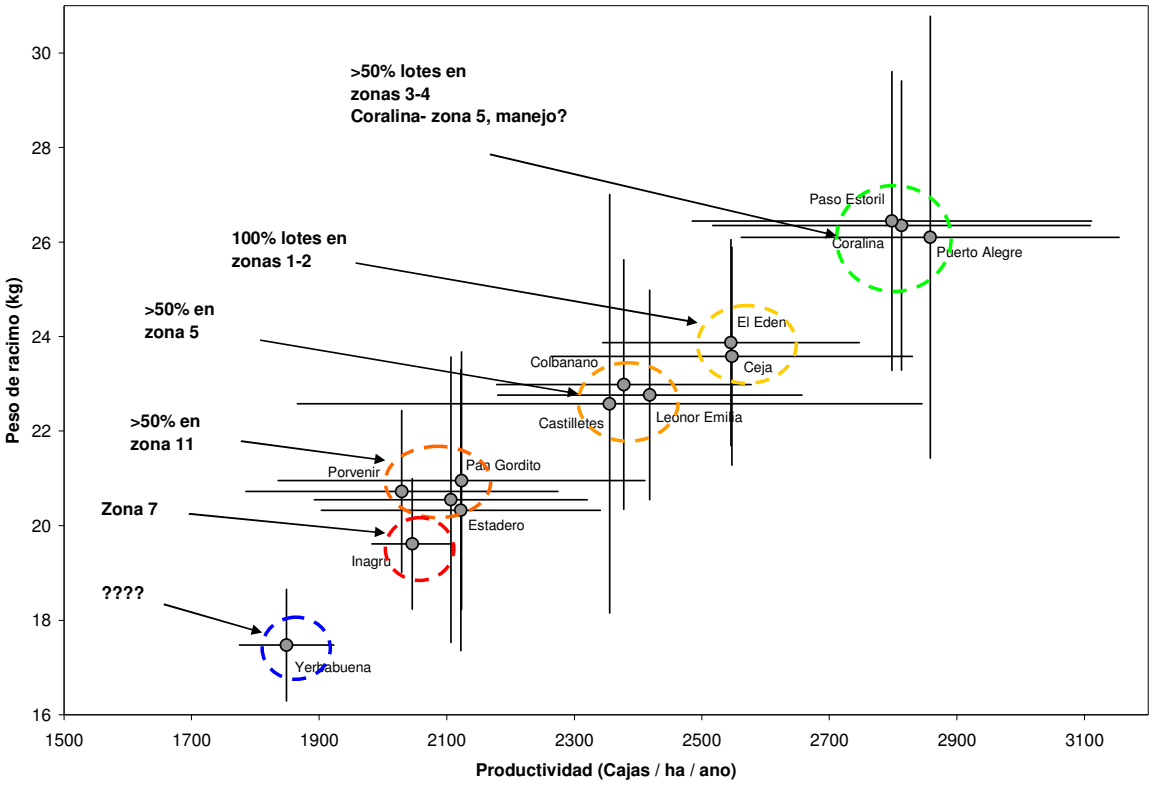

Figura 4 Grafico X-Y entre peso del racimo y productividad (cajas/ha/año) para las 14 fincas. Los puntos simbolizan el promedio y las líneas la desviación estándar alrededor de la media de lotes y años.

Con este tipo de gráfico, aunque es más difícil tomar una decisión que influya en la producción inmediatamente, puede analizarse el comportamiento de largo plazo de un conjunto de fincas, con miras a la priorización de problemas. Por ejemplo, la observación detallada y priorización de fincas con muchos problemas en su producción, antes que otras con pocos o sin problemas.

5. Análisis de iso-productividad

Un tipo de análisis comparativo que permite analizar tres variables de importancia al mismo tiempo es el de curvas de Iso-Productividad, desarrollado por Cenicaña, hace varios años, y con el que ellos han mejorado su producción, mediante el ajuste y enfoque de tecnologías en sitios en donde se comportan mejor. Aquí presentamos el método de generación de curvas de iso-productividad, y un *script* en R para tal fin.

a. Metodología de análisis

El método básicamente toma las dos variables de interés y las grafica en tres ejes: eje X, eje Y, y un eje transformado Y-2. En este caso se trabajó con el peso de racimo, el número de racimos embolsados, y la productividad. Las curvas de iso-productiviad tienen la particularidad de ser aplicables a cualquier par o tripleta de variables y cualquier grupo de datos que se deseen comparar (i.e. variedades, clones, fincas, entre otros). El procedimiento detallado es el que sigue:

1. Organizar los datos en una tabla con los siguientes campos:

Clase	V1	V2	V3(V1/V2)

En este caso, “Clase” sería la variable tipo “factor” base para comparación, en este caso serían las variedades, fincas, o lotes. V1 sería la primera variable medida, V2 es la segunda variable medida, y V3 es la razón (división) de las dos variables anteriores (V1 y V2).

2. Definir qué variables corresponden a cada eje, de la siguiente manera
   - V3 siempre debe corresponder al eje X
   - V2 (denominador) siempre debe ser el eje Y
   - V1 (numerador) siempre debe ser el eje Y-2

3. Dibujar los puntos X vs. Y en un gráfico normal X-Y, usando Excel, R, o cualquier paquete estadístico, u hoja de cálculo

4. Calcular rangos de variación en cada variable (máximo y mínimo)

5. A través de un proceso iterativo, dibujar las curvas de iso-productividad, manteniendo la variable V1 constante y variando las otras dos, se recomienda empezar desde el mínimo valor de V1. Supongamos los siguientes rangos de variación en nuestras variables objetivo (embolse, productividad y su división [productividad/embolse]):

Embolse: de 1000 a 3000 racimos/ha/año
Productividad: de 1500 a 4500 cajas/ha/año
Ratio: de 0.4 a 1.4 cajas/racimo

En este caso, el embolse corresponde al eje Y, la productividad al eje Y-2 y el ratio al eje X.

La primera curva se realiza para el valor mínimo de productividad: 1500 cajas/ha/año. Se define un intervalo de variación para el embolse, digamos, 100 unidades y se hace una tabla. Esto daría un total de 20 diferentes valores. Para cada valor de embolse (de los 20 totales), usando el valor de 1500 cajas/ha/año, se calcula el ratio como:

$$Ratio = \frac{Embolse}{1500}$$

Con los valores de embolse (Y) y ratio (X) se grafica la curva y se asigna el valor en la leyenda, o al final de la curva. La asignación del valor depende del software utilizado para graficarla. Así se obtiene la primera curva, luego se sigue el mismo proceso para la segunda. Si se define que la siguiente curva es, digamos 1700 cajas/ha/año, se realiza de nuevo el proceso, y se grafica la curva.

Basados en este método, hemos desarrollado 3 scripts de R que realizan este tipo de gráficos rápidamente y los almacenan en un PDF. Todos disponibles en el folder "./isoproductividad/_scripts":

"*plottingData-farmLevel.R*": Realiza tres tipos de gráficos para cada finca de las que se encuentran disponibles. Ejecutando, en la consola de R, el comando:

```
source("plottingData-farmLevel.R")
```

Se listarán las fincas (de 1 a N fincas en pantalla) y se harán disponibles tres funciones:

`PREIsoProd(Finca,Pasos)`: Realiza grafico de iso-productividad para productividad (Y-2) y Racimos Embolsados (Y). Los argumentos son "**`Finca`**", que corresponde al número (no el nombre) de la finca listado después de ejecutar el comando "**`source`**", y "**`Pasos`**", que corresponde al número de pasos para la variable Y-2, normalmente se

coloca un valor entre 30 y 50, mayor el valor, mayor número de isolíneas en el gráfico (figura 5).

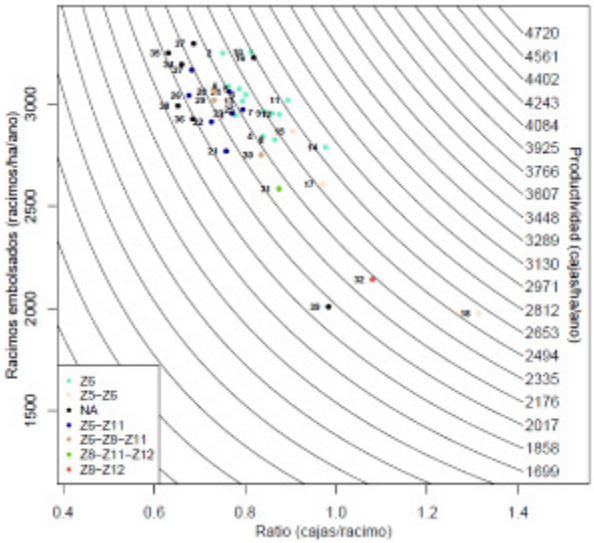

Figura 5 Grafico de isoproductividad para la finca Castilletes, comparando los diferentes lotes y su ubicación en zonas agroecológicas para las variables racimos embolsados y productividad

`PRPIsoProd(Finca,Pasos)`: Realiza grafico de iso-productividad para peso de racimo (Y) y Productividad (Y-2). Los argumentos son "**`Finca`**", que corresponde al número (no el nombre) de la finca listado después de ejecutar el comando "**`source`**", y "**`Pasos`**", que corresponde al número de pasos para la variable Y-2, normalmente se coloca un valor entre 30 y 50, mayor el valor, mayor número de isolíneas en el gráfico (figura 6).

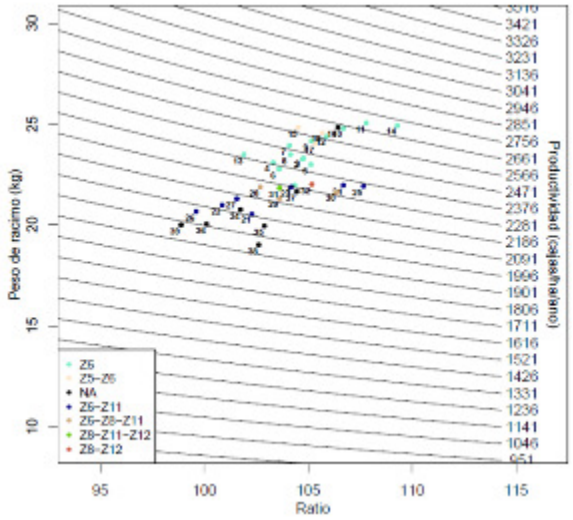

Figura 6 Grafico de isoproductividad para la finca Castilletes, comparando los diferentes lotes y su ubicación en zonas agroecológicas, para las variables peso de racimo y productividad

`PRREIsoProd(Finca,Pasos)`: Realiza grafico de iso-productividad para peso de racimo (Y) y racimos embolsados (Y-2). Los argumentos son "**`Finca`**", que corresponde al

número (no el nombre) de la finca listado después de ejecutar el comando "**source**", y "**Pasos**", que corresponde al número de pasos para la variable Y-2, normalmente se coloca un valor entre 30 y 50, mayor el valor, mayor número de isolíneas en el gráfico (figura 7).

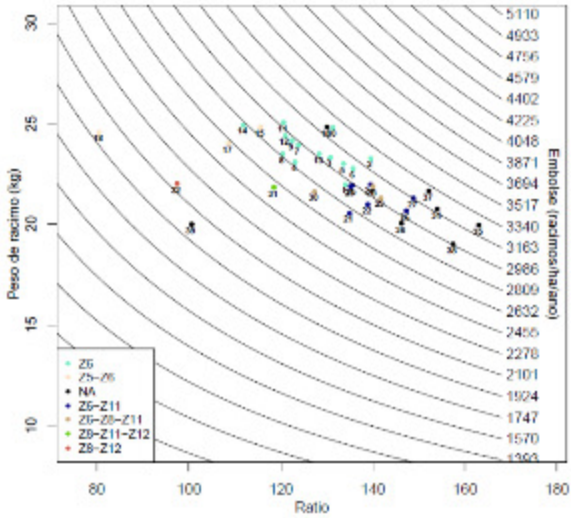

Figura 7 Grafico de isoproductividad para la finca Castilletes, comparando los diferentes lotes y su ubicación en zonas agroecológicas, para las variables peso de racimo y racimos embolsados

"*plottingData-zoneLevel.R*": Realiza tres tipos de gráficos para cada zona agroecológica de las que se encuentran disponibles. Ejecutando, en la consola de R, el comando:

```
source("plottingData-zoneLevel.R")
```

Se listarán las zonas (de 1 a N zonas en pantalla) y se harán disponibles tres funciones:

PREIsoProd(Zona,Pasos): Realiza grafico de iso-productividad para productividad (Y-2) y Racimos Embolsados (Y). Los argumentos son "**Zona**", que corresponde al número (no el nombre) de la finca listado después de ejecutar el comando "**source**", y "**Pasos**", que corresponde al número de pasos para la variable Y-2, normalmente se coloca un valor entre 30 y 50, mayor el valor, mayor número de isolíneas en el gráfico (figura 8).

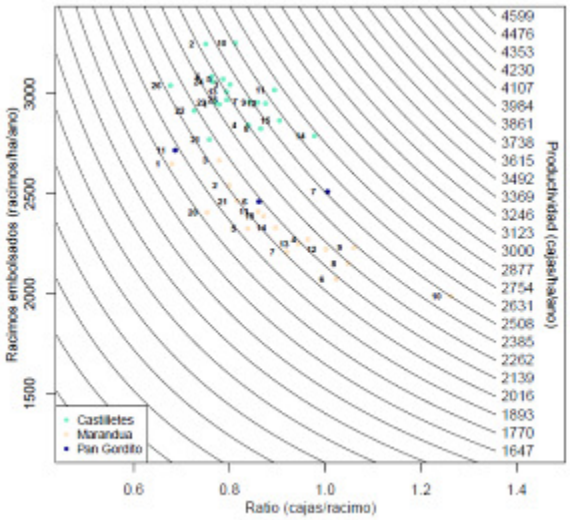

Figura 8 Grafico de isoproductividad para la zona agroecológica 6, comparando los diferentes lotes y fincas, para las variables racimos embolsados y productividad

`PRPIsoProd(Zona,Pasos)`: Realiza grafico de iso-productividad para peso de racimo (Y) y Productividad (Y-2). Los argumentos son "**`Zona`**", que corresponde al número (no el nombre) de la finca listado después de ejecutar el comando "**`source`**", y "**`Pasos`**", que corresponde al número de pasos para la variable Y-2, normalmente se coloca un valor entre 30 y 50, mayor el valor, mayor número de isolíneas en el gráfico (figura 9).

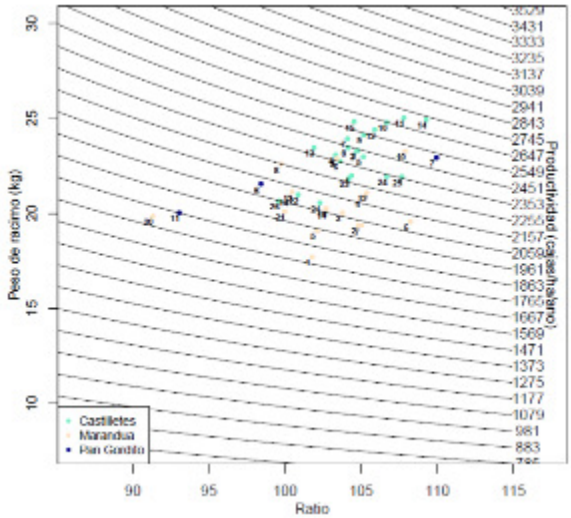

Figura 9 Grafico de isoproductividad para la zona agroecológica 6, comparando los diferentes lotes y fincas, para las variables peso de racimo y productividad

`PRREIsoProd(Zona,Pasos)`: Realiza grafico de iso-productividad para peso de racimo (Y) y racimos embolsados (Y-2). Los argumentos son "**`Zona`**", que corresponde al número (no el nombre) de la finca listado después de ejecutar el comando "**`source`**", y "**`Pasos`**", que corresponde al número de pasos para la variable Y-2, normalmente se

coloca un valor entre 30 y 50, mayor el valor, mayor número de isolíneas en el gráfico (figura 10).

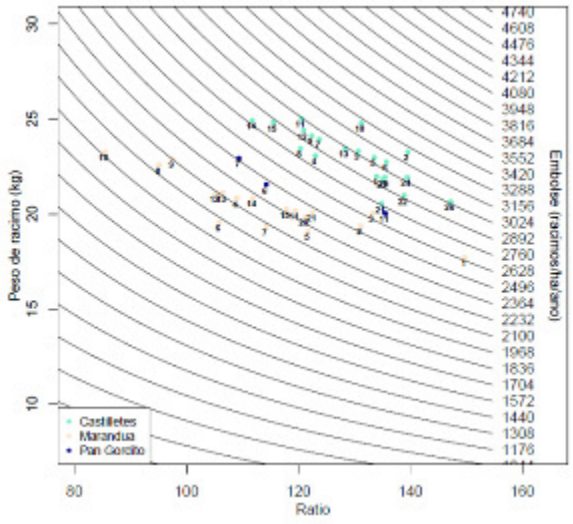

Figura 10 Grafico de isoproductividad para la zona agroecológica 6, comparando los diferentes lotes y fincas, para las variables peso de racimo y racimos embolsados

Un punto importante a considerar el gráfico por zonas es que algunos lotes pertenecen a más de una zona agroecológica, esto se debe a que la variabilidad en algunas de las variables (particularmente las de NDVI) es considerable al interior de ellos. Se entiende que estas variaciones son posibles y esto dificulta el procedimiento de análisis (separación de zonas). Por este motivo se han creado 3 campos de "pertenencia" a zona agroecológica: Zona, Zona50, y Zona70. Zona corresponde a la zona o zonas a las que el lote en particular corresponde, Zona50 hace referencia a la zona que ocupa 50% o más del área del lote y Zona70 hace referencia a la zona que ocupa 70% o más del lote. Esto se hace basados en la premisa de que las zonas afectan el comportamiento de la producción en cierto nivel, pero que si una zona ocupa más del 70% o incluso 50% de un lote, puede influenciar de manera tal el comportamiento que las variables medidas sean casi producto de esta zona en particular y no de las otras que también ocupan el lote. Se recomienda a UNIBAN que la información sea colectada de manera diferencial Lote-Zona para evitar este proceso, o asunción.

"*plottingData-cloneLevel.R*": Realiza tres tipos de gráficos para cada zona agroecológica de las que se encuentran disponibles como promedio de los clones que se encuentran en esa zona. Ejecutando, en la consola de R, el comando:

```
source("plottingData-cloneLevel.R")
```

Se listarán las zonas (de 1 a N zonas en pantalla) y se harán disponibles tres funciones:

`PREIsoProd(Zona,Pasos)`: Realiza grafico de iso-productividad para productividad (Y-2) y Racimos Embolsados (Y). Los argumentos son "**`Zona`**", que corresponde al número (no el nombre) de la finca listado después de ejecutar el comando "**`source`**", y "**`Pasos`**", que corresponde al número de pasos para la variable Y-2, normalmente se coloca un valor entre 30 y 50, mayor el valor, mayor número de isolíneas en el gráfico (figura 11).

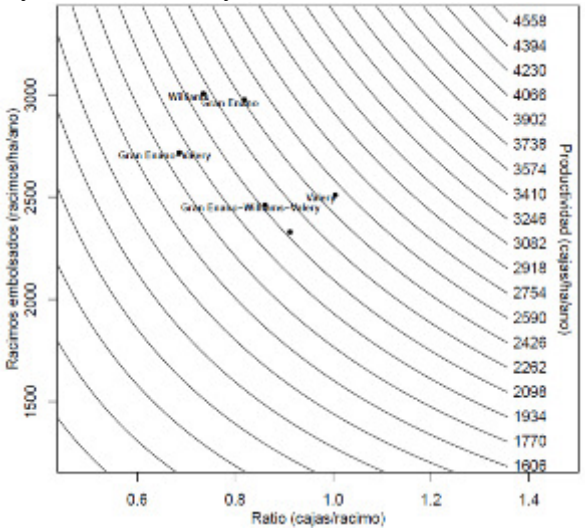

Figura 11 Grafico de isoproductividad para la zona agroecológica 6, comparando los diferentes lotes y fincas, para las variables racimos embolsados y productividad

`PRPIsoProd(Zona,Pasos)`: Realiza grafico de iso-productividad para peso de racimo (Y) y Productividad (Y-2). Los argumentos son "**`Zona`**", que corresponde al número (no el nombre) de la finca listado después de ejecutar el comando "**`source`**", y "**`Pasos`**", que corresponde al número de pasos para la variable Y-2, normalmente se coloca un valor entre 30 y 50, mayor el valor, mayor número de isolíneas en el gráfico (figura 12).

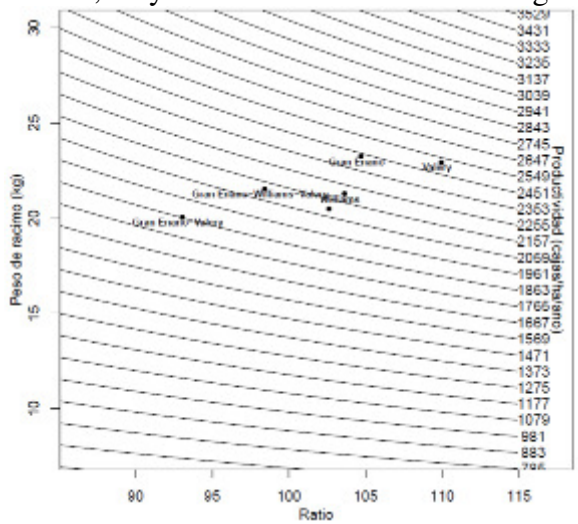

Figura 12 Grafico de isoproductividad para la zona agroecológica 6, comparando los diferentes lotes y fincas, para las variables peso de racimo y productividad

`PRREIsoProd(Zona,Pasos)`: Realiza grafico de iso-productividad para peso de racimo (Y) y racimos embolsados (Y-2). Los argumentos son "**`Zona`**", que corresponde al número (no el nombre) de la finca listado después de ejecutar el comando "**`source`**", y "**`Pasos`**", que corresponde al número de pasos para la variable Y-2, normalmente se coloca un valor entre 30 y 50, mayor el valor, mayor número de isolíneas en el gráfico (figura 13).

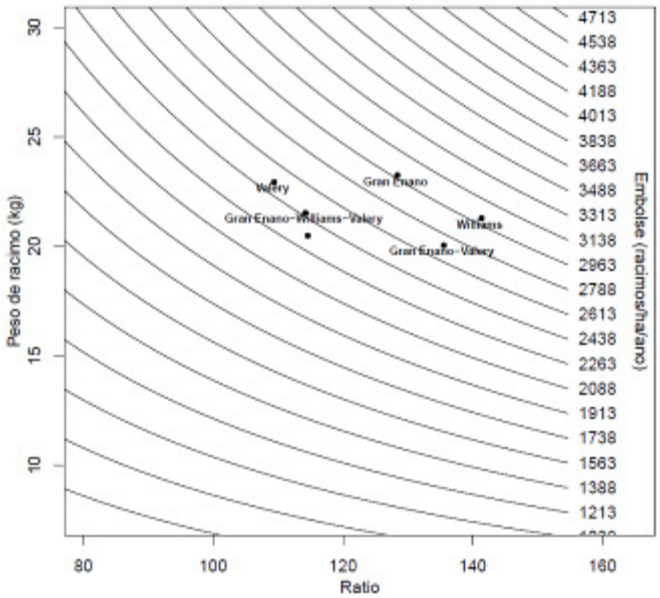

Figura 13 Grafico de isoproductividad para la zona agroecológica 6, comparando los diferentes lotes y fincas, para las variables peso de racimo y embolse

De la misma manera se establece un nivel de dominancia de al menos 50% para que un lote sea considerado dentro de una zona. De lo contrario, la variación es considerable y el agrupamiento en un solo gráfico se hace complicado, pues termina quedando el gráfico con un solo punto, o dos, lo que dificulta las comparaciones.

b. Resultados principales

Como puede observarse en las figuras 5 a la 13, el comportamiento de los clones o combinaciones de clones varía de acuerdo a la zona agroecológica en que se encuentran, y de la misma manera, bien sea porque el clon está mejor adaptado a ciertas condiciones que a otras, o porque las tecnologías utilizadas en promedio sobre ese clon lo hacen responder mejor, el clon termina teniendo un desempeño sustancialmente mejor.

