Revista Tecnológica ESPOL, Vol. 21, N. 1, 91-98, (Octubre, 2008)

Disminución de Tiempo y Costos en la Obtención de Información Biológica de Campo con Valoración Estadística

O. Ruiz* (1), M. Jiménez (1), S. Bauz (1)
(1)Centro de Investigaciones Biotecnológicas del Ecuador
Escuela Superior Politécnica del Litoral
Campus "Gustavo Galindo Velasco" La Prosperina Km. 30.5 vía Perimetral
*oruiz@espol.edu.ec

Resumen

En todo proceso de investigación desarrollado en el campo agronómico, uno de los principales contratiempos es el ingreso de los datos colectados a una base de datos, para su posterior análisis estadístico que ofrezca información oportuna. El presente trabajo propone que estas investigaciones utilicen herramientas tecnológicas para optimizar el proceso de captura, almacenamiento y análisis de los datos, para la obtención de información estadísticamente validada y en periodos cortos de tiempo. La propuesta es implementar dos aplicaciones informáticas con distintas plataformas de desarrollo, que permitan reducir el tiempo de este proceso, al restar errores de ingreso de los datos a la base y disminuir el tiempo de respuesta. La primera es una aplicación informática que utiliza plataforma WAP instalada en una agenda personal digital (PDA) para la captura de datos y la segunda es una aplicación que hace uso de tablas dinámicas, que realiza consultas de una base de datos en Microsoft SQL-Server, y que adicionalmente utiliza programación de Visual Basic en Microsoft Excel para obtener información del Data Warehouse y aplicar algoritmos de minería de datos extraídos de las librerías del software estadístico R en su versión 2.6.2. El uso de estas dos aplicaciones refleja una disminución en el tiempo de obtención de la información. Los datos utilizados fueron recopilados entre el 2004 y el 2006, las aplicaciones obtenidas fueron desarrolladas entre el 2006 y 2007.

Palabras claves: Cubo de datos, captura de datos de campo, comunicación inalámbrica, Minería de datos.

Abstract

Throughout the research process developed in the agronomic area, one of the major bottlenecks is the data entry collected for subsequent statistical analysis that provides timely information field. This work proposed that these investigations use technological tools to optimize the process of capture, storage and analysis of data, for obtaining information and statistically validated in short periods of time. This paper proposes the implementation of two software applications with different development platforms, to optimize the time of this process by reducing errors data entry and especially the response time. A software that uses Wireless Application Protocol platform and installed in a Personal Digital Assistant (PDA) for data capture and other software that uses of dynamic tables, that consults in Microsoft SQL-Server database, moreover uses Visual Basic programming in Microsoft Excel to obtained information of the Data Ware house and to implement algorithms for Data Mining extracted from libraries of the statistical software R 2.6.2. The use of these two applications reflects a decrease in time for obtaining the information. The data used were compiled between 2004 and 2006, The applications were developed between 2006 and 2007.

Key words: Data Cube, Data field capture, Wireless Application Protocol, Data Mining.

1. Introducción

La disminución de disponibilidad de tiempo para obtener información con sustento estadístico, crea la necesidad de renovar las metodologías actuales de captura y análisis de datos de estudios realizados en el campo agronómico, utilizando aplicaciones informáticas desarrolladas para plataformas de comunicación inalámbrica, las cuales son comercialmente utilizadas en otros países tecnológicamente más desarrollados.

Toda investigación genera datos que deben ser recolectados y almacenados de tal manera que permita su fácil acceso y uso, además de reducir al máximo el tiempo de respuesta y los posibles errores de almacenamiento. Cuando existen varias áreas de investigación involucradas, la cantidad de datos por recolectar se incrementa de manera sustancial, por tal motivo es necesario utilizar nuevas herramientas para optimizar su almacenamiento.

Cuando los datos son capturados de manera tradicional, el tiempo que se requiere para su

Recibido: Mayo, 2008
Aceptado: Agosto, 2008

preparación previa al análisis estadístico es muy largo. De igual manera cuando estos datos son analizados de forma tradicional, la información que de ellos se puede conseguir es limitada. Generalmente existe correspondencia entre la complejidad de los análisis y la relevancia de los resultados. Ausencia de análisis estadísticos matemáticos complejos genera escasa valoración de los datos obtenidos, desperdicio de información importante que en muchas ocasiones se encuentra oculta y que por falta de una herramienta informática más eficiente y amigable, no se pueda explorar a fondo para encontrar patrones conductuales de los entes investigados. Es claro que esta situación limita al investigador al momento de tomar decisiones al no contar con información ágil, fidedigna, demostrable y estadísticamente sustentada.

Para poder realizar lo anteriormente expuesto, existen agendas digitales personales PDA's que haciendo uso de tecnologías de comunicación inalámbrica WAP ofrecen