A partir de esto, por ejemplo, en la zona agroecológica 6, la variedad Valery, o la combinación entre Gran Enano y Valery en un lote, brinda un desempeño bastante bueno. Un número considerable de racimos embolsados anualmente, y un alto número de cajas exportadas al año, con una excelente relación entre racimos embolsados y número de cajas exportadas.

6. Conclusiones

En resumen, se ha desarrollado una metodología que permite comparar la producción de fincas, lotes, y variedades con base en la zonificación agroecológica desarrollada con anticipación. Se implementó el gráfico en varios scripts de R, que pueden ser analizados y re estructurados o traducidos a cualquier otro lenguaje que se desee. La metodología es sencilla, aunque se destaca que el punto crítico reside en la uniformización y confiabilidad de la información de entrada, que en la medida de lo posible debe ser a nivel de lote, e idealmente debe ser a nivel de lote-zona agroecológica, de manera tal que todos los análisis puedan efectuarse de la mejor manera

Referencias

Marquardt D (1963) An algorithm for least squares estimation of non-linear parameters. *J. Soc. Ind. Appl. Math.* 431-441.

Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393: 440-442

Parte 3. Estudio exploratorio para pronóstico bio-climático del ataque de Sigatoka Negra (*Mycosphaerella fijiensis* M.) en plantaciones bananeras de Urabá, y su posible efecto en la producción

Resumen

El presente documento describe los materiales usados, métodos aplicados y resultados obtenidos por el Centro Internacional de Agricultura Tropical (CIAT) respecto al estudio exploratorio para el pronóstico bio-climático del ataque de Sigatoka negra y sus posibles efectos en producción. El análisis consistió de 7 pasos básicos: (1) Organización de la información, (2) selección del lote y finca de análisis, (3) transformación matemática de las variables independientes, (4) aplicación de desfase a las variables independientes, (5) exploración de correlaciones entre los datos, (6) desarrollo de modelos de predicción usando regresiones multivariadas *stepwise* y sub-muestreo o *bootstrapping* de los datos, (7) análisis de grado de afección de la producción por causa de la enfermedad. Se analizaron variables de ataque de la enfermedad para el lote 1 de la finca Castilletes, ubicado en la zona agroecológica 6. Se encontró que el ataque de la enfermedad está altamente relacionado con la precipitación de 6 a 8 semanas anteriores, con la humedad relativa actual y de 1 a 2 semanas anteriores, con la temperatura media de 2 semanas anteriores, y el número de horas con humedad relativa por encima del 90%. Para cada una de 4 variables dependientes (número de hojas totales, hoja más joven infectada, hoja más joven manchada, y severidad) se encontró un modelo lineal multivariado para predecir el ataque futuro de la enfermedad con 1 a 8 semanas de adelanto en el tiempo (dependiendo de la variable). Estos modelos mostraron ser consistentes en el tiempo y ser capaces de predecir la enfermedad con hasta 80% de confiabilidad, e incluyeron entre 10 y 11 variables, con sus respectivos desfases. Hacia el futuro, se sugiere una validación extensiva de los modelos presentados aquí en otras fincas, lotes, y zonas agroecológicas, así como un mejoramiento de los modelos mismos usando un set de datos históricos mucho más completo.

Contenido

1. Introducción

En vista de la necesidad de un monitoreo adecuado de la producción, y de la alta prevalencia del hongo *Mycosphaerella fijiensis* Morelet (Sigatoka negra) en regiones de alta precipitación y humedad relativa tal como la región de Urabá, analizar el comportamiento histórico de la enfermedad se hace crítico, y aún más, relacionarlo con el comportamiento histórico de variables ambientales que influencian su crecimiento, desarrollo y dispersión. A priori, el conocimiento sobre Sigatoka negra, indica que las variables que más influencian su comportamiento son la humedad relativa, la temperatura al interior del cultivo, y la precipitación (Pérez-Vicente et al. 2000). Sin embargo, las relaciones entre estas variables y el desarrollo de la enfermedad no son en todas las ocasiones lineales ni en el mismo instante en el tiempo (Pérez-Vicente et al. 2000). En otras palabras, el comportamiento de la enfermedad en un instante determinado puede ser resultado de eventos climáticos de una o más semanas anteriores. Estas relaciones, sin embargo, tienden a mantenerse en el tiempo dado que se derivan del comportamiento fisiológico del hongo causante de la enfermedad.

En el presente estudio, se analizó la prevalencia y el comportamiento histórico de la Sigatoka negra en un lote representativo sobre una serie de lotes en la zona bananera de Urabá. Las variables con las que se midió el ataque de la enfermedad fueron: Número de hojas totales, hoja más joven infectada, hoja más joven manchada, y severidad. La escogencia del lote se hizo en base a la disponibilidad de datos de la enfermedad, datos climáticos (cercanía a la estación meteorológica más cercana), datos de producción histórica (peso de racimo, racimos embolsados). Se analizó el comportamiento de la enfermedad en intervalos de 2 semanas, desde Enero de 2007 hasta Diciembre de 2009, con cuatro desfases temporales diferentes respecto a las variables climáticas.

2. Datos de entrada

Los datos de entrada para el análisis aplicado fueron básicamente datos históricos de ataque de la enfermedad (hojas totales, hoja más joven infectada, hoja más joven manchada, severidad), datos de producción (peso de racimo, racimos embolsados), datos climáticos de la estación meteorológica, datos de lluvias diarias del satélite TRMM, y datos de índice de vegetación normalizado del satélite Terra-MODIS.

a. Datos históricos de Sigatoka negra

Los datos de Sigatoka negra, como ya se mencionó, son datos de campo que miden la severidad del ataque de la enfermedad en ciertos lotes de algunas de las fincas de la región. La frecuencia de toma de estos datos es cada 2 semanas, y las variables son:

- Número total de hojas (TNL)
- Hoja más joven infectada (YLI)
- Hoja más joven manchada (YLS)
- Severidad (S)

Se encontraron en la base de datos, 725 fincas con al menos 1 dato, entre los años 1998, y 2010. Siendo 2009 el año con mayor número de fincas disponibles (525), y 1998 el año con el menor número de datos (4). El número de años con datos fue variable a través de las fincas (Figura 1). Muy pocas fincas presentaron datos durante todo el período 1998-2010, y lo más usual fue que las fincas presentaran sólo un año de datos (usualmente el 2009).

Un total de 255 fincas presentaron datos para el período 2007-2009, que es el período óptimo de análisis, dada la disponibilidad de datos climáticos.

Se encontraron en la base de datos, 27 fincas con al menos 1 dato, entre los años 1998, y 2010 para uno de sus lotes, ubicado dentro de la región de estudio (indicado por las fincas presentes en el shapefile recibido de UNIBAN CI). Siendo 2009 el año con mayor número de fincas disponibles, y 1998 el año con el menor número de datos. El número de

años con datos fue variable a través de las fincas (Figura 1). Muy pocas fincas presentaron datos durante todo el período 1998-2010, y lo más usual fue que las fincas presentaran sólo un año de datos (usualmente el 2009).

El mismo total de 27 fincas (83 lotes) presentó datos para el período 2007-2009, que es el período óptimo de análisis dada la disponibilidad de datos climáticos. Aunque el número de fincas presentes tanto en el shapefile como en la base de datos de severidad y ataque de Sigatoka negra fueron 27, el número total de fincas en la base de datos de Sigatoka negra fue de 725, de las que 255 presentaron datos entre 2007 y 2009. No obstante esto, las fincas que no aparecen en el shapefile no se usaron debido a que no se conoce su ubicación geográfica, y por tanto no se puede conocer cuál es la estación meteorológica más cercana, ni la zona agroecológica a la que pertenecen.

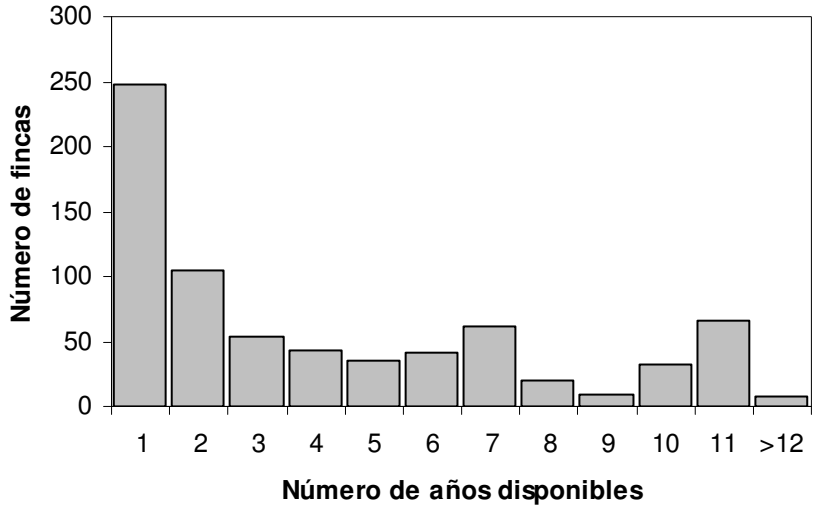

Figura 1 Histograma de número de años con datos disponibles por finca

En términos generales, aunque la disponibilidad de datos es limitada a unos cuantos años, es suficiente para realizar un análisis de variabilidad histórico. Además, la calidad de los mismos es adecuada para la realización de los análisis subsecuentes. Cabe anotar que el trabajo realizado por quienes se encargan de colectar esta información en el campo es notable y se sugiere que se siga realzando de esta manera, y que, aún más interesante, se expanda a todas las fincas de la región, y se aumente la frecuencia de las mediciones primero a una medición por semana, y luego a 2 o tres mediciones por semana. Esto ayudará a tener un sistema mucho más robusto de monitoreo.

b. Datos históricos climáticos (de estación meteorológica)

Los datos climáticos disponibles provinieron de siete estaciones meteorológicas que se encuentran ubicadas a través de la región. Se obtuvieron datos para el período 2007-2009 de las siguientes variables:

- Humedad relativa
- Temperatura máxima, mínima y media
- Precipitación
- Radiación solar
- Evapotranspiración

La frecuencia de toma de datos fue de 1 hora, pero dada la frecuencia de los datos de severidad de Sigatoka negra, se agregaron hasta diarios y luego cada 2 semanas. La calidad y abundancia de la información meteorológica es un punto fundamental a analizar antes de emprender cualquier análisis que involucre este tipo de información. En general, se encontró que ninguna de las estaciones contó con el 100% del total de datos. En el mejor de los casos (estación 1), las variables tuvieron aproximadamente el 90% de datos disponibles (Figura 2). Todas las variables presentaron niveles similares de disponibilidad de datos, excepto la evapotranspiración, que presentó muchos valores fuera del rango creíble (i.e. valores negativos). En algunos casos, los datos de precipitación y temperatura excedieron también los rangos creíbles (valores de más de 1000 mm en un solo día, y valores similares en temperatura).

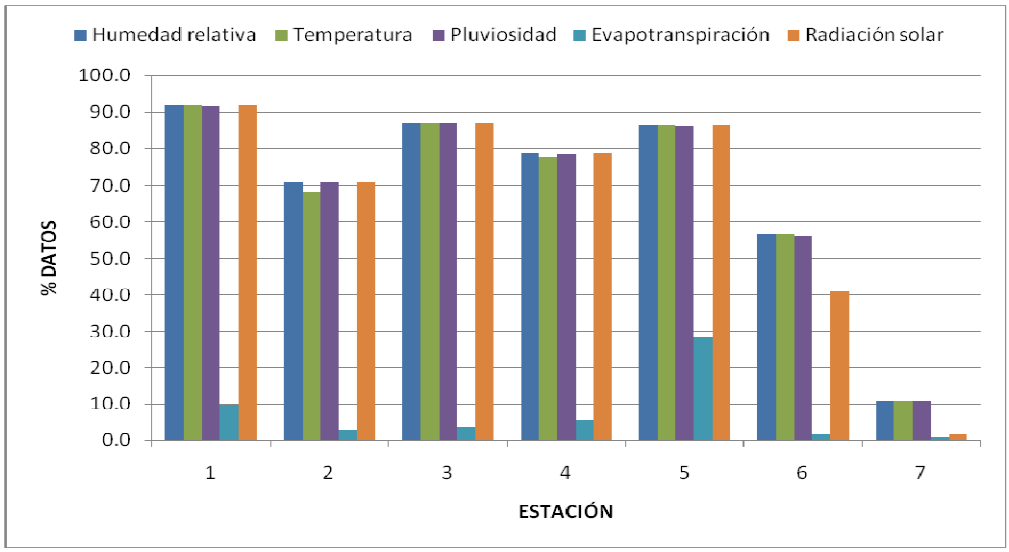

Figura 2 Porcentaje de datos por variable y estación meteorológica (sobre un total de 1096 potenciales).

Se recomienda, por tanto hacer una revisión extensiva y detallada a los datos provenientes de las estaciones y a las estaciones mismas y sus sensores, pues del mantenimiento de estos depende directamente la calidad de la información y su uso. No obstante lo anterior, la disponibilidad y calidad de algunas estaciones (estación 1, 3, 4, 5),

en variables como radiación solar, humedad relativa, temperatura y precipitación fue la suficiente como para la realización de los análisis que aquí se proponen. Los años a analizar son, por lo tanto entre 2007 y 2010.

c. Datos del satélite de lluvias TRMM

Se descargaron, procesaron y extrajeron datos del satélite de lluvias de la NASA TRMM (*Tropical Rainfall Measuring Mission*), producto 3b42, cuyos datos son horarios. Se agregaron los datos a nivel diario y se extrajeron para el área de estudio. La resolución espacial de los datos del satélite TRMM es de 15 arco-minutos (aproximadamente 28 km en el Ecuador). Aunque la resolución espacial del satélite es gruesa en comparación al tamaño del área de estudio, es una medida complementaria de alta resolución temporal que puede proveer información complementaria.

d. Datos de índice de vegetación (NDVI)

Los datos de NDVI se extrajeron a partir de imágenes del satélite TERRA MODIS, producto MOD13Q1. La resolución espacial de estos datos es de 7.5 arco-segundos (aproximadamente 250 metros en el Ecuador), y la frecuencia de los datos es de 16 días, esto significa que cada 16 días hay un nuevo dato de NDVI resultado de la medición del satélite, y que es publicado por la NASA, descargado y post-procesado por el CIAT. El período para el cual se extrajeron los datos de NDVI fue el período 2007-2009.

Se extrajeron datos para el período de análisis, para toda la región, con el objetivo de tener una serie de tiempo completa de datos que muestren el verdor de las plantaciones y que permitan relacionar dicho verdor con las variables de severidad de la enfermedad. Dado que los datos de NDVI tienen una periodicidad de 16 días, se ajustan fácilmente a los datos de severidad de la enfermedad, que son tomados cada 2 semanas.

e. Datos de producción

Se recibieron datos de peso de racimo (en kilogramos) a nivel de lotes, por semana, para 14 fincas: Castilletes, La Ceja, Colbanano, Coralina, El Edén, Estadero, Inagrú, Leonor Emilia, Marandua, Pan Gordito, Paso Estoril, Porvenir, Puerto Alegre, Yerbabuena. La disponibilidad de los datos también fue un factor importante en el peso de racimo. En general, las 14 fincas mostraron un número de años con datos muy variable entre ellas. En solo algunos casos el total de medidas por año y por finca fue del 100% (El Edén año 2007, entre otros), mientras que en otros hubo una evidente carencia de información, y el número de datos disponibles llegó sólo al 50% (Castilletes 1999, Pan Gordito 2008, entre otros). Los años para los que la mayoría de las fincas tuvo al menos 50% de los datos

fueron 2007, 2008 y 2009, aunque algunas fincas sólo presentaron datos para uno de estos tres años.

Adicionalmente, se recibieron datos de embolse a nivel de lote. Estos datos, al igual que los datos de peso de racimo, tienen una periodicidad semanal, y fueron agregados hasta dos semanas, para facilitar el análisis, debido a que los datos de severidad de Sigatoka negra tienen una periodicidad de dos semanas.

3. Selección de lote de análisis y organización de datos

Debido a la poca uniformidad que presentó cierta parte de la información recibida, y dado que el objetivo del presente estudio es desarrollar un método de análisis para los datos históricos de Sigatoka negra, se decidió trabajar con el "mejor caso" en términos de disponibilidad de datos históricos tanto de enfermedad, como de producción, y meteorológicos.

a. Criterios de selección

Para facilitar el desarrollo de un método robusto para el análisis de los datos, se seleccionó el "mejor lote", con base en los siguientes criterios:

- Disponibilidad de la serie de tiempo 2007-2009 de variables de Sigatoka negra
- Cercanía a la estación meteorológica más cercana con mayor cantidad de datos disponibles
- Disponibilidad de la serie de tiempo 2007-2009 de peso de racimo y embolse

Se seleccionó el lote 1 de la finca Castilletes (código 20117), ubicado en las proximidades de la estación meteorológica 1 (Figura 3). La estación meteorológica 1 fue la que contó con la mejor información histórica para el período de análisis (2007-2009), y este lote (1) de la finca Castilletes, contó con información suficiente para las variables peso de racimo, embolse y las cuatro variables usadas para medir el ataque de la enfermedad.

En la figura 3, además, se indican los demás lotes que hacen parte del programa de monitoreo de Sigatoka negra (lotes rojos), y los lotes que hacen parte del programa de monitoreo de producción (lotes grises). Asimismo, los lotes que hacen parte de ambos programas (lotes azules) y el lote seleccionado para el análisis (lote amarillo). Este lote, perteneciente a la finca Castilletes, pertenece a la zona agroecológica 6 (ver reporte de zonas agroecológicas. Cabe anotar que la distribución espacial de las estaciones meteorológicas es muy buena, y permite una cobertura adecuada de cada una de las fincas de la región. Se sugiere, sin embargo, que los sensores sean calibrados y revisados

con alta periodicidad, de tal manera que se asegure su calidad, y por tanto la relevancia de los resultados de cualquier análisis que con estos datos se efectúe.

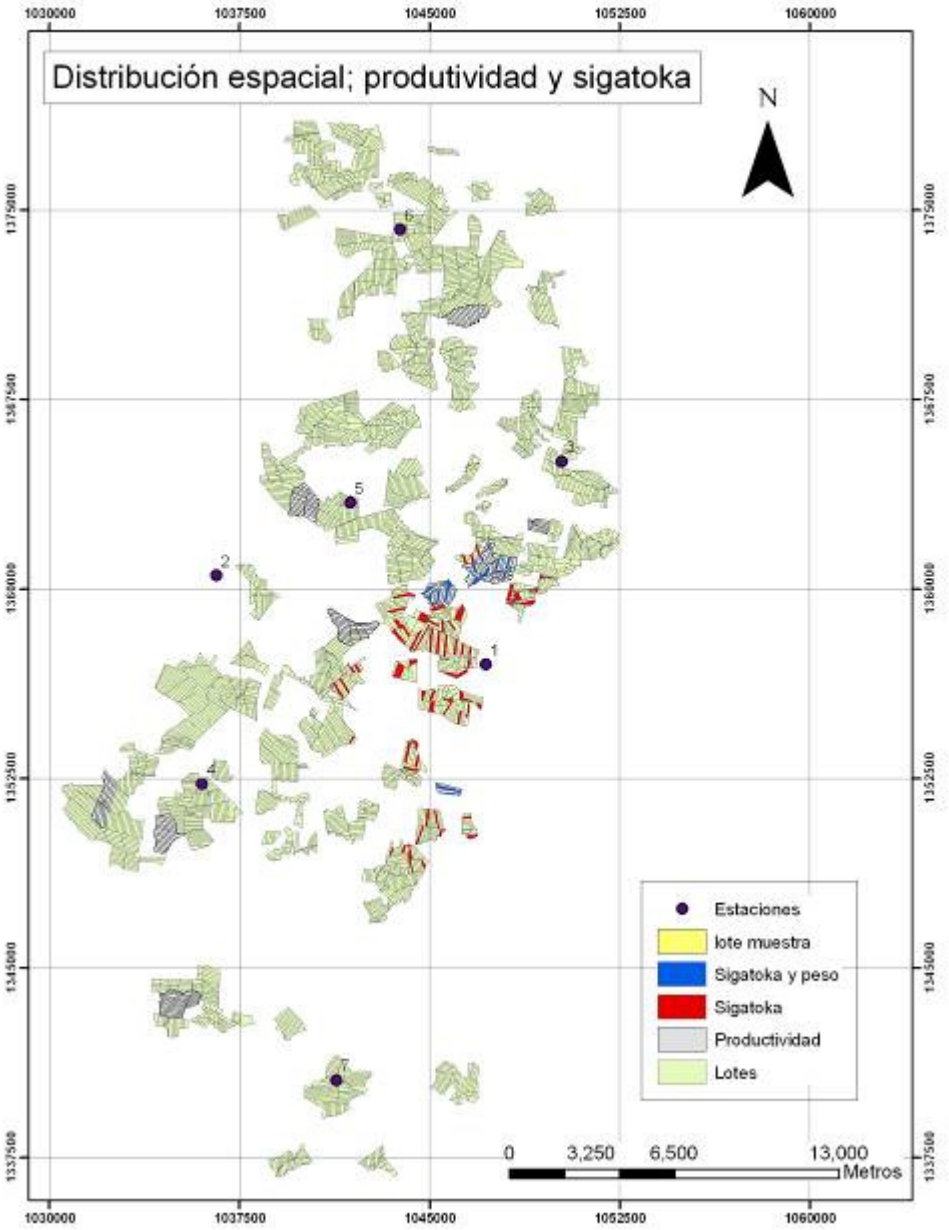

Figura 3 Distribución espacial de fincas, lotes y estaciones meteorológicas.

Del lote seleccionado, se extrajeron los datos de peso de racimo y embolse semanales, y los datos de ataque y severidad de Sigatoka negra para el período de análisis (2007-2009). De la misma manera, se extrajeron los datos climáticos de la estación meteorológica 1. La selección de un solo lote para el análisis obedece a que el principal objetivo de este estudio es la exploración de los datos, motivo por el cual se escogió el

"caso óptimo" en términos de disponibilidad y calidad. El método desarrollado, sin embargo, puede ser aplicado a cualquier lote con la suficiente disponibilidad de datos.

b. Límites temporales e intervalos de tiempo para análisis

Dado que la mayor disponibilidad de datos fue durante el período 2007-2009, será con este período con el que se trabajará, tanto para los datos de Sigatoka negra, como para los datos de producción, los datos de NDVI y los datos climatológicos.

Por otro lado, dado que el presente análisis es un análisis de variabilidad y tendencia histórica en el ataque de la enfermedad, es crítico definir un intervalo de tiempo para el análisis secuencial que sea congruente para todas las variables bajo análisis. La escogencia de este intervalo está limitada a la variable con la menor periodicidad. En este caso, tanto las mediciones de Sigatoka negra, como las de NDVI tienen la periodicidad más baja (2 semanas), y por tanto los análisis históricos se realizarán con datos cada 2 semanas. Las variables a utilizar son:

Variables dependientes
- Número de hojas totales
- Hoja más joven infectada
- Hoja más joven manchada
- Severidad

Variables independientes
- Humedad relativa (promedio de las 2 semanas)
- Humedad relativa (máximo de las 2 semanas)
- Humedad relativa (mínimo de las 2 semanas)
- Radiación solar
- Temperatura media (promedio de las 2 semanas)
- Temperatura máxima (promedio de las 2 semanas)
- Temperatura mínima (promedio de las 2 semanas)
- Precipitación (estación meteorológica, total de las 2 semanas)
- Precipitación (satélite TRMM, total de las 2 semanas)
- Promedio número de horas con humedad relativa por encima de 90%
- NDVI (promedio de las 2 semanas)
- NDVI (mínimo de las 2 semanas)
- NDVI (máximo de las 2 semanas)
- NDVI (desviación estándar de las 2 semanas)

Variables de producción
- Número de racimos embolsados (total de las 2 semanas)
- Peso de racimo (promedio de las 2 semanas)

El total de datos fue 77 para los 3 años (2007, 2008 y 2009). Con estas variables se realizaron análisis de variabilidad históricos, usando en parte la metodología propuesta por Pérez et al. (2000).

4. Metodología aplicada

La metodología aplicada, como se mencionó con anterioridad, obedece a un análisis de relación entre las variables ambientales (NDVI, clima) y el ataque de la enfermedad. Todo basado en la premisa de que el ambiente modifica el comportamiento, crecimiento, y dispersión del patógeno *Mycosphaerella fijiensis*, causante de la enfermedad. Sin embargo, el grado de influencia del clima en la enfermedad es variable tanto en el tiempo como en el espacio, y por lo tanto es crítico analizarlas en los períodos apropiados y con los desfases adecuados. Además de esto, las relaciones entre estas variables no siempre son lineales, por lo que a veces se requieren transformaciones a estas variables antes de analizarlas. Todos los procedimientos estadísticos aplicados en el presente estudio fueron realizados con el paquete estadístico R, de libre acceso y gratuito, disponible en http://www.r-project.org/.

a. Transformación de variables

Se aplicaron tres transformaciones diferentes a las 16 variables independientes. Estas transformaciones se aplicaron con el fin de explorar relaciones que no fueran necesariamente lineales, y fueron:

Polinomial de segundo orden a cada dato centrado (Ecuación 1)

$$X_{i-T} = \left(X_i - \overline{X}\right)^2 \qquad \text{(Ec. 1)}$$

Donde el dato transformado (X_{i-T}) se calcula como la resta entre el dato original (X_i) y el promedio de los datos ($\overline{X}$), elevado al cuadrado.

Exponencial a cada dato normalizado (Ecuación 2)

$$X_{i-T} = e^{\left(\frac{X_i - \overline{X}}{\sigma}\right)} \qquad \text{(Ec. 2)}$$

Donde el dato transformado (X_{i-T}) se calcula como la constante de Euler (e), elevada a la resta entre el dato original (X_i) y el promedio de los datos ($\overline{X}$), dividido por la desviación estándar (σ).

Logaritmo en base 10 a cada dato crudo (Ecuación 3)

$$X_{i-T} = Log(X_i) \qquad \text{(Ec. 3)}$$

Donde el dato transformado (X_{i-T}) se calcula como el logaritmo en base 10 del dato original sin transformar (X_i).

Con esto, las 16 variables independientes iniciales aumentan a 64 variables. Con estas 64 variables se realizaron todos los análisis subsiguientes.

b. Desfase de mediciones

Como se mencionó con antelación, aunque hay una relación entre el comportamiento del clima y el comportamiento del patógeno, estas relaciones no se presentan en el mismo instante en el tiempo. En ocasiones la precipitación de algunos días, o incluso semanas antes del máximo ataque de la enfermedad es la que influencia dicho ataque, mediante el humedecimiento excesivo o encharcamiento del suelo, el aumento de la humedad relativa al interior del cultivo, o la formación de una lámina de agua en la hoja que facilita la germinación y desarrollo del hongo.

Por todo lo anterior, para cada una de las medidas de ataque de la enfermedad, se realizaron cuatro desfases en los datos climáticos, estos desfases consistieron en movimientos de 1 intervalo hacia atrás. Es decir, se construyeron parejas entre datos de la misma semana, y de 1, 2, 3 y 4 fechas anteriores. A cada dato de ataque de la enfermedad le correspondieron, por tanto, cinco mediciones diferentes de clima (una para cada fecha de desfase). Las únicas variables que se consideró innecesario realizar desfase fueron las de producción, puesto que es la enfermedad quien las influencia y no al revés, por tanto el desfase, de existir, existiría hacia el futuro, y no hacia el pasado.

c. Correlaciones

Para cada variable dependiente, y para cada desfase, se realizaron correlaciones con las 64 variables independientes. De esta manera, se logró relacionar el ataque de la enfermedad con diferentes eventos climáticos, no sólo de la semana actual, sino de semanas anteriores (hasta 8 semanas, o T=4).

Estas correlaciones se encuentran en el archivo “Correlaciones.xls”. Cada fila indica una pareja de variables, y las columnas “Variable dependiente” y “Variable independiente” indican entre qué pareja de variables fue realizada la correlación. Las columnas subsiguientes cada una presenta el coeficiente de correlación de Pearson (R), el coeficiente de determinación (R^2), el número de parejas con las que se efectuó el procedimiento, y la significancia estadística del coeficiente de correlación. Para los diferentes desfases se encontraron diferencias en la significancia estadística de las correlaciones. Para cada variable dependiente se encontraron los siguientes resultados:

Número de hojas totales (NHT)

Las correlaciones más fuertes para el número total de hojas se encontraron, dependiendo de las variables independientes, con diferentes desfases. Así por ejemplo, con la humedad relativa promedio y máxima, las correlaciones más fuertes se observaron con 4 fechas de desfase, mientras que para la humedad relativa mínima, las correlaciones más fuertes se observaron con 1 fecha de desfase. El NDVI mínimo se encontró sólo ligeramente correlacionado con el total de hojas con 2 a 3 fechas de desfase, las demás variables de NDVI sólo tuvieron correlaciones significativas al nivel $p<0.1$, lo que se considera bajo.

- Correlaciones negativas con
  - La humedad relativa mínima sin transformar
  - La humedad relativa mínima con transformación exponencial, y
  - La humedad relativa mínima con transformación logarítmica

  Esto significa que a medida que la humedad relativa mínima aumenta, el número de hojas totales tiende a disminuir, probablemente a causa del daño severo en las hojas a causa de la enfermedad (que se estimula por la presencia de alta humedad en el ambiente)

- Correlación positiva con
  - La temperatura media con transformación polinomial
  - La temperatura media con transformación exponencial
  - La precipitación total de estación meteorológica con transformación logarítimica
  - La precipitación total de TRMM con transformación logarítimica
  - El peso del racimo sin transformar
  - El peso del racimo con transformación exponencial
  - El peso del racimo con transformación logarítmica

  Estas correlaciones indican que probablemente una mayor temperatura media incrementa la tasa fotosintética y probablemente influencia la generación de nuevas hojas, no obstante, dado que estas relaciones ocurren con las transformaciones polinomiales y exponenciales en algunos casos, es muy probable que a partir de ciertos rangos la relación cambie su sentido

Hoja más joven infectada (HMJI)

Las correlaciones más fuertes con la hoja más joven infectada se encontraron con la humedad relativa promedio con 4 fechas de desfase, con la humedad relativa máxima con 2 fechas de desfase, con la humedad relativa mínima con 3 fechas de desfase, con la temperatura media con 0 o 1 fecha de desfase, con la temperatura mínima con 1 o 2 fechas de desfase, con el número de horas con humedad relativa mayor a 90% con 4

fechas de desfase. El número de racimos embolsados y el peso del racimo estuvieron altamente correlacionados con la HMJI ($p<0.01$). Muy poco significativas fueron las correlaciones con el NDVI ($p<0.1$) en la mayoría de los casos.

- Correlaciones negativas con
  - Humedad relativa promedio, máxima y mínima sin transformar
  - Humedad relativa promedio, máxima y mínima con transformación logarítmica
  - Humedad relativa máxima y mínima con transformación exponencial
  - Número de horas con humedad relativa mayor a 90%, todas sus transformaciones excepto la polinomial
  - Racimos embolsados sin transformar y con transformación logarítmica

- Correlaciones positivas con
  - Temperatura media y todas sus transformaciones, excepto la polinomial
  - Temperatura máxima sin transformar y con transformación logarítmica
  - Peso de racimo con todas sus transformaciones, excepto la polinomial

Hoja más joven infectada (HMJM)

La HMJM se encontró, nuevamente, altamente correlacionada con la humedad relativa promedio con desfase de 4 fechas, con la humedad relativa máxima sin desfase, con la radiación solar ($p<0.05$) con desfase de 3 a 4 fechas, con la temperatura media y máxima, con la precipitación de 4 y 3 fechas anteriores, y con el número medio de horas por encima de 90% de humedad relativa con 4 fechas de desfase. El número de racimos embolsados se encontró sólo ligeramente correlacionado con la HMJM ($p<0.05$ y $p<0.1$).

- Correlaciones negativas con
  - La humedad relativa máxima sin transformar
  - La humedad relativa máxima con transformación exponencial, y
  - La humedad relativa máxima con transformación logarítmica

- Correlación positiva con
  - La temperatura media sin transformar
  - La temperatura media con transformación logarítmica

Severidad (SEV)

La severidad de la enfermedad fue la variable que menos correlación mostró con las variables climáticas, independientemente del desfase aplicado. La severidad estuvo relacionada, sin embargo, con la humedad relativa promedio y máxima (con 2 a 4 fechas de desfase), y con la temperatura mínima (2 fechas de desfase) y la precipitación medida en la estación meteorológica y medida por el satélite TRMM (4 fechas de desfase), y lo

mismo ocurrió con el número de horas con humedad relativa por encima de 90% ($p<0.05$).

- Correlación negativa con
  - Humedad relativa promedio con transformación polinomial

- Correlación positiva con
  - Humedad relativa máxima sin transformar y con transformación logarítimica

La severidad de la enfermedad, al igual que la HMJM, estuvo correlacionada con el número de racimos embolsados en la fecha (semana) de medición, aunque sólo ligeramente ($p<0.05$).

En general se encontró que la variabilidad histórica de las variables que miden el ataque de la enfermedad se puede explicar mediante algunas otras variables, aunque sólo hasta cierto nivel. Se encontró que una alta humedad relativa en general favorece el desarrollo del patógeno, y por tanto causa una disminución en el número de hojas totales, y causa que la hoja más joven infectada y manchada sean más jóvenes. La precipitación causa un efecto de fortalecimiento del patógeno, pero este se observa después de cierto tiempo (2 a 4 fechas después del evento de precipitación en cuestión). Las transformaciones de variables en algunos casos estuvieron correlacionadas de manera más significativa con las variables dependientes que las variables originales (sin transformar), lo que se debe a la presencia de no-linealidades entre las variables. Todos estos resultados coinciden con lo reportado en la literatura (Pérez et al. 2000; Gauhl 1994; Calpouzos et al. 1962; Orozco-Santos et al. 2001; Valadares et al. 2007; entre otros).

Se seleccionó la correlación más fuerte para cada pareja variable independiente-variable dependiente a través de los diferentes desfases (Tabla 1), para el desarrollo de modelos multivariados.

Tabla 1 Correlaciones (R^2) más fuertes considerando todos los posibles desfases para cada pareja posible de variable dependiente-variable independiente, para todas las transformaciones realizadas, exceptuando las variables de producción.

ID	Variable	Transf.*	NHT**	HMJI**	HMJM**	SEV**
HR	Humedad relativa (promedio)	ST	**4 (0.076 **)**	**4 (0.307 ***)**	**4 (0.133 ***)**	**4 (0.085 **)**
HR2	Humedad relativa (promedio)	PSO	**0 (0.075 **)**	4 (0.058 *)	**1 (0.094 ***)**	**0 (0.155 ***)**
HREXP	Humedad relativa (promedio)	EXP	0 (0.057 *)	**4 (0.246 ***)**	**4 (0.071 **)**	4 (0.054 *)
HRLOG	Humedad relativa (promedio)	LOG	**4 (0.078 **)**	**4 (0.303 ***)**	**4 (0.137 ***)**	**4 (0.087 **)**
HRX	Humedad relativa (maximo)	ST	**4 (0.077 **)**	**1 (0.279 ***)**	**0 (0.119 ***)**	**0 (0.104 ***)**
HRX2	Humedad relativa (maximo)	PSO	NS	2 (0.055 *)	**3 (0.062 **)**	**2 (0.065 **)**
HRXEXP	Humedad relativa (maximo)	EXP	**4 (0.102 ***)**	**0 (0.295 ***)**	**0 (0.11 ***)**	**1 (0.074 **)**
HRXLOG	Humedad relativa (maximo)	LOG	**4 (0.077 **)**	**0 (0.298 ***)**	**0 (0.119 ***)**	**0 (0.104 ***)**
HRN	Humedad relativa (minimo)	ST	**1 (0.227 ***)**	**3 (0.236 ***)**	NS	NS
HRN2	Humedad relativa (minimo)	PSO	**1 (0.11 ***)**	NS	NS	NS
HRNEXP	Humedad relativa (minimo)	EXP	**1 (0.231 ***)**	**1 (0.117 ***)**	NS	NS

HRNLOG	Humedad relativa (minimo)	LOG	**1 (0.218 ***)**	**3 (0.246 ***)**	NS	NS
RS	Radiacion solar	ST	NS	NS	**3 (0.061 **)**	NS
RS2	Radiacion solar	PSO	NS	NS	4 (0.052 *)	NS
RSEXP	Radiacion solar	EXP	NS	NS	NS	NS
RSLOG	Radiacion solar	LOG	NS	NS	3 (0.079 **)	NS
T	Temperatura media	ST	3 (0.053 *)	**0 (0.16 ***)**	**2 (0.115 ***)**	**2 (0.09 ***)**
T2	Temperatura media	PSO	**0 (0.137 ***)**	**1 (0.072 **)**	NS	NS
TEXP	Temperatura media	EXP	**0 (0.086 ***)**	**1 (0.11 ***)**	0 (0.055 *)	NS
TLOG	Temperatura media	LOG	3 (0.051 *)	**0 (0.162 ***)**	**2 (0.12 ***)**	**2 (0.095 ***)**
TX	Temperatura maxima	ST	**3 (0.074 **)**	**0 (0.153 ***)**	4 (0.043 *)	NS
TX2	Temperatura maxima	PSO	**1 (0.072 **)**	NS	NS	NS
TXEXP	Temperatura maxima	EXP	3 (0.058 *)	NS	NS	NS
TXLOG	Temperatura maxima	LOG	**3 (0.072 **)**	**0 (0.158 ***)**	4 (0.043 *)	NS
TN	Temperatura minima	ST	NS	**2 (0.108 ***)**	**3 (0.113 ***)**	**2 (0.098 ***)**
TN2	Temperatura minima	PSO	NS	**1 (0.141 ***)**	NS	1 (0.058 *)
TNEXP	Temperatura minima	EXP	NS	**2 (0.115 ***)**	**2 (0.089 ***)**	**2 (0.088 ***)**
TNLOG	Temperatura minima	LOG	NS	**2 (0.105 ***)**	**3 (0.113 ***)**	**2 (0.097 ***)**
P	Precipitacion	ST	**4 (0.067 **)**	4 (0.05 *)	1 (0.046 *)	4 (0.048 *)
P2	Precipitacion	PSO	NS	NS	2 (0.043 *)	**2 (0.062 **)**
PEXP	Precipitacion	EXP	NS	NS	NS	NS
PLOG	Precipitacion	LOG	**1 (0.206 ***)**	NS	4 (0.088 **)	**4 (0.095 ***)**
TRMM	TRMM	ST	**4 (0.115 ***)**	4 (0.059 *)	**4 (0.11 ***)**	**4 (0.145 ***)**
TRMM2	TRMM	PSO	MS	3 (0.05 *)	**2 (0.073 **)**	2 (0.071 **)
TRMMEXP	TRMM	EXP	**4 (0.071 **)**	NS	4 (0.049 *)	4 (0.057 *)
TRMMLOG	TRMM	LOG	**1 (0.266 ***)**	**4 (0.076 **)**	**4 (0.175 ***)**	**4 (0.249 ***)**
HR90	Horas HR>90%	ST	4 (0.043 *)	**4 (0.334 ***)**	**4 (0.153 ***)**	**4 (0.125 ***)**
HR902	Horas HR>90%	PSO	1 (0.041 *)	**4 (0.093 ***)**	NS	NS
HR90EXP	Horas HR>90%	EXP	NS	**4 (0.197 ***)**	**4 (0.078 **)**	**4 (0.074 **)**
HR90LOG	Horas HR>90%	LOG	4 (0.054 *)	**4 (0.303 ***)**	**4 (0.2 ***)**	**3 (0.067 **)**
NDVIN	NDVI (min)	ST	2 (0.057 *)	**4 (0.064 **)**	NS	NS
NDVIN2	NDVI (min)	PSO	3 (0.048 *)	NS	NS	NS
NDVINEXP	NDVI (min)	EXP	NS	**4 (0.063 **)**	NS	NS
NDVINLOG	NDVI (min)	LOG	**2 (0.073 **)**	4 (0.048 *)	NS	NS
NDVIP	NDVI (promedio)	ST	NS	NS	NS	NS
NDVIP2	NDVI (promedio)	PSO	NS	NS	NS	NS
NDVIPEXP	NDVI (promedio)	EXP	NS	0 (0.052 *)	NS	2 (0.044 *)
NDVIPLOG	NDVI (promedio)	LOG	NS	NS	NS	NS
NDVIX	NDVI (max)	ST	NS	NS	NS	NS
NDVIX2	NDVI (max)	PSO	NS	NS	NS	NS
NDVIXEXP	NDVI (max)	EXP	NS	0 (0.049 *)	0 (0.04 *)	0 (0.045 *)
NDVILOG	NDVI (max)	LOG	NS	NS	NS	NS
NDVISD	NDVI (std)	ST	2 (0.055 *)	NS	NS	NS
NDVISD2	NDVI (std)	PSO	NS	NS	NS	NS
NDVISDEXP	NDVI (std)	EXP	NS	NS	NS	NS
NDVISDLOG	NDVI (std)	LOG	NS	NS	NS	NS

*Transformaciones de las variables: ST: sin transformar, PSO: Polinomial de segundo orden, EXP: exponencial, LOG: logarítmica. **Desfase (R^2 Significancia) de las correlaciones, (***=p<0.01; **=p<0.05; *=p<0.1).

Aunque las variables de NDVI no tuvieron correlaciones demasiado fuertes con los datos de Sigatoka negra, es claro que la Sigatoka influencia el grado de verdor del cultivo, puesto que afecta directamente las hojas. Probablemente el NDVI tiene una relación con desfases hacia el futuro, más que hacia el pasado.

d. Gráficos de variabilidad histórica

Aunque los análisis de correlaciones revelaron relaciones significativas entre las variables, un simple análisis de correlaciones a veces puede enmascarar resultados interesantes, puesto que limita los datos a un solo valor numérico. Por este motivo, se presentan gráficos de variabilidad históricos para las variables climáticas más importantes (humedad relativa mínima, y promedio, temperatura media, precipitación total [estación y TRMM], número de horas con humedad relativa por encima de 90%, y NDVI promedio)

Se observa una relación histórica entre el número de hojas totales y la humedad relativa promedio (Figura 4). Con aumentos en la humedad relativa, se observa un efecto a largo plazo (de 2 a 4 fechas) de disminución en el número de hojas totales (dado por el aumento en la prevalencia de la enfermedad). Cuando se observa una disminución en la humedad relativa promedio, con el respectivo desfase, se observa un aumento en el número de hojas totales (producto del aumento en la tasa fotosintética y tasa de emisión foliar).

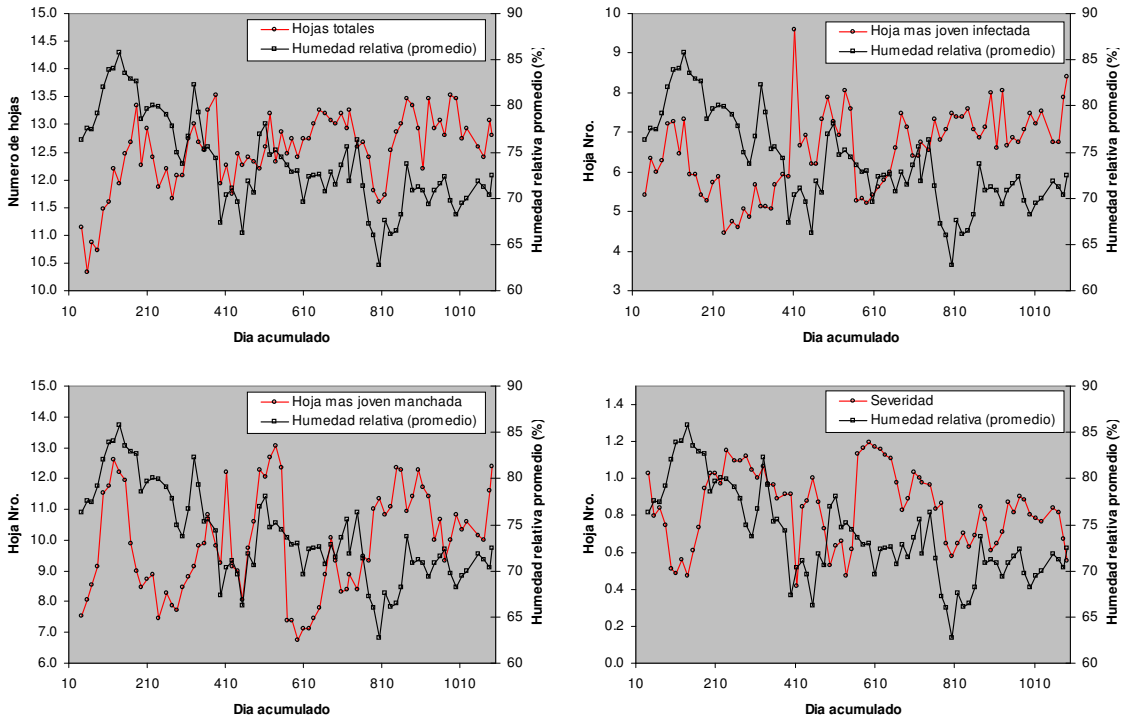

Figura 4 Comportamiento histórico de las cuatro variables de ataque de la enfermedad y la humedad relativa promedio.

De otro lado, la hoja más joven manchada y la hoja más joven infectada tienen un comportamiento relativamente similar, a través del tiempo, aunque la hoja más joven manchada tiende a presentar mucha mayor variabilidad a través del tiempo. En términos generales, el aumento en la humedad relativa promedio, causa que la hoja más joven

manchada y la hoja más joven infectada se hagan más jóvenes, pero con un desfase de algunas semanas (entre 4 y 8 semanas). La severidad también muestra una respuesta similar, con un desfase de varias semanas y siendo relativamente sensible a los cambios a medida que avanza el desarrollo de la enfermedad en el tiempo.

Por otro lado, la humedad relativa mínima (Figura 5) también parece afectar el comportamiento de las variables, aunque en ciertos períodos parecen poco sensibles a la misma. El número de hojas totales tiende a disminuir a medida que la humedad relativa aumenta con un desfase entre 1 y 3 fechas (una fecha representa 2 semanas), aunque hay períodos de poca variación en la humedad en los que parece que tiene poca influencia en el desarrollo del patógeno.

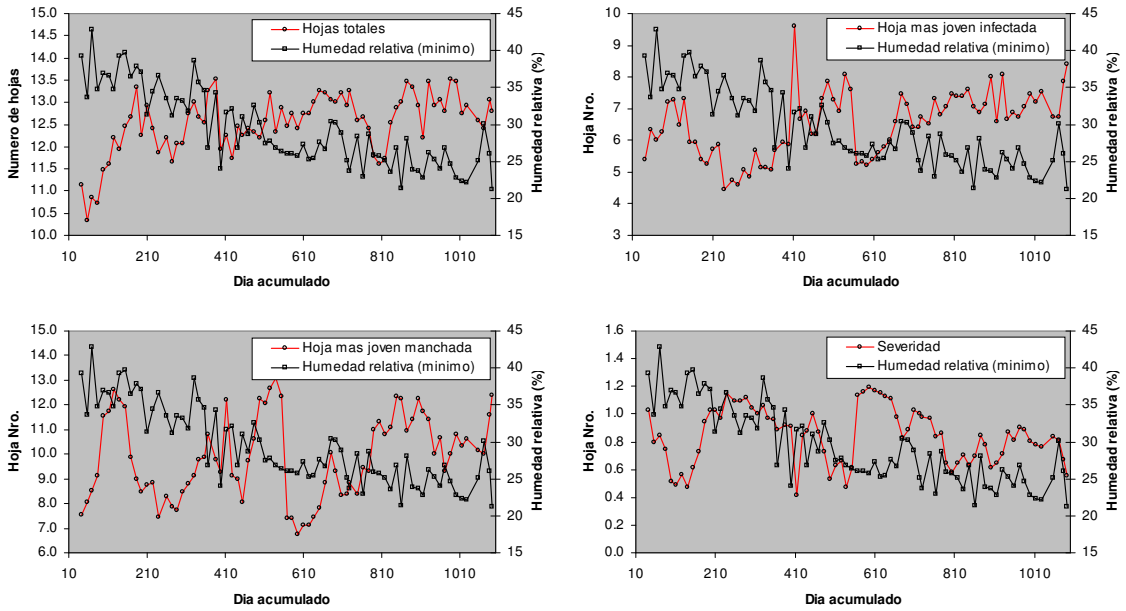

Figura 5 Comportamiento histórico de las cuatro variables de ataque de la enfermedad y la humedad relativa mínima.

La HMJI y la HMJM parecen estar influenciadas a un menor nivel por la humedad relativa mínima, aunque algunas variaciones (disminuciones especialmente) en la humedad relativa mínima, parecen causar un incremento en el número de la hoja, no obstante, la tendencia no es tan clara como en el caso de la humedad relativa promedio.

La temperatura media (Figura 6) tiene una influencia a pequeña escala en las variaciones cada 2 semanas de las variables que miden el ataque de la enfermedad. Sin embargo, grandes cambios en la temperatura media (como ocurre después del día 810), no afectan de manera significativa la Sigatoka negra. De otro lado, hay un fenómeno observable de pequeña escala temporal en el que la enfermedad tiende a verse afectada por la temperatura media, pero sólo después de cierto umbral. Las disminuciones en la

temperatura media, y muy probablemente, el número de horas con temperaturas muy bajas tienden a hacer la hoja más joven manchada e infectada menos jóvenes.

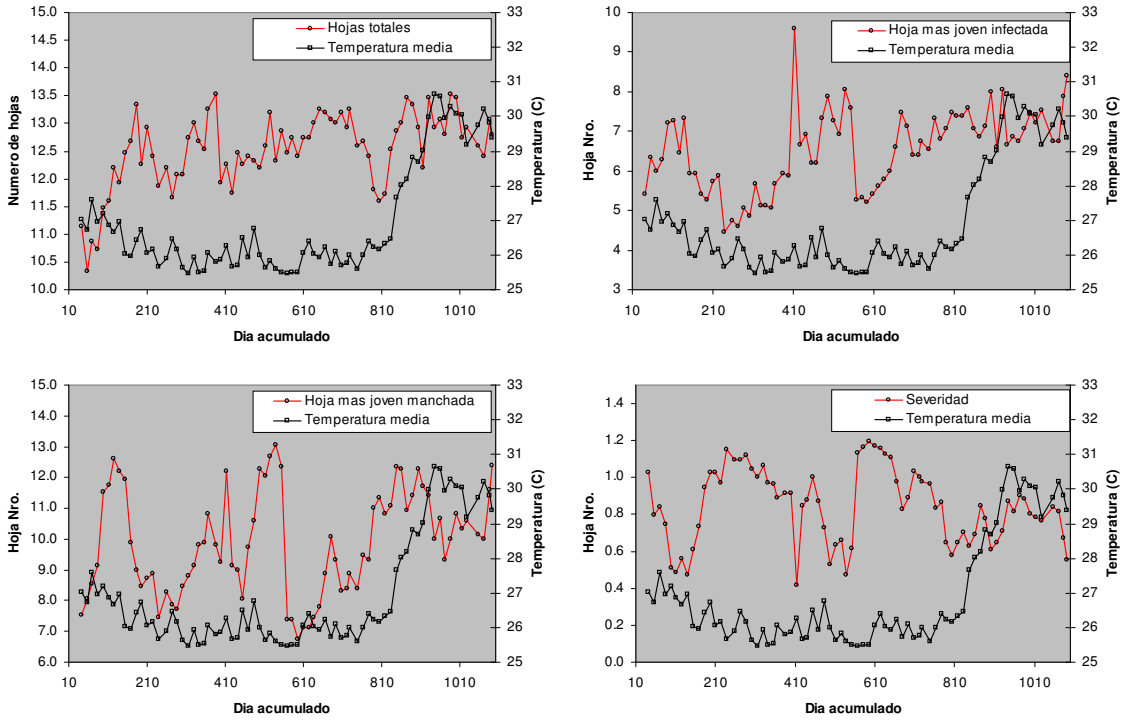

Figura 6 Comportamiento histórico de las cuatro variables de ataque de la enfermedad y la temperatura media.

La severidad de la enfermedad tiende a verse ligeramente afectada en particular por las disminuciones sustanciales en la temperatura media, que hacen disminuir la severidad de la enfermedad con 1 a 2 fechas de desfase.

La precipitación medida por la estación meteorológica y la precipitación medida por el satélite TRMM, presentan una tendencia bastante similar, lo que indica que, al menos en la forma del patrón de precipitación, ambas fuentes de información presentan muy alta similitud (Figura 9). Sin embargo, los valores crudos de precipitación difieren de manera significativa, pues el satélite tiende a subestimar la precipitación real.

En general, la precipitación parece tener un efecto indirecto en la enfermedad, puesto que influencia otras variables como la humedad relativa, y la evapotranspiración, así como la formación de rocío en las hojas. Disminuciones en precipitación parecen estar relacionadas con aumentos en la prevalencia de la enfermedad entre 2 a 4 fechas después del evento de precipitación, esto es observable en las disminuciones sustanciales que se observan en la HMJI y la HMJM entre 4 y 8 semanas después de uno o varios eventos de precipitación.

La enfermedad no responde inmediatamente, y tampoco responde con un solo evento. Generalmente se requiere un período de altas precipitaciones para que el hongo germine, y luego de este, el nivel de ataque se mantiene relativamente constante. Si las precipitaciones se mantienen bajas durante ese período, el crecimiento del hongo se obstaculiza, sin embargo en Urabá, las precipitaciones se mantienen en niveles relativamente altos mensualmente, lo que ocasiona que la respuesta del patógeno ante eventos específicos de precipitación sea relativamente baja.

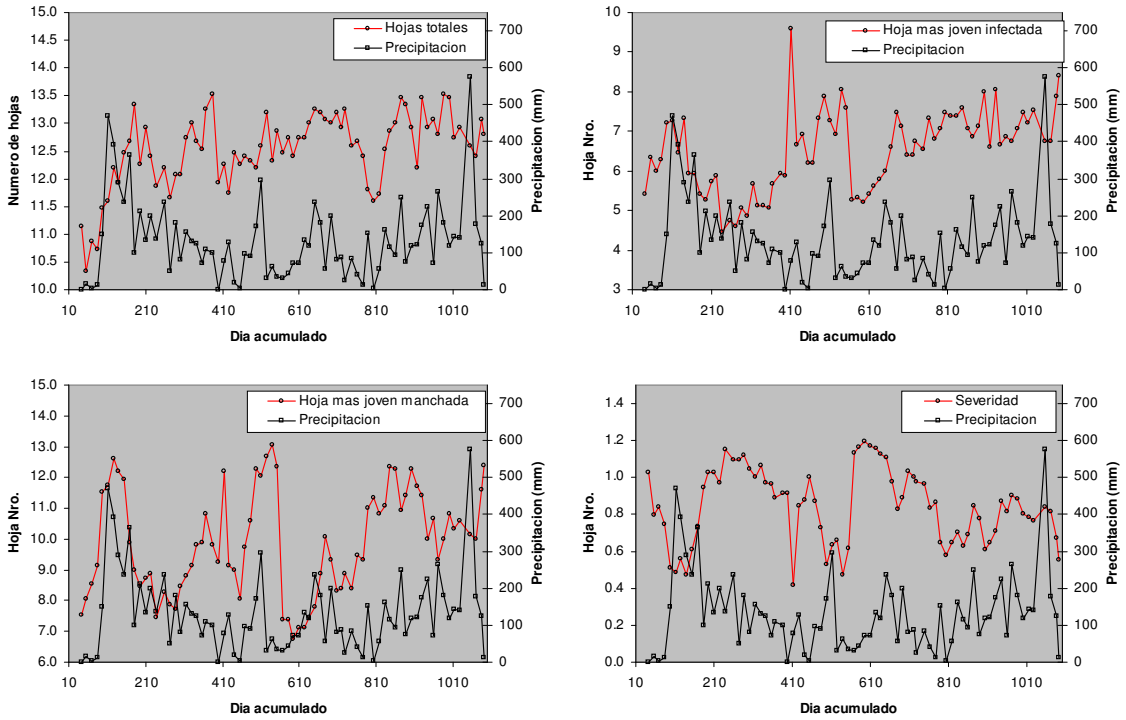

Figura 7 Comportamiento histórico de las cuatro variables de ataque de la enfermedad y la precipitación medida en la estación meteorológica.

El índice de infección, por otro lado, responde a las variaciones de precipitación de varias semanas, y hasta un mes antes. Esto implica que a partir de ciertas tendencias en la precipitación es posible detectar y predecir tendencias en el ataque de la enfermedad, y pronosticar, con cierto nivel de confianza, el ataque.

De la misma manera, las precipitaciones medidas por el satélite tienen un impacto relativo a la frecuencia de eventos significativos de precipitación, y de la misma manera influyen el desarrollo y prevalencia del patógeno. Dada la similitud en comportamiento entre la precipitación medida en la estación meteorológica y la medida por el satélite, las mismas tendencias pueden ser observadas: varias semanas después de un evento significativo de precipitación se liberan mayor cantidad de ascosporas, y con el medio adecuado para su germinación, empiezan su ataque continuado hasta un nuevo evento significativo de precipitación, o hasta que la sequía cause la disminución de poblaciones del hongo.

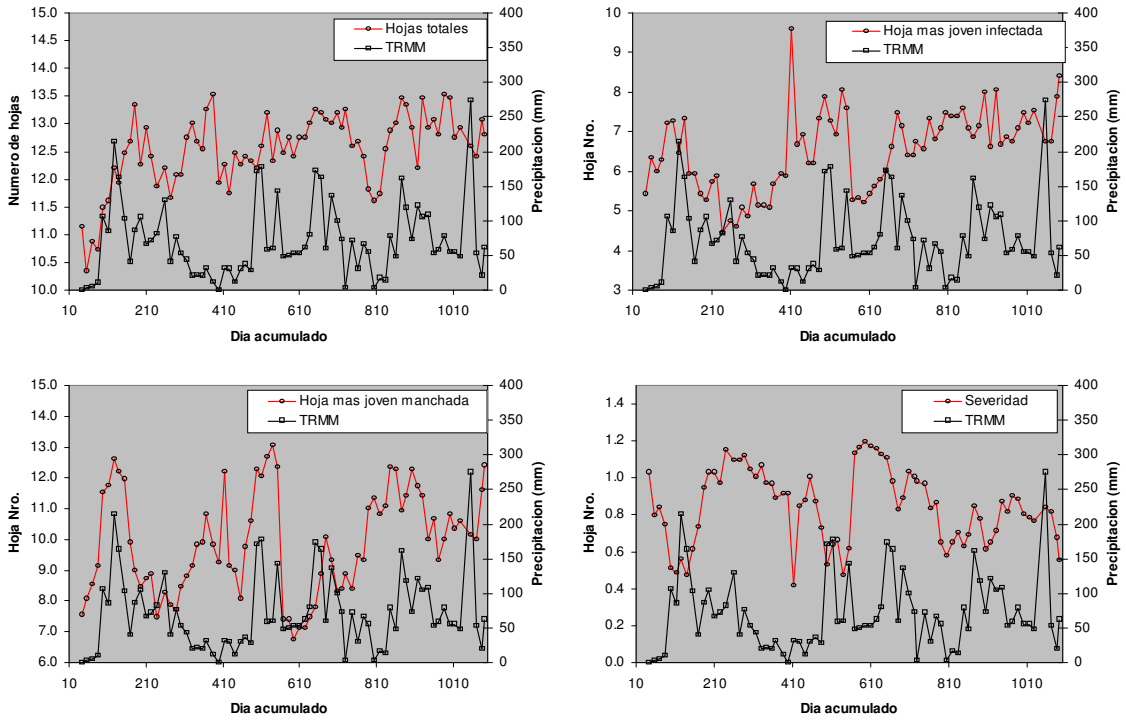

Figura 8 Comportamiento histórico de las cuatro variables de ataque de la enfermedad y la precipitación medida por el satélite TRMM.

Cualquier sistema de preaviso biológico debe tener en cuenta la precisión de las variables utilizadas así como su resolución temporal. En general, las variables climáticas utilizadas en el presente estudio tienen una alta influencia en el comportamiento de la enfermedad, según lo indicado por la literatura, el conocimiento experto, y los datos de campo colectados.

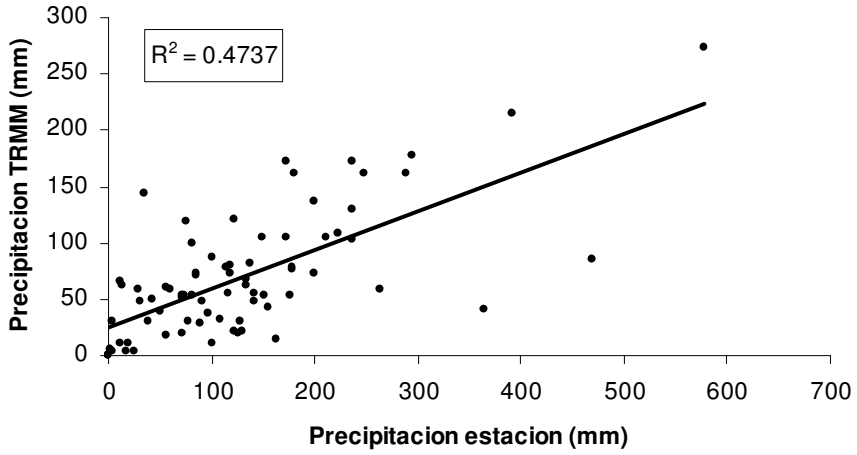

Figura 9 Gráfico X-Y de la precipitación medida en la estación meteorológica

El número de horas con humedad relativa mayor a 90% parece incrementar el ataque de la enfermedad en tanto el valor mismo de esta variable se incrementa, con 1-3 semanas de desfase, dependiendo del año analizado.

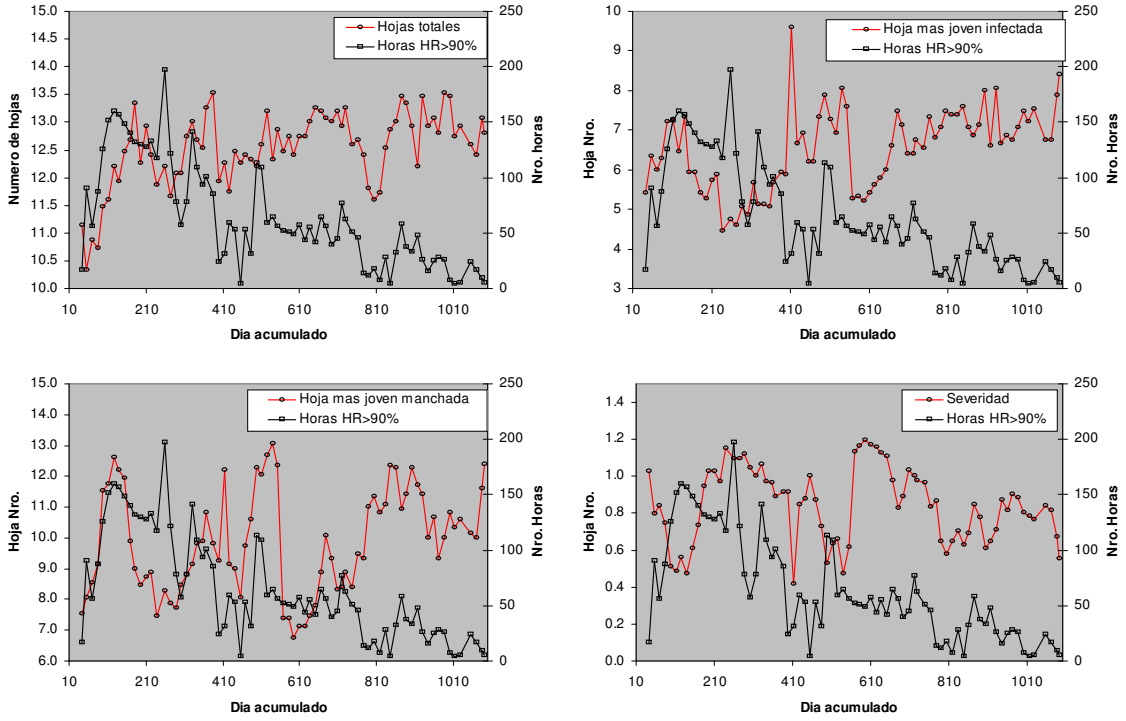

Figura 10 Comportamiento histórico de las cuatro variables de ataque de la enfermedad y el número de horas con humedad relativa mayor a 90%

En general el número de horas no parece influenciar de manera notable pequeños eventos en los que la enfermedad aumenta considerablemente, pero sí la variación estacional de largo plazo. En este sentido, se observa cómo la HMJI y la HMJM se hacen menos jóvenes a medida que el número de horas con HR > 90% se hace menor.

En relación con el NDVI, se observa una respuesta mucho más inmediata de una variable y la otra. El NDVI tiende a aumentar a medida que el número de hojas totales aumenta, como respuesta inmediata al aumento en la cobertura del cultivo (*canopy*). De la misma manera, cuando la HMJM y la HMJI se hacen muy jóvenes, el NDVI tiende a disminuir, indicando una disminución en el verdor del lote analizado.

Estas tendencias, sin embargo, no son tan claras cuando no hay eventos significativos (disminución abrupta, aumento vertiginoso) respecto a la prevalencia de la enfermedad. Es probable que la alta variabilidad temporal del NDVI esté causando este efecto, para lo cual, el uso de promedios a través de varias fechas resulta de utilidad.

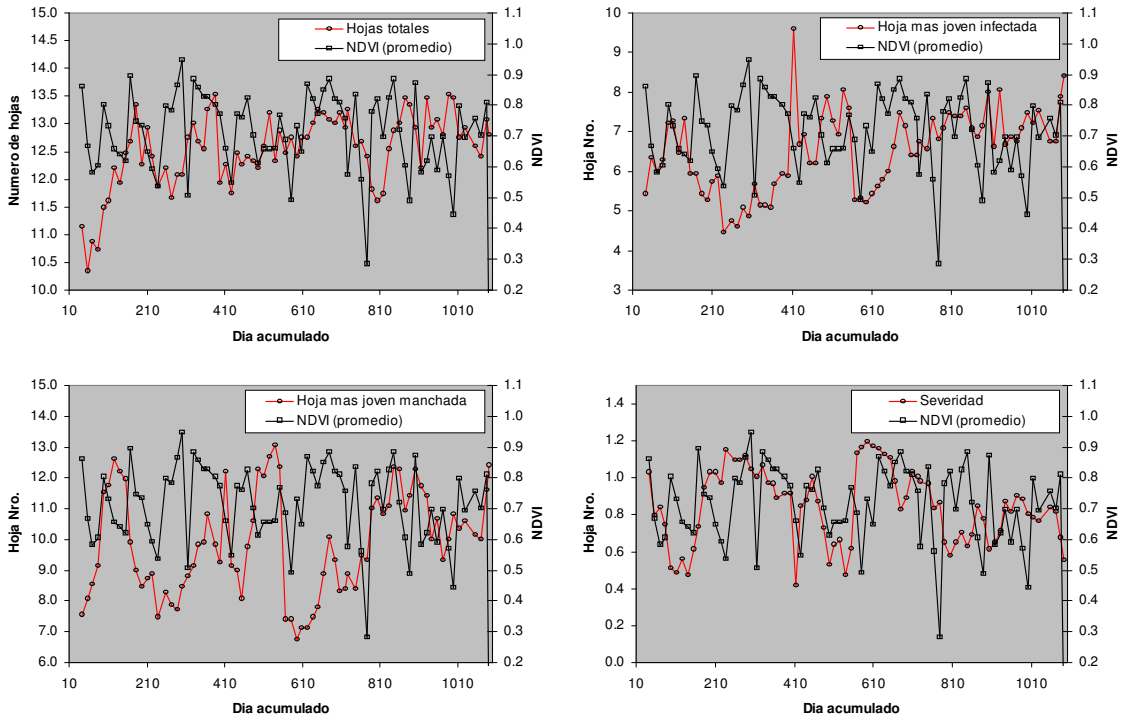

Figura 11 Comportamiento histórico de las cuatro variables de ataque de la enfermedad y el NDVI promedio

Todos los gráficos aquí suministrados se encuentran también en el archivo de Excel "matriz-de-datos-resultados.xls", adjunto al presente documento. La hoja "Graficos" contiene los gráficos presentados en este documento, y algunos adicionales.

d. Regresiones *stepwise* usando *bootstrapping* de muestras: Modelos de predicción, evaluación y precisión

Con las correlaciones estadísticamente significativas al nivel 0.05 o menor (negrilla) de la Tabla 1, para cada una de las 4 variables dependientes, se realizaron regresiones multivariadas con método de selección de variables por pasos hacia ambos lados (*stepwise*). Se desarrolló un modelo completo con todos los datos y se realizaron 100 iteraciones con una sub-selección de datos del 80%, y el restante 20% se usó para validación. Este procedimiento se conoce como *bootstrapping*, y es útil cuando se desea incrementar la confiabilidad de los modelos estadísticos y matemáticos desarrollados.

i. Número de hojas totales (NHT)

Para el NHT se usaron un total de 21 variables independientes para el ajuste, de las que 11 resultaron siendo relevantes según el procedimiento stepwise cuando se usaron todos

los datos. Al usar sub-muestras de los datos (*bootstrapping*), hubo diferente número de variables seleccionadas. El número de variables seleccionado varió entre 5 y 16, aunque el promedio, la mediana (dato central) y la moda (dato más frecuente) todos coincidieron en 11 variables. El número óptimo de variables, por tanto, para la predicción del NHT es 11. Dependiendo de la variable y su importancia para el modelo (grado en el que se encuentra relacionada con la variable dependiente), el procedimiento *stepwise* la seleccionó un número diferente de veces (Figura 12). Sólo dos variables fueron seleccionadas en todas las oportunidades: PLOG y NDVINLOG (ver Tabla 1 para IDs de las variables). Ambas presentaron muy bajos valores p promedio, lo que indica una altísima significancia estadística, y por tanto un alto porcentaje de explicación de la varianza de la variable dependiente. Las variables que menor cantidad de veces fueron seleccionadas fueron HRX2, T2, TRMMLOG.

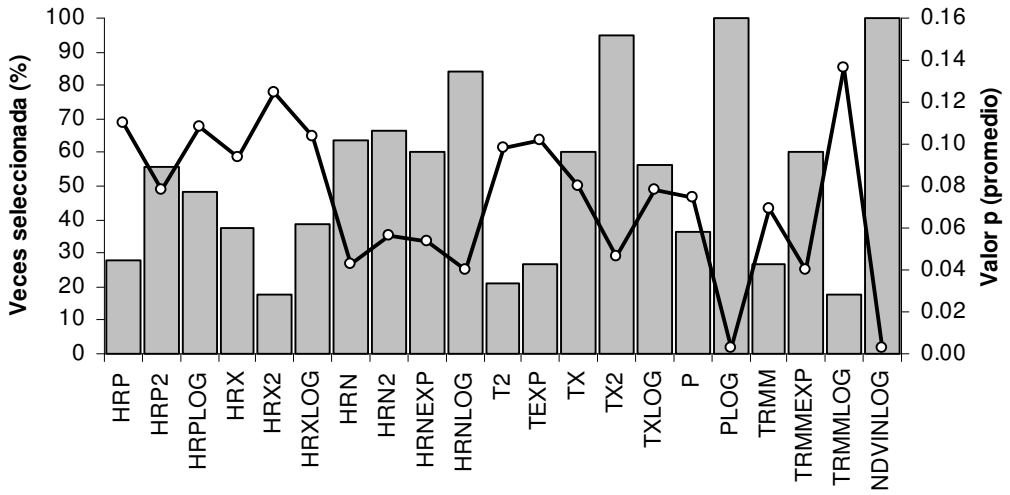

Figura 12 Porcentaje de veces (barras) que cada variable fue seleccionada como importante en la regresión multivariada, y valor p (significancia estadística, línea negra) promedio de cada variable sobre las 100 repeticiones o *bootstraps*.

Las variables que fueron seleccionadas en muy pocas ocasiones, generalmente presentaron baja significancia estadística (p>=0.1), lo que indica que una relativamente alta correlación no siempre es indicadora de alta significancia estadística para el modelo. De la misma manera se evaluó el coeficiente de determinación R^2, tanto para los datos de ajuste (entrenamiento, 80% de los datos de cada iteración), como para los datos de validación (20% de los datos de cada iteración). Usando los datos de prueba, se encontró que el mínimo R-cuadrado fue de 0, mientras que el máximo fue de 0.81 (Figura 13). En la mayoría de los casos, el R-cuadrado estuvo entre 0.3 y 0.6 (entre 54 y 77% de explicación de la varianza en la variable dependiente), lo que resulta en un considerablemente buen desempeño del algoritmo, aún sobre los datos de prueba. Los coeficientes de determinación de las regresiones stepwise (con menos variables) fueron ligeramente menores a los R-cuadrado desarrollados con todas las variables.

Por parte de los datos de ajuste (entrenamiento de los modelos), los coeficientes R-cuadrado variaron entre 0.5 y 0.85 (Figura 14), siendo los del procedimiento stepwise ligeramente menores. Una mayor cantidad de variables aumenta la correlación, pero también incrementa el riesgo de sobre-estimación (*overfitting*) en los modelos, por esto una reducción de variables resulta siendo de gran utilidad.

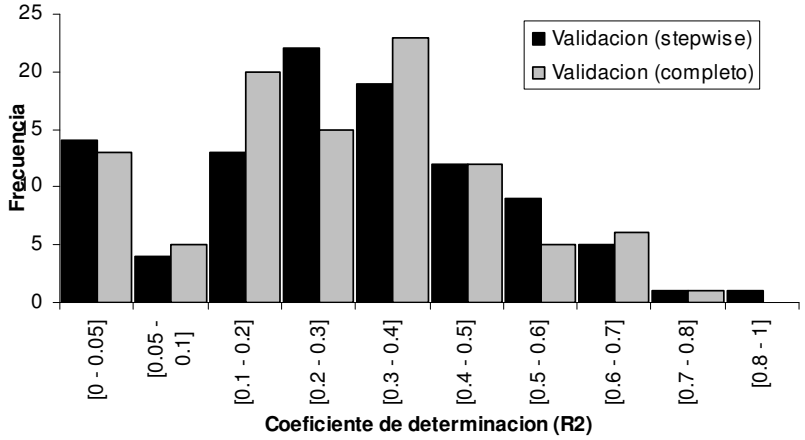

Figura 13 Histograma del coeficiente de determinación para el total de 100 sub-muestras (*bootstrap samples*), sobre los datos de validación para los modelos usando todas las variables (barras grises) y solo las que fueron seleccionadas con el procedimiento stepwise (barras negras)

En la mayoría de las iteraciones, el R-cuadrado estuvo entre 0.6 y 0.75 (70 y 85% de explicación de la varianza), lo que indica, dado el número de grados de libertad de la muestra (67, n=69), que de una manera bastante precisa, las variaciones temporales en el ataque de Sigatoka negra.

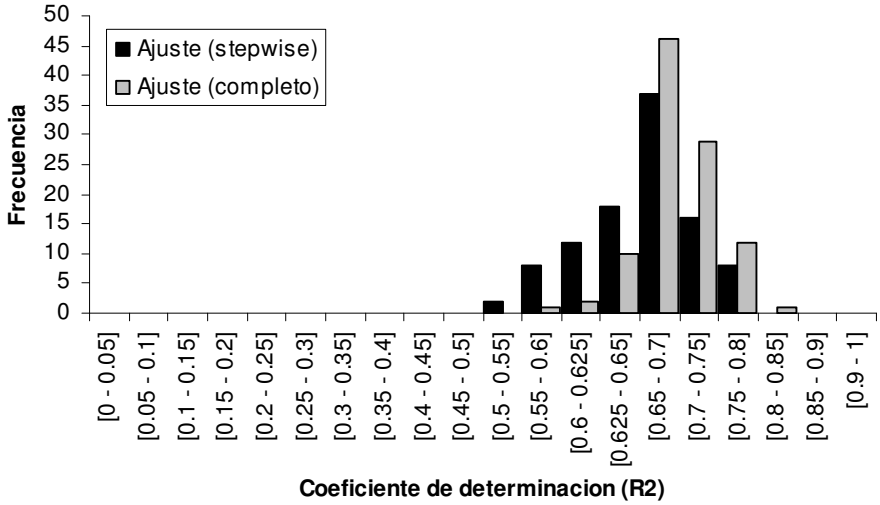

Figura 14 Histograma del coeficiente de determinación para el total de 100 sub-muestras (*bootstrap samples*), sobre los datos de ajuste (entrenamiento) para los modelos usando

todas las variables (barras grises) y solo las que fueron seleccionadas con el procedimiento stepwise (barras negras)

Todos los modelos desarrollados se proveen en la hoja de cálculo de Excel "**NHT-regresiones.xls**". Existen dos opciones para implementar estas regresiones en el pronóstico de la enfermedad. La primera es usando todos los modelos posibles (*stepwise*), en la página "**NHT-Estimates**", los que en la primera columna están marcados como "**STEP**", y obteniendo de ellos un resultado que se promedia y a su vez se generan también intervalos de confianza para la predicción. La segunda opción es usar el primer modelo desarrollado y dejar los demás solo como parte de la evaluación de precisión. En este último caso, el modelo lineal con el que debe calcularse el número de hojas totales de la planta es:

$$NHT = 220.042 - 0.004 * \left(HRP_0 - \overline{HRP}\right)^2 - 0.01835 * \left(HRN_1 - \overline{HRN}\right)^2 + 0.607 * e^{\left(HRN_1 - \overline{HRN}\right)}$$
$$-13.12 * Log\left(HRN_1\right) - 0.0293 * \left(T_0 - \overline{T}\right)^2 + 2.05 * TX_3 - 169.11 * Log\left(TX_3\right) + 0.507 * Log\left(P_1\right)$$
$$+ 0.0553 * e^{TRMM_4 - \overline{TRMM}} + 1.816 * Log\left(NDVIN_2\right)$$

Este modelo fue desarrollado con 69 parejas de datos, y presentó un R-cuadrado de 0.6341 (81% de explicación de la varianza) (Figura 15). Los sub-índices indican el desfase que se debe aplicar a los datos para predecir la variable dependiente (0 es ningún desfase, y 4 es 4 fechas de desfase).

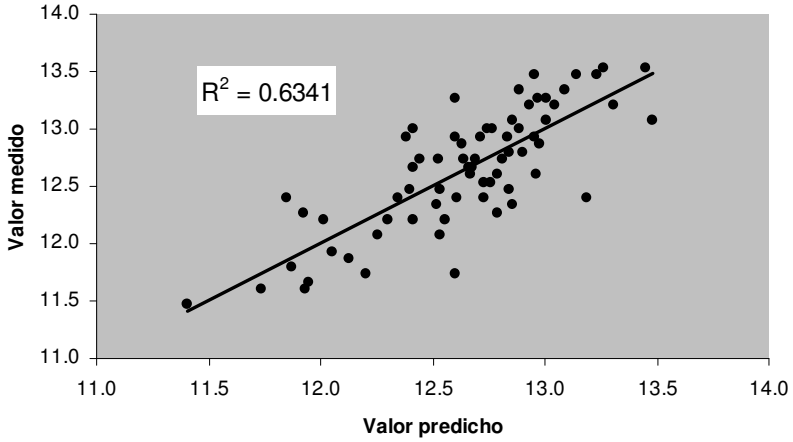

Figura 15 Comparación entre el valor predicho por el modelo y el valor medido de número de hojas totales

El modelo funciona para predecir el comportamiento hacia futuro de la enfermedad con al menos 80% de confiabilidad.

ii. Hoja más joven infectada (HMJI)

Para la HMJI se usaron un total de 24 variables independientes para el ajuste, de las que 11 resultaron siendo relevantes al nivel p=0.05. Las variables seleccionadas

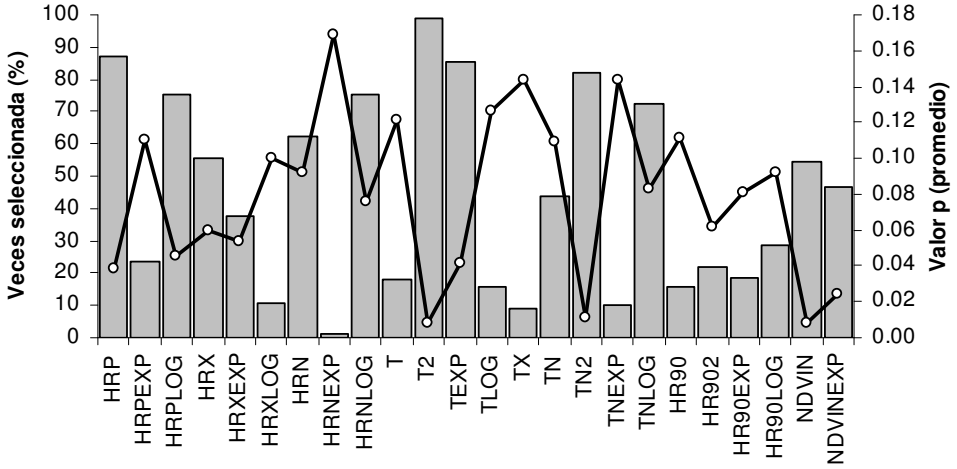

Figura 15 Porcentaje de veces (barras) que cada variable fue seleccionada como importante en la regresión multivariada, y valor p (significancia estadística, línea negra) promedio de cada variable sobre las 100 repeticiones o *bootstraps*.

Las variables que con mayor frecuencia fueron seleccionadas fueron T2, TEXP, TN2, TNLOG, HRN, HRP. Estas variables también mostraron un muy bajo valor p, lo que indica una alta significancia estadística. El coeficiente de determinación a través de las sub-muestras presentó una variación significativa para los datos de validación (valores entre 0 y 1, aunque la mayor cantidad de muestras estuvo concentrada en el rango 0.3-0.6, lo que indica un buen desempeño del algoritmo para predecir esta variable (Figura 16).

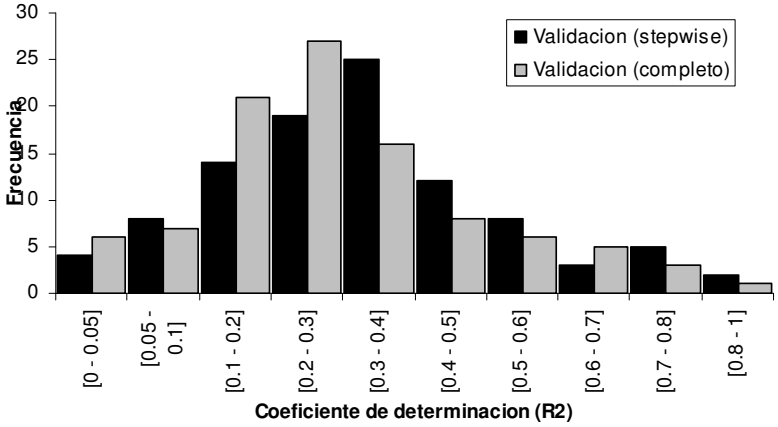

Figura 16 Histograma del coeficiente de determinación para el total de 100 sub-muestras (*bootstrap samples*), sobre los datos de validación para los modelos usando todas las variables (barras grises) y solo las que fueron seleccionadas con el procedimiento stepwise (barras negras)

En los datos de entrenamiento del modelo, se encontraron más altos coeficientes de determinación R-cuadrado (Figura 17). La mayoría de las sub-muestras presentaron coeficientes de determinación por encima de 0.6 (75% de explicación de varianza), indicando una alta precisión del algoritmo. La predicción de la hoja más joven infectada, por tanto, es posible.

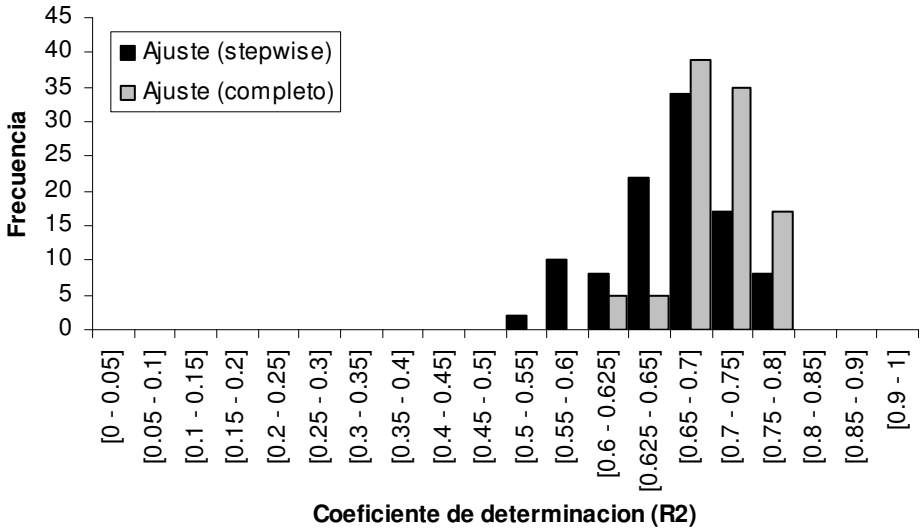

Figura 17 Histograma del coeficiente de determinación para el total de 100 sub-muestras (*bootstrap samples*), sobre los datos de ajuste (entrenamiento) para los modelos usando todas las variables (barras grises) y solo las que fueron seleccionadas con el procedimiento stepwise (barras negras)

Los modelos se encuentran en el archivo de Excel "HMJI-regresiones.xls", en la página "HMJI-Estimates". Aquellos marcados con "STEP", son aquellos que solo tienen variables estadísticamente significativas. El modelo final para predicción de la HMJI fue:

$$HMJI = -493.144 - 1.23 * HRP_4 + 201.17 * Log(HRP_4) - 0.1076 * HRX_1 + 0.377 * HRN_3$$
$$- 27.95 * Log(HRN_3) - 0.41 * (T_1 - \overline{T})^2 + 0.402 * e^{\left(\frac{T_1 - \overline{T}}{\sigma}\right)} - 5.06 * TN_2 + 0.272 * (TN_1 - \overline{TN})^2$$
$$+ 271.25 * Log(TN_2) + 2.022 * NDVIN_4$$

Este modelo presentó un R-cuadrado de 0.635 (Figura 18), y se desarrolló con un total de 71 parejas de datos. Nuevamente, el desfase de cada variable se indica como subíndice. La implementación de este modelo debe hacerse con el uso de la base de datos de clima, para cada una de las estaciones meteorológicas, o incluso, para los datos de campo de pluviómetros. Con este modelo puede detectarse el comportamiento hacia futuro y predecir futuros ataques.

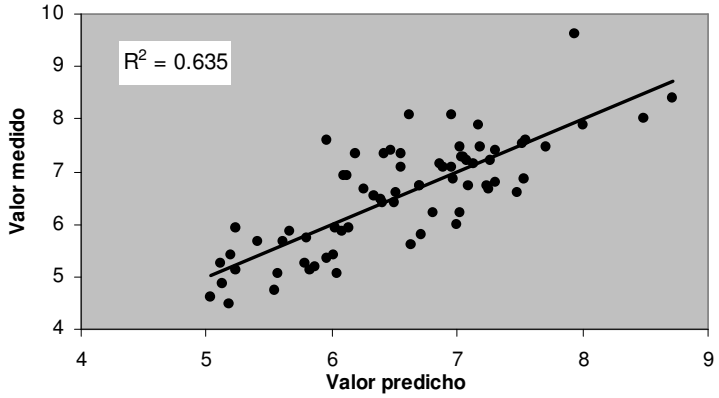

Figura 18 Comparación entre el valor predicho por el modelo y el valor medido de hoja más joven infectada

Cabe destacar que el modelo, aunque se desempeña bien para el set de datos usados, no está construido para extrapolar datos, y por tanto debe aplicarse siempre y cuando los datos estén dentro de los rangos establecidos durante el desarrollo de los modelos.

iii. Hoja más joven manchada (HMJM)

Para la HMJM se usaron 22 variables independientes, de las que 10 variables resultaron estadísticamente significativas (Figura 19). Las variables que con mayor frecuencia resultaron seleccionadas fueron la HRXEXP, la HRNLOG, la T, la TLOG, la TRMMLOG y la TRMM2.

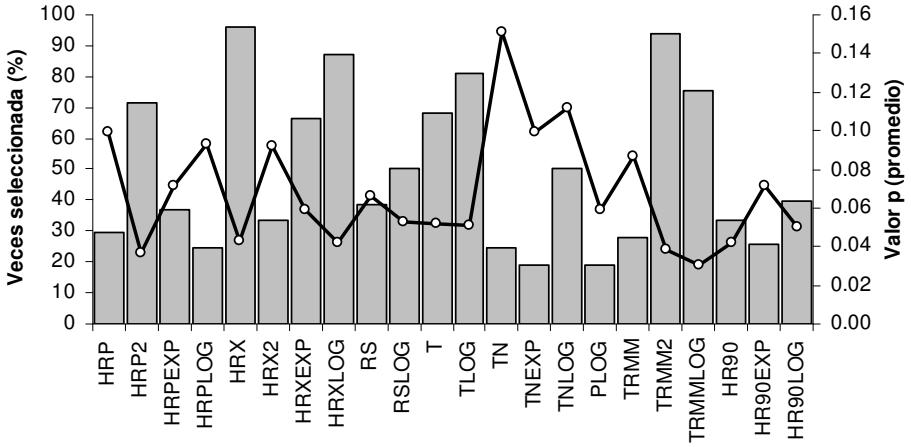

Figura 19 Porcentaje de veces (barras) que cada variable fue seleccionada como importante en la regresión multivariada, y valor p (significancia estadística, línea negra) promedio de cada variable sobre las 100 repeticiones o *bootstraps*.

Todas estas variables tienen una muy alta correlación con la enfermedad, según la literatura. La distribución de frecuencias del R-cuadrado mostró que la mayoría de las repeticiones tuvieron un R-cuadrado de validación (Figura 20) por encima de 0.3, lo que indica que la mayoría de los modelos, a ciegas están evaluando correctamente sobre la variable de interés HMJM.

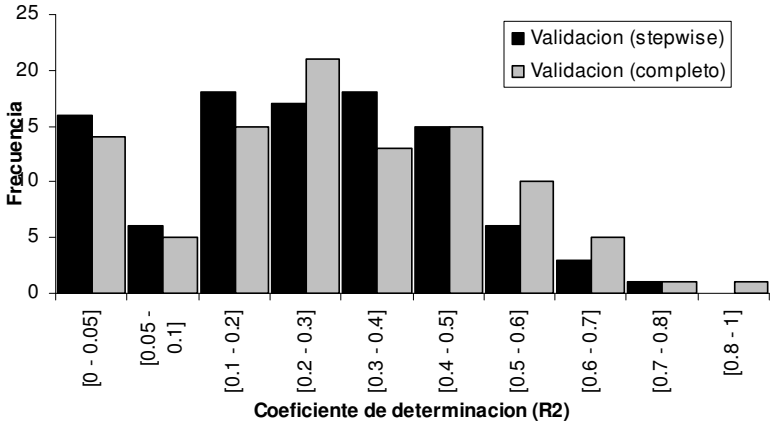

Figura 20 Histograma del coeficiente de determinación para el total de 100 sub-muestras (*bootstrap samples*), sobre los datos de validación para los modelos usando todas las variables (barras grises) y solo las que fueron seleccionadas con el procedimiento stepwise (barras negras)

Los datos de ajuste mostraron un coeficiente de determinación mucho más alto que los datos de validación (Figura 21). En general por encima de 0.5 (60% de explicación de la varianza). El ajuste de los datos al modelo fue bastante bueno en la mayoría de las repeticiones realizadas.

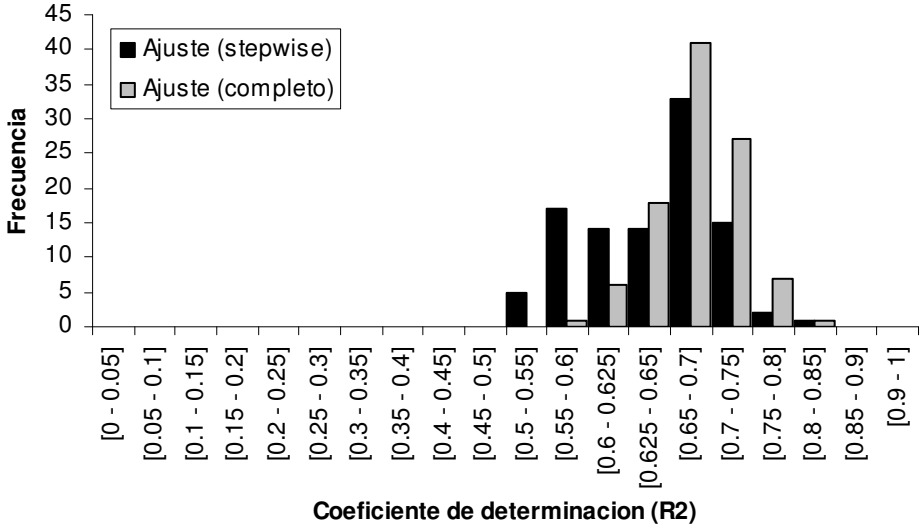

Figura 21 Histograma del coeficiente de determinación para el total de 100 sub-muestras (*bootstrap samples*), sobre los datos de ajuste (entrenamiento) para los modelos usando

todas las variables (barras grises) y solo las que fueron seleccionadas con el procedimiento stepwise (barras negras)

El modelo final, como se ha dicho para las demás variables dependientes (i.e. NHT, HMJI), puede implementarse mediante la aplicación de los 100 modelos de las submuestras y presentando una medida de incertidumbre (rango). Para esto, las ecuaciones a aplicar están en el archivo "HMJM-regresiones.xls", en la hoja "HMJM-Estimates". Sin embargo, de manera alternativa se puede aplicar la ecuación del modelo desarrollado usando todos los datos, y el procedimiento *stepwise*.

$$HMJM = -17355.7 + 0.01143 * \left(HRP_1 - \overline{HRP}\right)^2 - 50.6 * HRX_0 + 3.418 * e^{\left(\frac{HRX_0 - \overline{HRX}}{\sigma}\right)}$$
$$+ 10933.6 * Log\left(HRX_0\right) + 2.798 * Log\left(RS_3\right) - 8.218 * T_2 + 531.48 * Log\left(T_2\right)$$
$$+ 5.8E^{-5} * \left(TRMM_2 - \overline{TRMM}\right)^2 - 1.174 * Log\left(TRMM_4\right) - 1.095 * Log\left(HR90_4\right)$$

Este modelo presentó un R-cuadrado de 0.603 (70% de explicación de la varianza), y salvo algunos casos aislados, predice con alta precisión comportamiento de la variable dependiente HMJM.

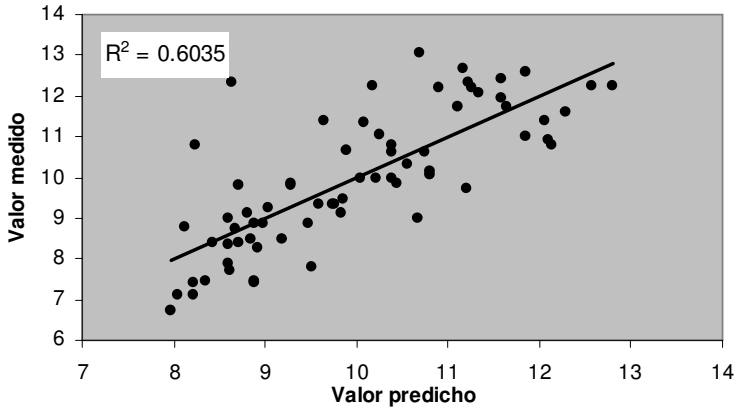

Figura 22 Comparación entre el valor predicho por el modelo y el valor medido de hoja más joven manchada

Se sugiere que el modelo sea probado en algunas otras áreas de la finca, con datos complementarios climáticos. Se constata que funciona bastante bien para el lote seleccionado, que es un lote con muy buena disponibilidad de datos. Probablemente el modelo es aplicable a otros lotes de la misma zona agroecológica (Zona 6), pero esto debería ser probado en un futuro, en tanto se tengan datos climáticos suficientes para tal fin.

iv. Severidad

Para la SEV se usaron 18 variables independientes, de las que 11 resultaron siendo estadísticamente significativas y muy altamente relacionadas con el comportamiento histórico de la Sigatoka negra. El promedio de variables incluidas en el modelo fue de 10, aunque los valores variaron entre 6 y 10, a través de las 100 repeticiones o *bootstraps* (Figura 23). El porcentaje de veces que cada variable fue seleccionada fue muy variable a través de las mismas, y aunque hubo casos en los que las variables presentaron muy baja significancia estadística, en otras ocasiones, las variables fueron seleccionadas más del 90% de las veces, indicando una muy alta significancia y relación con la variable dependiente

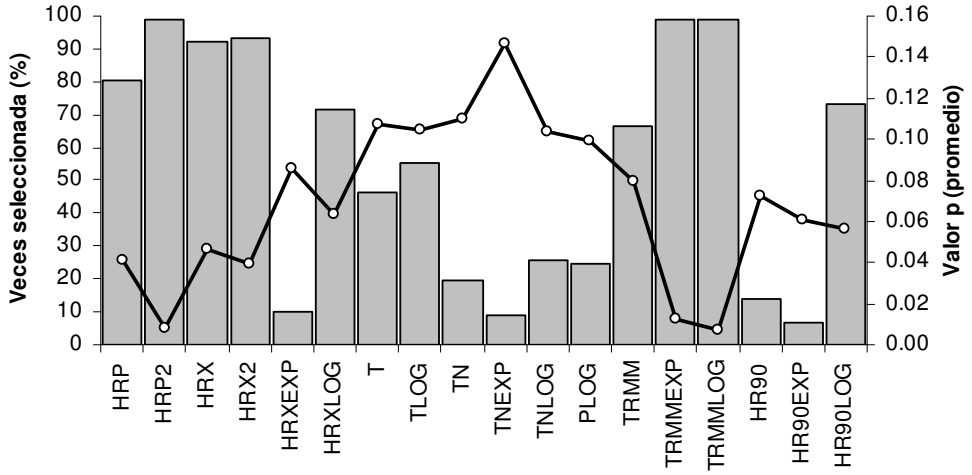

Figura 23 Porcentaje de veces (barras) que cada variable fue seleccionada como importante en la regresión multivariada, y valor p (significancia estadística, línea negra) promedio de cada variable sobre las 100 repeticiones o *bootstraps*.

El coeficiente de determinación (R2) tanto para los datos de validación (Figura 24) como para los datos de ajuste o entrenamiento (Figura 25), fue bastante alto en la mayoría de los casos, indicando un muy buen ajuste entre la predicción y los datos medidos en campo.

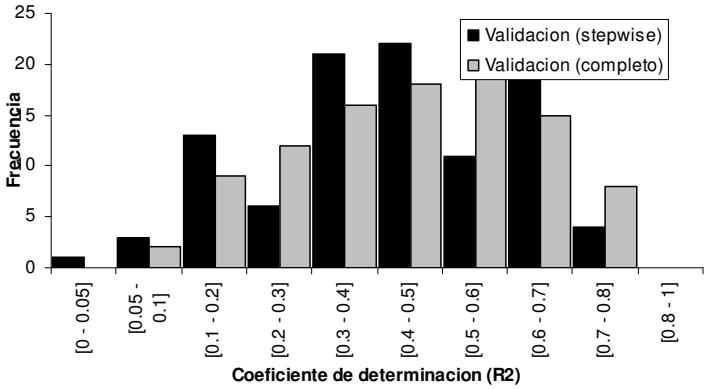

Figura 24 Histograma del coeficiente de determinación para el total de 100 sub-muestras (*bootstrap samples*), sobre los datos de validación para los modelos usando todas las variables (barras grises) y solo las que fueron seleccionadas con el procedimiento stepwise (barras negras)

Similar a las otras variables, los coeficientes de determinación en la validación estuvieron en su mayoría por encima de 0.3, mientras que en el ajuste estuvieron por encima de 0.55. Esto, en general indica que la explicación de la variable dependiente no sólo es consistente a través de las repeticiones, sino que es lo suficientemente precisa como para ajustar datos no usados e el entrenamiento (datos de validación, 20% aleatorio sobre los datos iniciales).

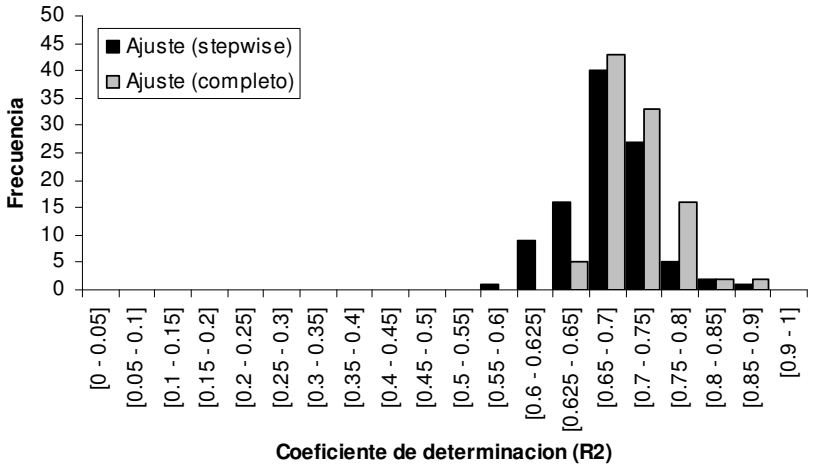

Figura 25 Histograma del coeficiente de determinación para el total de 100 sub-muestras (*bootstrap samples*), sobre los datos de ajuste (entrenamiento) para los modelos usando todas las variables (barras grises) y solo las que fueron seleccionadas con el procedimiento stepwise (barras negras)

Sería deseable contar con un set de datos históricos de mayor duración, para lograr validar de una manera mucho más robusta las predicciones. No obstante esto, la validación realizada (rigurosamente) en este estudio, muestra que la variable dependiente puede predecirse con bastante precisión en muchos casos. Para este caso, se podrá desarrollar un modelo que predice para más áreas el comportamiento de la enfermedad. El modelo final, entonces, puede ser aplicado usando los datos históricos con el desfase indicado en cada variable.

$$SEV = 463.07 + 0.0134 * HRP_4 - 0.0022 * \left(HRP_o - \overline{HRP}\right) + 1.203 * HRX_0 - 0.0075 * \left(HRX_2 - \overline{HRX}\right)^2$$
$$- 261.32 * Log\left(HRX_0\right) + 0.921 * T_2 - 59.88 * Log\left(T_2\right) - 0.00108 * TRMM_4$$
$$- 8.5E^{-6} * \left(TRMM_2 - \overline{TRMM}\right)^2 + 0.256 * Log\left(TRMM_4\right) - 0.1636 * Log\left(HR90_3\right)$$

Este modelo presentó un R-cuadrado de 0.666 (Figura 26), y fue realizado con la totalidad de los datos disponibles para tal fin (69). El modelo puede ser validado con

otras zonas, para verificar su consistencia, no obstante, tal como se observa en la Figura 26, representa con bastante precisión el comportamiento de la enfermedad a través del tiempo para el lote seleccionado (Zona agroecológica 6). Este modelo, junto con todos los demás (de las 100 repeticiones realizadas) se encuentran en el archivo "SEV-regresiones.xls", página "SEV-Estimates".

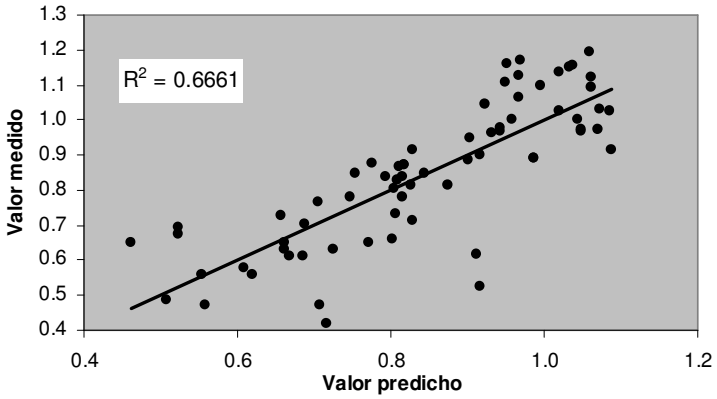

Figura 26 Comparación entre el valor predicho por el modelo y el valor medido de severidad

Todos los modelos desarrollados y presentados en el presente documento hacen parte de un estudio profundo de las relaciones entre las variables ambientales y las variables que miden el ataque de la Sigatoka negra. Se encontraron modelos multivariados que logran explicar el comportamiento de las 4 variables analizadas (NHT, HMJI, HMJM, y Severidad) en más del 70% de su variabilidad, con 67 grados de libertad, lo que indica una muy alta significancia estadística de los modelos.

Los modelos requieren validación adicional antes de ser implementados para otros lotes o zonas agroecológicas. Sin embargo, tal como se presentan aquí, sirven para pronosticar el ataque de la enfermedad en el Lote 1 de la finca Castilletes (zona agroecológica 6).

6. Efectos de la Sigatoka negra en producción

La Sigatoka Negra (*Mycosphaerella fijiensis* M.) tiene un efecto negativo en el desarrollo del banano, especialmente en la tasa de fotosíntesis (µmoles CO_2 reducidos por $m^2 \cdot s^{-1}$) y la tasa de transpiración foliar (Hidalgo et al, 2006). El porcentaje de severidad y el estadío de la enfermedad se correlacionan negativamente tanto con la tasa fotosintética como con la tasa de transpiración foliar (Hidalgo et al, 2006).

La Sigatoka negra (SN) es una de las enfermedades más destructivas que afectan los cultivos de banano y plátano en el mundo (Fullerton 1994; Fullerton & Stover 1990; Stover 1980). Causa manchas amarillas en la hoja que rápidamente se vuelven necróticas. La unión de estas manchas necróticas lleva a la destrucción total de la hoja (Du Mois

2003). Los ataquese de *M. fijiensis*, por lo tanto, llevan a una disminución de la capacidad fotosintética de la planta, y por tanto una pérdida del rendimiento bruto. En adición, la Sigatoka negra acelera la maduración del fruto resultando en una prematura maduración y por tanto pérdida del rendimiento exportable (Fouré 1994). Muchos estudios y autores indican que la presión de la Sigatoka negra depende de un amplio rango de factores ambientales (Valadares et al. 2007; Cordeiro et al. 2005; Guzmán 2003; Hernández et al. 2003; Pérez et al. 2000; Porras y Pérez. 1997; Smith et al. 1997; Gauhl 1994; Jiménez et al. 1994; Mobambo 1994; Capouzos et al. 1962). La Sigatoka negra y su variabilidad espacio-temporal es por tanto un factor crítico puesto que es la mayor amenaza para los cultivos de banano. En general, el desarrollo y prevalencia de la Sigatoka negra dependen de condiciones específicas de temperatura y humedad. Sitios con alta precipitación, con alta humedad relativa, y con un alto número de horas con humedad relativa por encima del 90% tienden a tener alta presión del patógeno. Áreas muy secas y/o muy calientes tienden a inhibir el desarrollo del patógeno.

Los genotipos susceptibles tienen un corto período de incubación, de evolución y de desarrollo de la enfermedad. Esto indica que una vez la SN infecta estos genotipos, las lesiones iniciales se convierten rápidamente en manchas necróticas, resultando en una extensiva muerte de la hoja, defoliación de la planta, corto período de llenado de fruto, reduciéndose significativamente el peso de racimo de estos genotipos (Craenen & Ortiz 1998).

Los cultivares del grupo Cavendish son susceptibles a la Sigatoka negra, motivo por el cual, en los cultivos de exportación, se hace necesario un manejo de la enfermedad. Este manejo, en la mayoría de los casos se hace mediante fungicidas, lo que implica un alto costo para el productor. El uso de estos fungicidas en algunos casos es indiscriminado, y por este motivo se hace necesario, como primera medida, conocer el comportamiento espacio-temporal de la enfermedad, y como segunda medida, conocer el efecto de dicho comportamiento en la producción y productividad de los genotipos.

El lote de la finca seleccionada para el presente estudio está cultivado con la variedad Gran Enano, y sometido a cierta presión de la enfermedad. Esta presión causa una disminución en el rendimiento potencial, por los argumentos anteriormente expuestos. En el presente estudio se analizó la variación temporal de la Sigatoka y su relación con el clima, pero de la misma manera se quiso analizar la variación histórica en la enfermedad y dos variables de producción: el número de racimos embolsados y el peso del racimo.

En general se puede observar (Figura 27) que el número de racimos embolsados (línea negra) es sensible a las variaciones en el ataque de la enfermedad (línea roja). Sin embargo, debe reconocerse que las variables de producción también están influenciadas por el clima, el manejo agronómico, el fungicida aplicado y por factores genéticos propios del genotipo (respuesta específica ante presión de la enfermedad, o presión ambiental), características de quien decide embolsar. Disminuciones en el número de

hojas totales tienden a generar una disminución en la cantidad de racimos embolsados, con un desfase de 2 a 4 fechas hacia delante respecto a la variable de producción.

Cuando la HMJM y la HMJI se hacen muy jóvenes, el número de racimos embolsados de 2 a 3 fechas (4 a 6 semanas) después tiende a disminuir. En ocasiones, esta disminución es de hasta 40% respecto a la fecha anterior. Lo contrario ocurre cuando la HMJI y la HMJM se hacen menos jóvenes, aunque la respuesta del genotipo no es tan rápida, probablemente porque el desarrollo del hongo continúa.

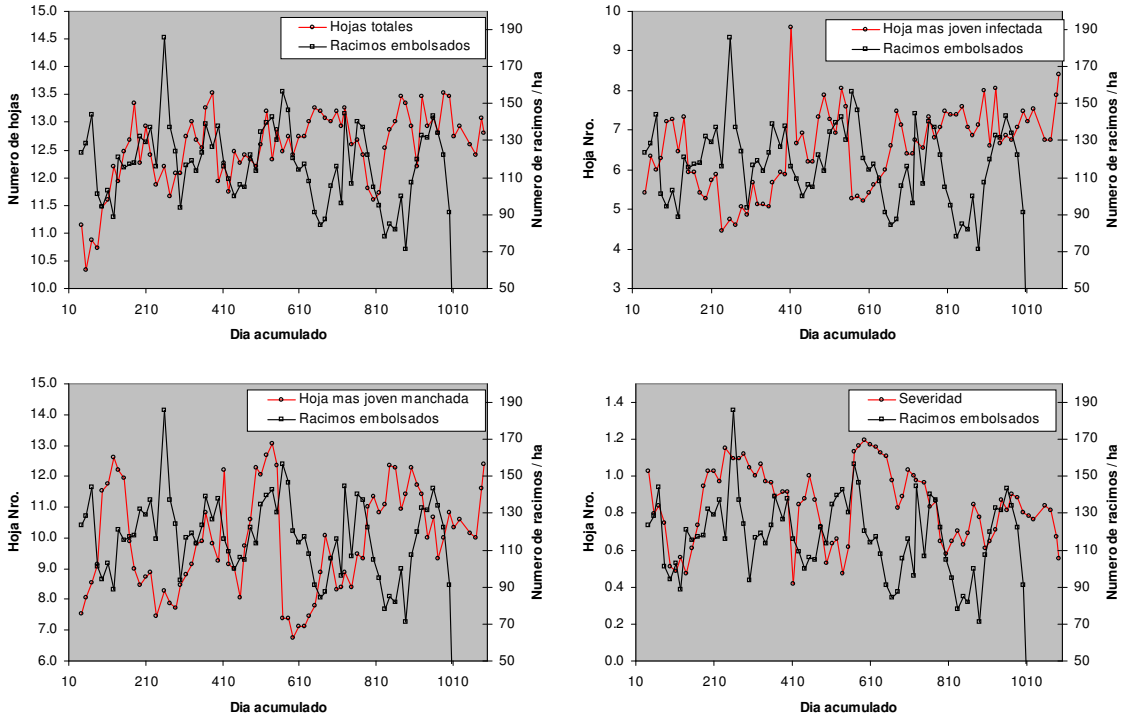

Figura 27 Variación temporal en el número de racimos embolsados y el ataque de Sigatoka negra, expresado como medida de cuatro variables (NHT, HMJI, HMJM y Severidad)

Estos comportamientos (rápida respuesta de la variable de producción) se observan especialmente durante la mitad del año 2008. El grado de afección de la producción por causa de la Sigatoka, sin embargo, aún puede ser indeterminado, porque el rendimiento es un factor complejo que depende de muchas variables (tal como se mencionó anteriormente). No obstante esta condición, se pueden detectar algunas tendencias en los datos, tanto en el peso del racimo como en el número de racimos embolsados existe una relación entre los extremos de ataque de la enfermedad y los extremos de caída de rendimientos.

Es probable que una vez infectada la planta, el desarrollo de la enfermedad se vea frenado por la aplicación de fungicidas, cuyo uso no fue tenido en cuenta en el presente estudio.

Se encuentra una relación, pero esta relación no se mantiene en el tiempo (después del día 610).

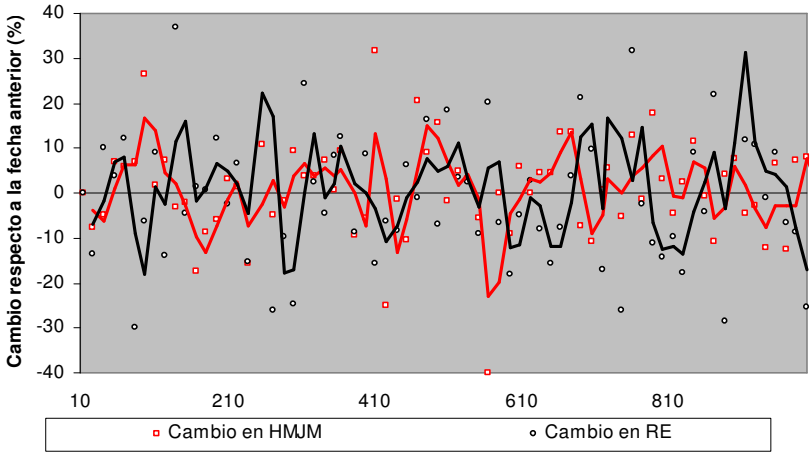

Figura 28 Tasa de cambio por cada fecha respecto a la fecha anterior, usando el promedio cada 2 fechas para suavizar las curvas, para la HMJM (línea roja) y el número de racimos embolsados (RE, línea negra)

Por parte del peso de racimo, se observa una tendencia estacional en el comportamiento de la producción, con el año 2007 teniendo en general bajo peso promedio, y los años subsiguientes valores al menos 50% más altos.

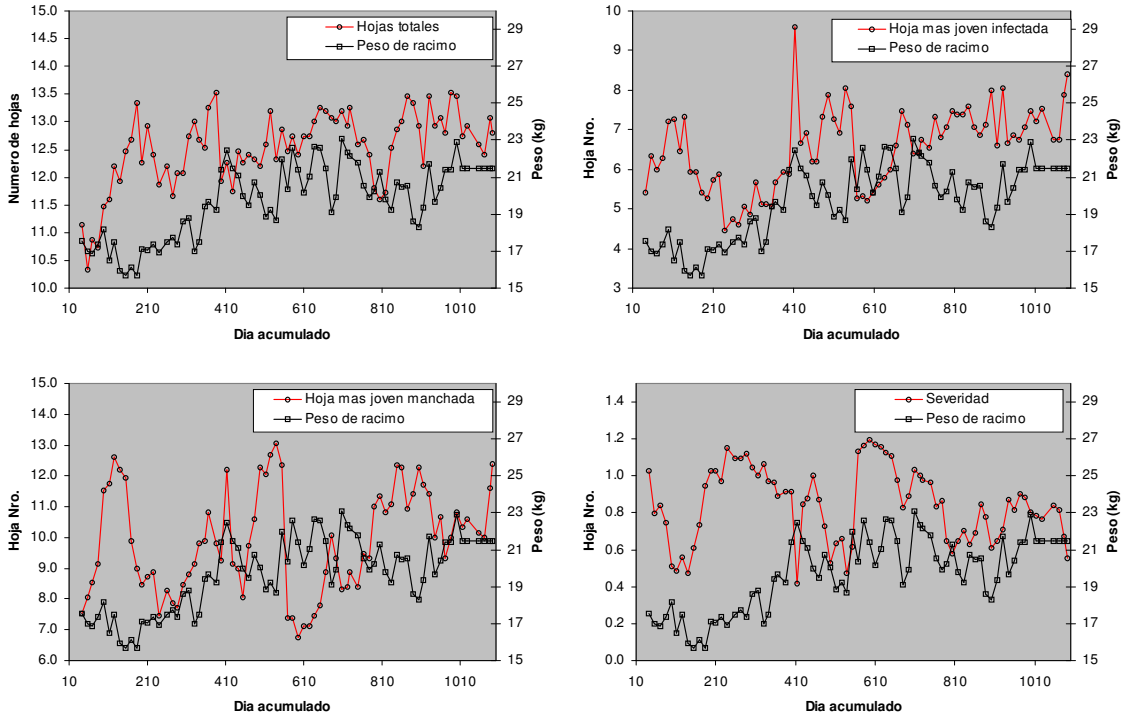

Figura 29 Variación temporal en el peso del racimo y el ataque de Sigatoka negra, expresado como medida de cuatro variables (NHT, HMJI, HMJM y Severidad)

Los períodos de mayor prevalencia de la enfermedad, al igual que con el número de racimos embolsados, resultan en una disminución del peso del racimo, aunque esta tendencia no se mantiene en el tiempo y su interpretación es confusa, probablemente debido a la influencia de otros factores (manejo, clima, entre otros)

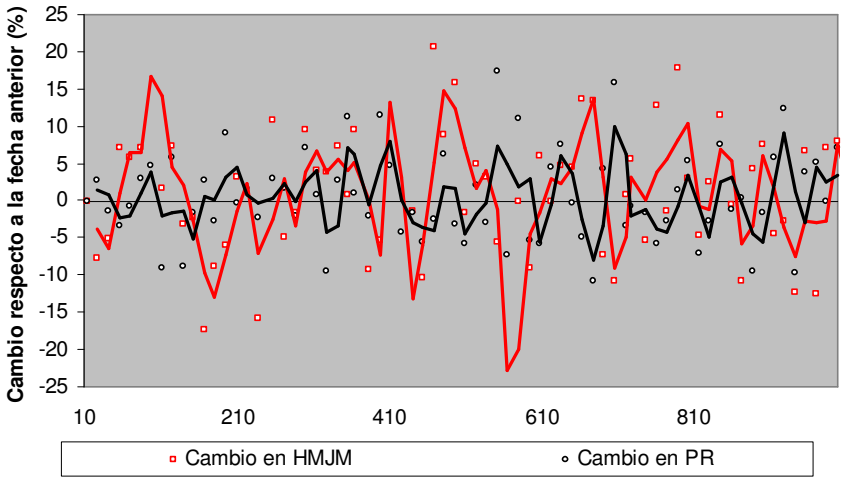

Figura 30 Tasa de cambio por cada fecha respecto a la fecha anterior, usando el promedio cada 2 fechas para suavizar las curvas, para la HMJM (línea roja) y el número de racimos embolsados (RE, línea negra)

Las disminuciones en el peso de racimo en algunos casos se observaron del orden de 15 a 20%, mientras que los incrementos, en la mayoría de los casos estuvieron por debajo del 10%. A partir de los indicios generales encontrados en el presente estudio, es posible detectar tendencias en la productividad y relacionarlas con las tendencias en la prevalencia y ataque de la enfermedad. De lo anterior se puede concluir que, aunque es complicado establecer un sistema de pronóstico de reducciones en la productividad, si se obtienen datos de manejo (aplicación de fungicidas, fertilización, entre otros), sería muy posible predecir tanto las afecciones del cultivo (ataques de la enfermedad), como sus efectos en la producción. En general, la disminución del peso del racimo se debe (en parte) a aumentos en la presión y prevalencia de la enfermedad, y esto se evidencia en las series de tiempo (con el debido desfase).

7. Conclusiones

En general, se ha desarrollado una metodología para predecir el ataque de la Sigatoka negra en banano Gran Enano, en un lote de la finca Castilletes, ubicado en la zona agroecológica 6. Esta metodología probó ser útil y suficientemente precisa cuando se evaluó usando el método conocido como *bootstrapping*, que permite evaluar todos los posibles comportamientos de un modelo aleatoriamente, mediante la selección de sub-muestras en los datos. Para cada una de 4 variables dependientes (número de hojas totales, hoja más joven infectada, hoja más joven manchada, y severidad) se encontró un

modelo lineal multivariado para predecir el ataque futuro de la enfermedad con 1 a 8 semanas de adelanto en el tiempo (dependiendo de la variable). Estos modelos mostraron ser consistentes en el tiempo y ser capaces de predecir la enfermedad con hasta 80% de confiabilidad. Hacia el futuro, se sugiere una validación extensiva de los mismos, así como un mejoramiento usando un set de datos mucho más completo.

Referencias

Capouzos, L; Theis, T; Colberg, C; Rivera, C. M; Santiago, A. 1962. Relationship between climatic factors and incidence of the Sigatoka leaf-spot disease of bananas. Plant disease Reporter 46:758-761.

Cohan JP, Abadie C, Tomekpé K, Tchango Tchango J (2003) Evaluación del desempeño agronómico y de la resistencia a la Sigatoka negra del híbrido de plátano 'CRBP-39'. Infomusa Vol. 12 No. 1. 29-32.

Cordeiro ZJM, Matos AP, Ferreira DMV, Abreu KCLM (2005) Manual para identificaçao e controle da Sigatoka-negra da bananeira. EMBRAPA Mandioca e Fruticultura Tropical: Cruz das Almas, 2005. 36p.

Craenen K, Ortiz R (2003) Genetic improvement for a sustainable management of resistance. In: Jacome L, et al (Eds.) *Mycosphaerella* leaf spot diseases of bananas: Present status and outlook. Proceedings of the 2nd International workshop on *Mycosphaerella* leaf spot diseases held in San José, Costa Rica, 20-23 May 2002.

Du Mois D (2003) Bananes for ever. Fruitrop. No. 99. February 2003.

Fouré E (1994) Leaf Spot Diseases of Banana and Plantain caused by Mycosphaerella musicola and M. fijiensis. In: Jones DR (ed) The improvement and testing of Musa: a Global Partnership. Proceedings of the first global conference of the International Musa testing program held at FHIA, Honduras 27-30 April 1994.

Fullerton RA (1994) Sigatoka leaf diseases. In: . C. Ploetz, G. A. Zentmyer, W. T. Nishijima, K. G. Rohrbach and H. D. Ohr (eds) Compendium of Tropical Fruit Diseases. APS Press, St. Paul, MN, USA. pp. 12-14

Fullerton RA, Stover RH (eds.) (1990) Sigatoka leaf spot diseases of bananas. Proceedings of an International workshop held at San José, Cost Rica, 29.3-1.4.1989, INIBAP, Montpellier

Gauhl F (1994) Epidemiology and Ecology of Black Sigatoka (Mycosphaerella fijiensis Morelet) in Plantain and Banana (Musa spp.) in Costa Rica, Central America. PhD thesis originally presented in German. INIBAP, Montpellier, France. 120pp.

Guzmán M (2003) Epidemiología de Sigatoka negra y el sistema de preaviso biológico. Actas del taller "Manejo convencional y alternativo de la Sigatoka negra, nematodos y otras plagas asociadas al cultivo de Musáceas", celebrado en Guayaquil, Ecuador. 11-13 de agosto.

Hernández L, Hidalgo W, Linares B, Hernández J, Romero N, Fernández S (2003) Estudio preliminar de vigilancia y pronóstico para Sigatoka negra (*Mycosphaerella fijiensis* Morelet) en el cultivo de plátano (Musa AAB cv Hartón) en Managua-Jurimiquire, estado Yaracuy.

Hidalgo M, Tapia A, Rodríguez W, Serrano E (2005) Efecto de la Sigatoka Negra (Mycosphaerella fijiensis) sobre la fotosíntesis y la transpiración foliar del banano (Musa sp. AAA, cv. Valery). Agronomía Costarricense 30(1): 35-41.

Jiménez F, Tapia AC, Gribuis N, Escalant JV (1994) Relation entre la dure´e de la pluie et le developpement de la Cercosporiose noir sur le banane et le plantain. Proposition d'un syst´eme d'avertissement biometeorologique . Fruits 50, 87.

Mobambo KN (1994) Factores que influyen sobre el desarrollo de la Sigatoka negra en el plátano y en los híbridos de plátano. INFOMUSA: Vol 4 Nº 1.

Orjeda, G (ed) (2000) Evaluating bananas: a global Partnership. Results of IMTP Phase II. International Plant Genetic Resources Institute. International Network for the Improvement of Banana and Plantain pp. 15.

Orozco-Santos M, Farías-Larios J, Manzo-Sánchez G, Guzmán-González S (2001) Black Sigatoka disease (Mycosphaerella fijiensis Morelet) in México. INFOMUSA – Vol 10, Nº 1.

Pérez-Vicente L, Mauri-Mollera F, Hernández-Mancilla A, Abreu-Antúnez E, Porras-González A (2000) Epidemiología de la Sigatoka Negra (Mycosphaerella fijiensis Morelet) en Cuba. I. Pronóstico Bio-Climático de los tratamientos contra le enfermedad en Bananos (Musa acuminata spp. AAA). Revista Mexicana de Fitopatología. Volumen 18, Número 1.

Porras A, Pérez L (1997) The role of temperature in the growth of the germ tubes of ascospores of Mycosphaerella spp., responsible for leaf spot diseases of banana. Infomusa – vol 6, Nº2.

Smith MC, Rutter J, Burt PJA, Ramírez F, González EH (1997) Black Sigatoka disease of banana: spatial and temporal variability in disease development. Association of Applied Biologists.

Stover RH (1980) Sigatoka leaf spots in banana and plantain. Plant disease 64 750-755.

Valadares R, Cintra de Jesus W, Avelino R (2007) Influencia das mudanças climáticas na distribuiçao espacial da Mycosphaerella fijiensis no mundo. Anais XIII Simpósio Brasileiro de Sensoramiento Remoto, Forianópolis, Brasil, 21-26 abril 2007, INPE, p 443-447

Parte 4. Determinación del nivel freático histórico a través de la zona bananera de Urabá para establecimiento de mejor criterio para diseño de sistemas de drenaje

Resumen

El presente documento describe los materiales usados, métodos aplicados y resultados obtenidos por el Centro Internacional de Agricultura Tropical (CIAT) respecto a la determinación del nivel freático histórico (comportamiento) a través de la zona bananera del Urabá de cara al establecimiento de un mejor criterio para el diseño de sistemas de drenaje así como para la detección de problemas de anegamiento en los suelos, y de épocas donde hay un aumento considerable del nivel freático. El análisis consistió de 4 pasos básicos: (1) Organización e inventario de la información, (2) interpolación de superficies de precipitación y nivel freático, (3) mapeo de zonas problema, (4) observación del comportamiento histórico como herramienta para detección de problemas. Se contó con datos de pozos de observación y de pluviómetros en las fincas tomados directamente en campo durante las épocas 2006-2010. Estos datos se agregaron a nivel mensual y se interpolaron usando el algoritmo *Thin Plate Spline* (TPS) por medio del software R2.11.1 (librerías *fields* y *raster*). Aunque se encontraron algunas imprecisiones en las interpolaciones (especialmente en aquellas áreas que se encontraron fuera del rango donde se hallaban los pozos) reflejados en valores demasiado altos o demasiado pequeños (incluso negativos) en algunas áreas. Se lograron, sin embargo, crear superficies de precipitación y de nivel freático mensuales para los años 2007, 2008 y 2009. Usando estas superficies se produjeron algunos mapas de relevancia, que se presentan en el presente documento, así como también algunos ejemplos de comportamiento histórico de ambas variables. Se recomienda la continuada colección de datos de pozos y pluviómetros, incluyendo la georreferenciación de los mismos en la totalidad de fincas, puesto que con este tipo de información geográfica se logran optimizar los procesos de toma de decisiones.

Contenido

1. Introducción

La agricultura es una actividad considerablemente dependiente de una clima propicio para su adecuado desarrollo y por este motivo (sumado a la dificultad para predecir cualquier fenómeno climático), ocurren (en ocasiones) descensos sustanciales en la producción de bienes derivados de la agricultura. La actividad bananera no es la excepción. La zona del Urabá es una zona de baja altitud (entre 0 y 100 metros sobre el nivel del mar) donde prevalecen las altas precipitaciones (entre 2,500 y 3,500 mm/año) y donde la temperatura se mantiene en un rango entre 26 y 28 grados centígrados, como promedio. De la misma manera, a través de la región prevalece la alta humedad relativa (promedio) y sumado a esto, hay tanto muchos días con lluvia como muchas horas con alta humedad relativa durante el día. Los suelos en general tienen altos contenidos de arcilla y son relativamente profundos, aunque se hace necesario contar con sistemas de drenaje eficiente debido a las altas precipitaciones que se reciben en la región. En general, hay muy buena aptitud para la producción bananera en la región, empero hay algunos factores que causan pérdidas de producción y/o aumentan los costos. Dentro de estos se destacan la Sigatoka negra y algunos problemas de anegamiento de los suelos. El anegamiento en los suelos es un factor que, además de ser relativamente difícil de controlar, aumenta el riesgo de pérdidas productivas.

En el presente informe, se analizó el comportamiento mensual del nivel freático y la precipitación con base a medidas tomadas en campo en pluviómetros y pozos de observación. Durante el período 2007-2009 se generaron superficies de nivel freático y de precipitación con base en medidas de campo usando el algoritmo de interpolación *Thin Plate Spline* (TPS) (Hutchinson, 1984; Hutchinson & de Hoog, 1985). Usando estas superficies se generaron mapas de zonas problema para ciertas fechas del año como ejemplo, y se provee un mapa en formato MXD para visualización de los resultados. De la misma manera se analizó el comportamiento histórico del nivel freático y se creó una plantilla para análisis de los datos históricos. El análisis provee una línea base para la

detección de problemas de anegamiento en los suelos y por tanto para la definición de mejores sistemas de drenaje agrícola.

En general se encontró que el nivel freático en la mayor parte de la región osciló entre 50 y 150 centímetros, aunque durante la mayor parte del año y en la mayor parte del área se mantuvo por debajo de los 100 cm. Sin embargo, las interpolaciones fueron desarrolladas con datos sólo de unas pocas fincas, y además de eso, no todas las fincas presentaron datos en todas las fechas. Se recomienda un continuo trabajo de captura de datos así como de georreferenciación de lo demás pozos de observación y pluviómetros, para mejorar los resultados de las interpolaciones y de la misma manera las conclusiones de los análisis. El establecimiento de un equipo de trabajo en Sistemas de Información Geográfica (GIS) es fundamental para este tipo de tareas.

2. Datos de entrada

Los datos de entrada para el análisis aplicado fueron básicamente datos históricos de precipitación y nivel freático, tomados en campo en pluviómetros y pozos de observación respectivamente. Las mediciones fueron realizadas entre el 2006 y 2010, sin embargo, para las fincas en las que se realizó georreferenciación (Figura 1), sólo se logró trabajar para los años 2007 a 2009.

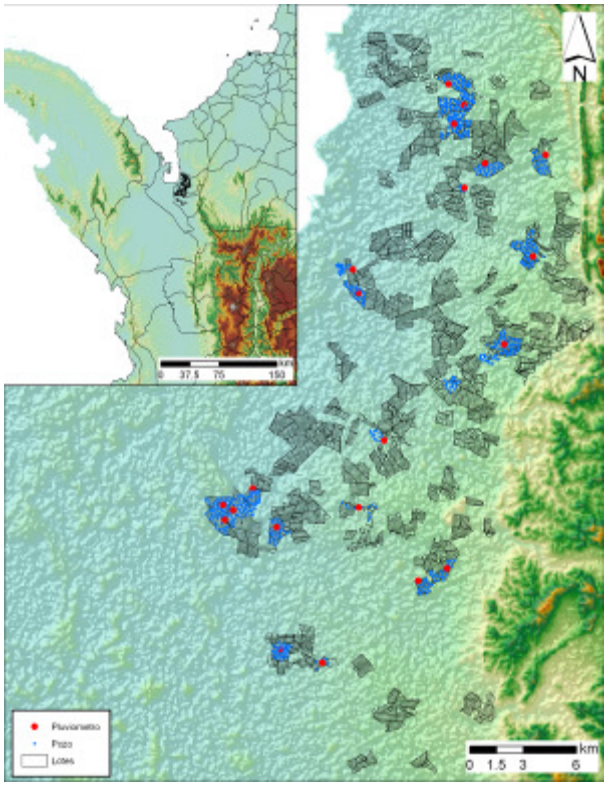

Figura 1 Distribución geográfica de pozos de observación (puntos azules) y pluviómetros (puntos rojos) cuyos datos se usaron en el análisis.

a. Datos de pluviómetros

Se contó con datos de 23 pluviómetros distribuidos de a uno a través de 23 de las 179 fincas que se encuentran georreferenciadas en la cartografía base provista por UNIBAN CI. La localización de los pluviómetros (puntos rojos, Figura 1) se encuentra en el archivo "pozos-pluviometros-UTM-18N.shp", que se encuentra en el folder "./nivel-freatico/datos-entrada/shapefiles". La columna "Tipo" indica si el punto en cuestión es un pozo o un pluviómetro.

Se contó con datos diarios de precipitación para cada uno de los 23 pluviómetros cubriendo desde el 2006 hasta mediados del 2010. La Figura 2 muestra un ejemplo con la finca Bahía. Los datos diarios se agregaron a nivel mensual, para incrementar el número de pluviómetros muestreados en una sola fecha, y así aumentar la confiabilidad de las superficies climáticas interpoladas.

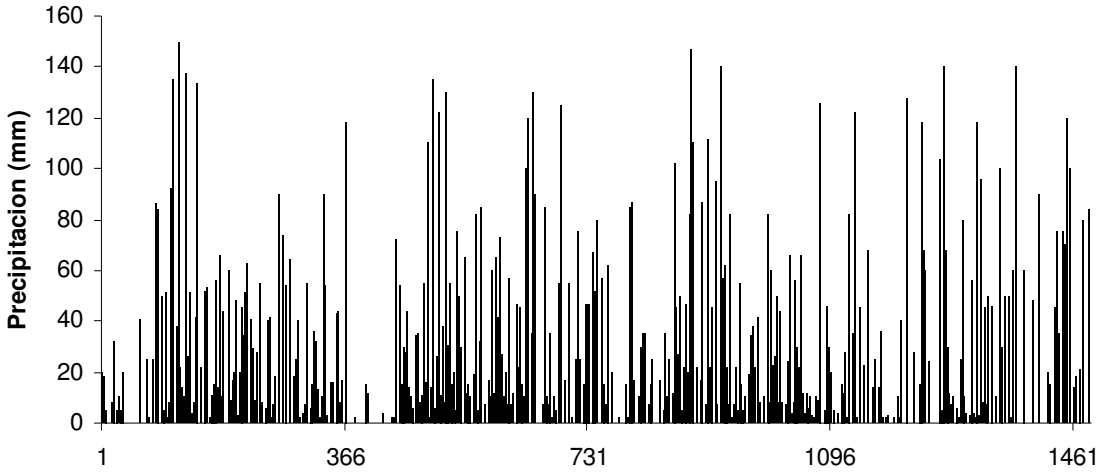

Figura 2 Comportamiento histórico de la precipitación en la finca Bahía entre los años 2006 y 2010

Como puede observarse (Figura 1), los pluviómetros están distribuidos adecuadamente a través de la región, aunque en algunas áreas existen vacíos, lo que ciertamente causa imprecisiones en las interpolaciones posteriores. Debido a esto, se recomienda que hacia el futuro se incluyan todos lo pluviómetros de la región en el análisis.

Los datos de pluviómetros se agregaron hasta el nivel mensual (suma de precipitación diaria) de tal manera que se logró tener un dato por cada mes para los 3 años con datos disponibles. Todos estos datos se encuentran en los archivos "ppt-original.xls" y "ppt-matrices.xls" en el folder "./nivel-freatico/datos-entrada/archivos-xls/".

b. Datos de pozos de observación

Los datos de pozos de observación provinieron de 726 pozos georreferenciadas en las mismas 23 fincas en las que se georreferenciaron los pluviómetros. Los datos de pozos de observación fueron colectados con una frecuencia de 2 semanas, lo que en general resultó en dos medidas por mes. La Figura 3 muestra el comportamiento del nivel freático histórico para los años 2007 a 2009, para el lote 14 de la finca Claudia Sofía.

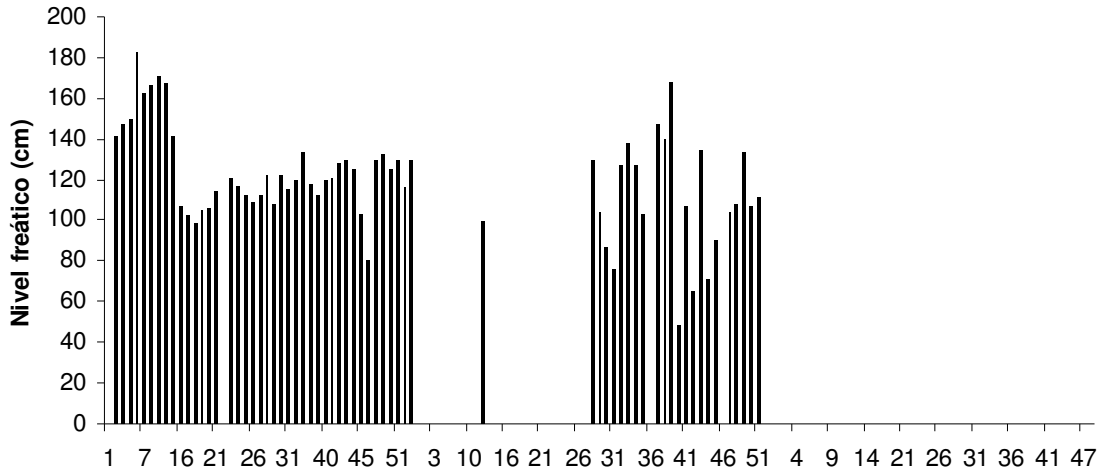

Figura 3 Comportamiento del nivel freático semanal para el pozo de observación de la finca Claudia Sofía, lote 14.

Al igual que los datos de precipitación de los pluviómetros, los datos de nivel freático se agregaron a nivel mensual, usando el promedio de los datos disponibles para cada mes. Estos datos se encuentran en los archivos Excel "nf-matrices.xls", "nf-original.xls", y "nf-original-organizado.xls", que se encuentran dentro del folder "./nivel-freatico/datos-entrada/archivos-xls/".

3. Organización de datos de entrada y observaciones generales

Como se dijo con anterioridad, se agregaron los datos diarios (caso de precipitación) y semanales (caso observaciones freatimétricas) hasta el nivel mensual para facilitar todos los análisis ulteriores. La Figura 4 muestra el comportamiento del pozo 1 de la finca Fragata y su respectivo pluviómetro, para los años 2007 a 2009, a nivel mensual. Como se puede observar, la precipitación tiene una alta variabilidad a través del anho, siendo el período Junio-Septiembre el más lluvioso del año, seguido por el período Octubre-Noviembre. El nivel freático presenta sus valores más bajos (es decir, es menos profundo) durante las épocas inmediatamente posteriores a aquellas de mayor precipitación, mostrando la influencia directa de la precipitación en el anegamiento del suelo. Esto también, claro está, depende del tipo de suelo, y de la variabilidad espacial de la precipitación.

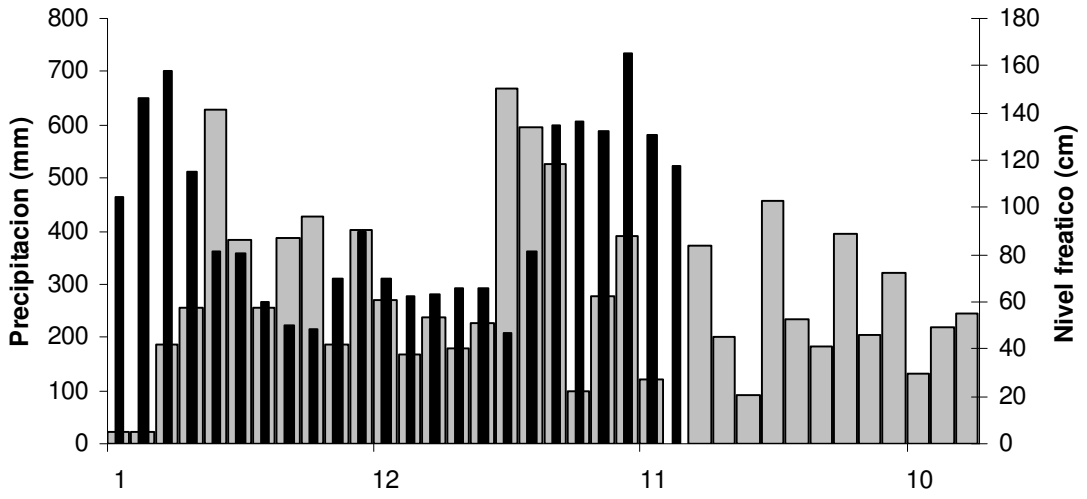

Figura 4 Comportamiento mensual de la precipitación (barras grises) y del nivel freático (barras negras)

Dado que no todos los pozos de observación tuvieron la misma disponibilidad de información, es decir, que en algunas fechas el número de pozos disponibles (con datos) fue considerablemente menor al número de pozos totales georreferenciados, se decidió implementar un algoritmo que lograra aumentar la cobertura espacial de las medidas. En este caso, se usó la interpolación de lecturas tanto de pluviómetros como de pozos usando el algoritmo *Thin Plate Spline* (TPS).

4. Metodología aplicada y resultados principales

La metodología aplicada consistió básicamente en la interpolación de superficies mensuales de precipitación y de nivel freático. Estas superficies se produjeron en formato raster (grillas), y con ellas se mapearon las zonas problema, y se realizó un análisis histórico del comportamiento tanto de la precipitación como del nivel freático.

a. Interpolación de superficies usando el DEM de 90 metros

Usando los valores correspondientes a cada uno los puntos (pozos o pluviómetros), y con el fin de obtener una superficie que describiese el gradiente espacial de cada una de las dos variables, se usó el método de interpolación *Thin Plate Spline* (TPS) (Hutchinson, 1984; Hutchinson & de Hoog, 1985) que ha sido empleado por diversos autores para ajustar la distribución de variables con muy alta varianza, o cuyas mediciones de campo tienen ruido. El método, como se aplicó a los datos recibidos, usa cada la precipitación y el nivel freático como variables dependientes, y ajusta su distribución a una función multivariada usando la altitud (en metros, derivada de SRTM, Jarvis et al. 2008), la latitud (unidades dependiendo del sistema de coordenadas) y la longitud (unidades

dependiendo del sistema de coordenadas). Una vez se encuentra la función de ajuste, dicha función se proyecta sobre toda el área bajo análisis, incluyendo aquellas áreas en las que no se han tomado datos. Debido a la limitada disponibilidad de datos en algunas de las fechas, el algoritmo tuvo un pobre desempeño en algunas áreas en esas fechas específicas, produciendo valores negativos durante esas fechas y por tanto reduciendo la aplicabilidad y escalabilidad del método. La única solución a esto sería

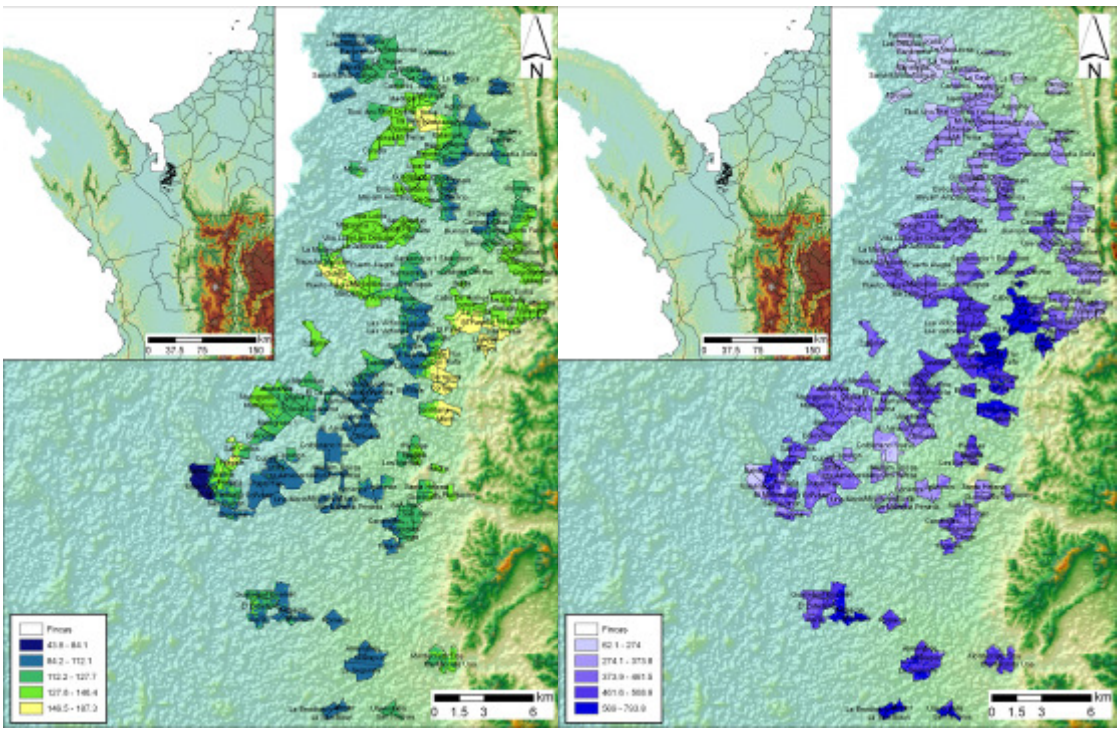

Figura 5 Resultados de la interpolación para el mes de Julio de 2008 para nivel freático (izquierda) y precipitación (derecha)

Idealmente, sin embargo, con la suficiente disponibilidad de datos, no sería necesario realizar interpolaciones. No obstante, la interpolación es un método que suele proveer información adicional en áreas pequeñas, con un riesgo relativamente pequeño de error. Esto, de una manera eficiente y rápida, permite realizar análisis que los datos de campo *per se* no permiten realizar.

Para cada una de las dos variables bajo análisis, se ajustó una función TPS, y se calculó una superficie de distribución de la variable a través de la geografía de la región. Este proceso se realizó usando el software R (http://www.r-project.org), con las librerías *fields*, *spam*, *rgdal*, *sp* y *raster*. Tanto R, como las librerías en cuestión pueden descargarse de la página http://www.r-project.org. Para más información sobre instalación, notas y librerías de R, referirse al manual de este software. Este proceso de ajuste puede ser realizado con el mismo script que entregado en el primer reporte (Zonificación Agroecológica).

La interpolación de los datos en cuestión permitió determinar de una manera aproximada, los patrones de distribución espacial de las dos variables mensualmente (Figura 5), y que probablemente influyen la producción de las diferentes fincas. Además de esto, permitió la obtención de superficies continuas a través de las diferentes fincas (incluyendo aquellas que no presentaron datos) con las que se pueden detectar los problemas en campo de una manera más eficiente, y de la misma manera, a la postre, aplicar los correctivos necesarios.

En los folders "./nivel-freatico/nivel-freatico/interpolaciones" y "./nivel-freatico/nivel-freatico/interpolaciones-corte-fincas" se encuentran los rasters en dos sistemas de coordenadas: Universal Transverse Mercator (UTM) (zona 18N) y Geographic Coordinate System (GCS), ambos con datum WGS84 que son resultado de la interpolación de datos de nivel freático. Los datos se encuentran en formato ESRI-GRID y en formato ASCII-GRID (AAIGrid)

En los folders "./nivel-freatico/pluviometria/interpolaciones" y "./nivel-freatico/ pluviometria/interpolaciones-corte-fincas" se encuentran los rasters en dos sistemas de coordenadas: Universal Transverse Mercator (UTM) (zona 18N) y Geographic Coordinate System (GCS), ambos con datum WGS84 que son resultado de la interpolación de datos de precipitación.

b. Mapeo de zonas problema

Usando las interpolaciones realizadas, se mapearon las zonas problema usando diferentes rangos respecto a la tabla de agua. Aquellas áreas en las que se encontró un valor de tabla de agua muy bajo (i.e. tabla de agua muy cercana a la superficie del suelo) se señalan como críticas (NF entre 0 y 50 cm); aquellas áreas con NF entre 50 y 100 cm se señalan como de alto riesgo, aquellas con NF entre 100 y 150 como de riesgo intermedio, y aquellas con NF mayor a 150, como de riesgo bajo (Figura 6).

Para 2007 (Figura 6), por ejemplo, en general, la zona norte y este de la región (exceptuando el mes de Abril), el nivel freático se encuentra en la zona de riesgo bajo; mientras que en la zona oeste, en general el riesgo de anegamiento tiende a ser mayor, probablemente debido a las propiedades físicas de los suelos. La zona de mayores problemas se encuentra en las fincas Bahía, Coralina, Fragata, Ensenada, El Manantial, San Quintín y Gorgona. Durante los meses de Agosto y Noviembre, se encontró que podría haber problemas en áreas muy al norte de la región.

Este tipo de visualización histórica provee una herramienta útil y de fácil acceso y aplicación para la detección de problemas de drenaje. En general se encontró Que el nivel freático presenta muy alta variabilidad, en particular en ciertas épocas en las que se hace crítica una intervención, para evitar una posible pudrición de raíces y el consecuente detrimento de la producción.

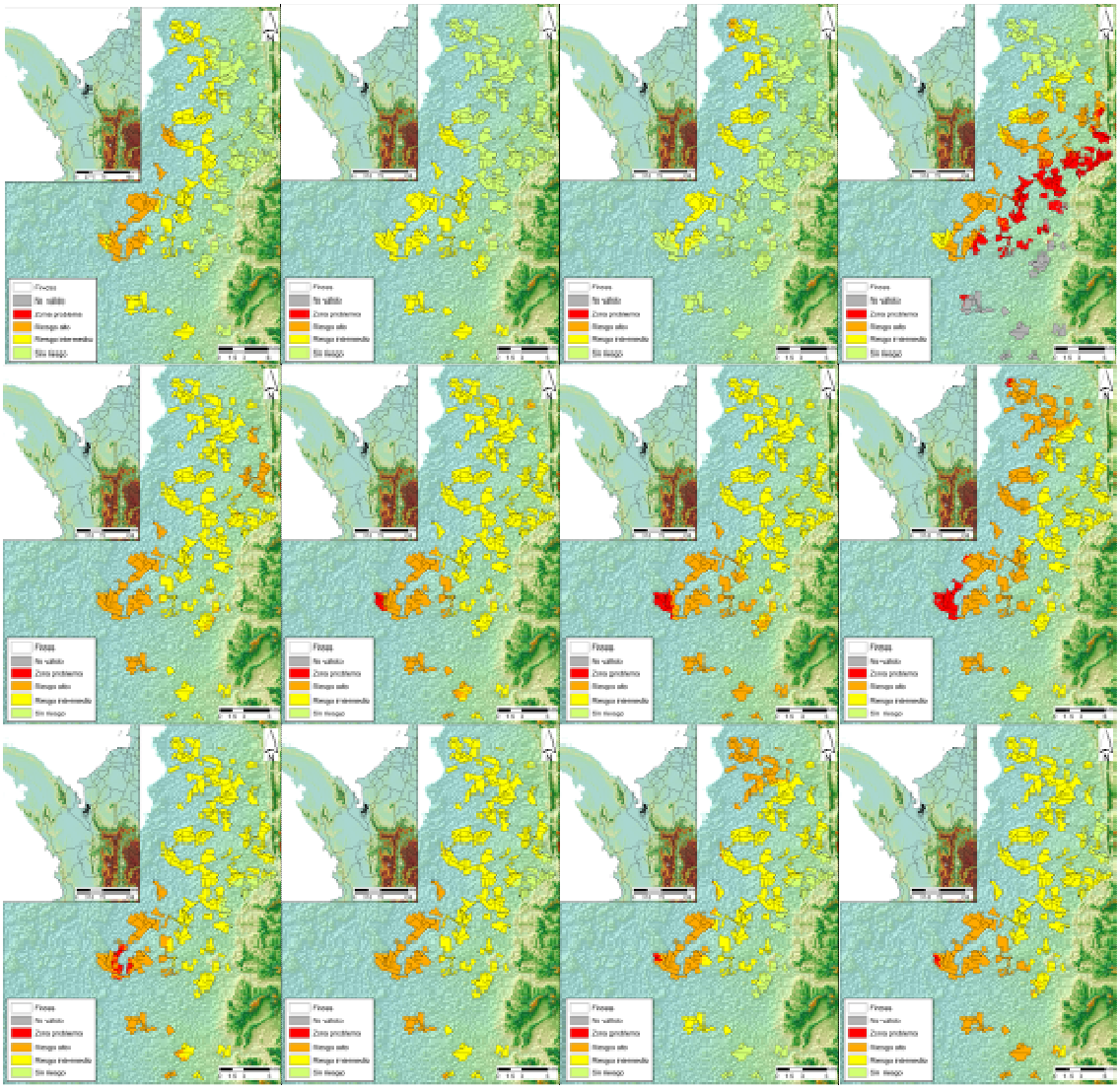

Figura 6 Variación mensual en el nivel freático en la zona bananera de Urabá durante el año 2007. De izquierda a derecha y de arriba hacia abajo cada figura representa un mes de enero hasta diciembre.

Se observaron algunas zonas grises (zonas con datos no válidos o fuera del rango), y esto se debió a que los datos usados en el ajuste de la función TPS para ese mes en particular no cubrieron el rango total de variabilidad actitudinal dentro de la región. Esto produce extrapolación, que al final se deriva en errores en el resultado. Se recomienda el incremento en el número de datos para aquellas áreas con baja disponibilidad en el número de pozos. Esto, a la postre, reducirá los errores de interpolación y aumentará la confiabilidad de los resultados.

Como parte del análisis realizado, se provee una superficie de incertidumbre basada en la densidad de los puntos de muestreo (el óptimo o máximo número de puntos disponibles).

Las incertidumbres (Figura 7). Las áreas más oscuras son aquellas en las que las incertidumbres tienden a ser mucho mayores, lo que aumenta la probabilidad de error.

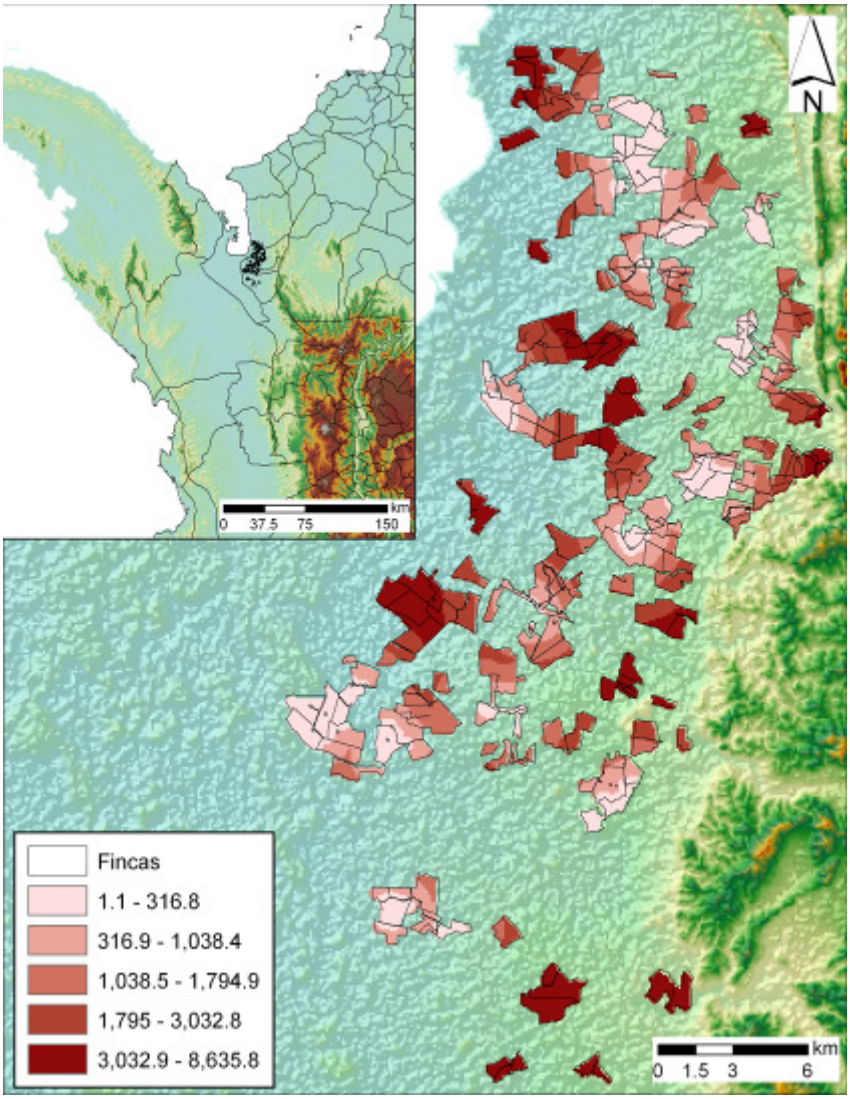

Figura 7 Incertidumbres de interpolación como distancias de cada píxel al punto más cercano con datos disponibles, clasificados en quintiles.

La excelente disponibilidad en la zona este (cerca a las fincas Fragata y Bahía) hace que los resultados de esta zona sean bastante buenos. Idealmente, la misma disponibilidad debería promoverse hacia toda la región.

c. Análisis histórico usando dos casos de estudio

Usando el script "extractByCoordinate.R" que se encuentra en el folder "./nivel-freático/plantilla", se pueden extraer los valores correspondientes a cualquier coordenada

dentro del área de estudio, a nivel mensual, para los 3 años (2007 hasta 2009). Estos valores se ingresaron en una plantilla de análisis históricos. La plantilla "00-plantilla-analisis-nf-ppt-limpia.xls", se encuentra en el mismo folder que el script y sólo requiere que los datos se copien y se peguen desde el archivo de datos extraído para visualizar el comportamiento histórico en ese sitio en particular. Como ejemplo, se ha tomado el lote 3 de la finca Castilletes (Figura 8)

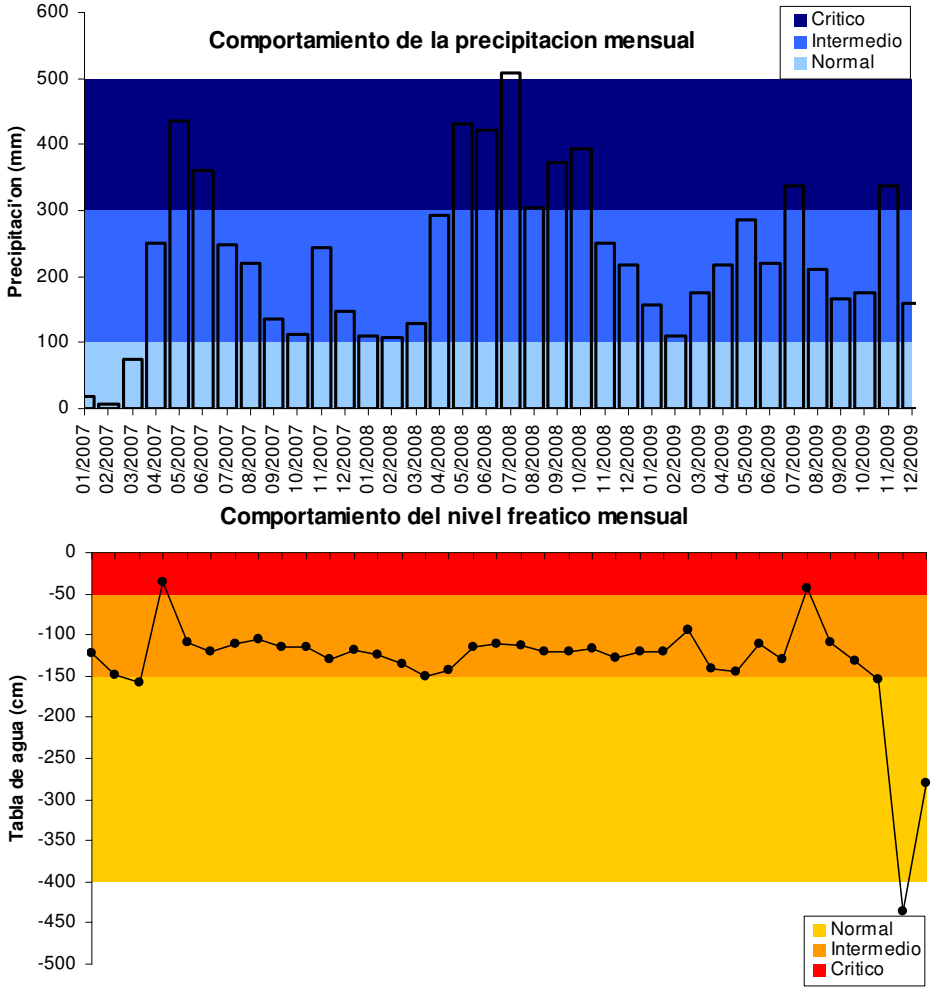

Figura 8 Comportamiento histórico del nivel freático y la precipitación de la finca Castilletes, lote 3.

Como se puede observar, hay una alta variabilidad temporal en la precipitación mensual. Las épocas más húmedas se encuentran en los meses de Mayo y finalizan en Agosto. Las épocas de finales y principios de año son en general de bajas precipitaciones. En general toda la zona tiene un comportamiento similar, siendo las épocas más húmedas aquellas de la mitad del año. El nivel freático tiene un comportamiento, entretanto, al menos en la finca Castilletes (lote 3), relativamente constante, manteniéndose en el rango seguro (franja naranja) y con tendencia hacia un NF muy profundo (zona amarilla). Solamente

durante el mes de Abril de 2007 y el mes de Agosto de 2009 se encontró que el nivel freático subió hasta la zona crítica (franja roja).

De la misma manera pueden extraerse los datos correspondientes a otros lotes de otras fincas. Sólo se requiere extraer los valores de los rasters con el script proveído (aunque también puede hacerse con algún otro paquete SIG, o incluso con otro lenguaje de programación, y pegar los datos en la plantilla mencionada anteriormente. Para extraer datos se debe ejecutar el comando *source*:

```
source("extractByCoordinate.R")
resultado <- extractTimeSeries(longitud, latitud, datafile="datos-salida.csv")
```

El script escribirá un archivo de datos (.csv) y un archivo de imagen (.jpeg) mostrando la localización del punto en el área de estudio (área de las fincas bananeras). La variable resultado contiene los datos extraídos en forma matricial.

5. Conclusiones y recomendaciones

En resumen, se ha realizado un análisis de variabilidad espacio-temporal del nivel freático, usando datos de campo colectados en pluviómetros y en pozos de observación. Se contó con datos de 23 pluviómetros y 726 pozos de observación georreferenciados. Con estos datos, se generaron superficies continuas mensuales usando el método de interpolación TPS (*Thin Plate Spline*). A partir de las superficies interpoladas, se identificaron las zonas problema usando una clasificación subjetiva sobre el nivel freático. Con esto, se sugieren las zonas críticas en las que debería intervenirse para mejorar el drenaje agrícola y por tanto la producción. En adición se plantea un análisis histórico de zonas problema basado en gráficos de variabilidad, y un script automatizado para extracción de los datos. Los resultados presentados en este estudio pueden sobreponerse con el diseño actual de los sistemas de drenaje de tal manera que se puedan detectar problemas de diseño y se puedan re-dimensionar y re-diseñar aquellas áreas bajo alto riesgo de anegamiento. Todas estas, al final, mejorarán los procesos de toma de decisiones en campo.

Como recomendaciones generales, se insta a una expansión en la colecta de información de campo, incluyendo el entrenamiento de personal de fincas para la correcta colecta de la información, incluyendo la georreferenciación de puntos de toma de datos, que resultan siendo fundamentales cuando se tratan de realizar análisis geográficos. Todo esto, incrementará la disponibilidad de la información y a la postre, mejorará los procesos de toma de decisiones en campo.

Adicionalmente, como próximos pasos, se recomienda el establecimiento de un sistema de toma de decisiones que involucre actores en diferentes niveles: productores, analistas SIG y estadística, expertos, y técnicos de campo. En adición a una infraestructura que

optimice el almacenamiento, flujo y análisis de información colectada en campo, incluyendo el establecimiento de una base de datos tanto de campo como de experiencias, problemas y soluciones. De esta manera, UNIBAN CI y los productores podrán tener un mejor monitoreo de la producción y podrán orientar las soluciones de la mejor manera.

Referencias

Hutchinson MF (1984) A summary of some surface fitting and contouring programs for noisy data.*CSIRO Division of Mathematics and Statistics, Consulting Report ACT* 84/6. Canberra, Australia.

Hutchinson MF, de Hoog FR (1985) Smoothing noisy data with spline functions. *Numerische Mathematik* 47: 99-106.

Jarvis A, Reuter HI, Nelson A, Guevara E (2008) Hole-filled seamless SRTM data V4, International Centre for Tropical Agriculture (CIAT), available from http://srtm.csi.cgiar.org.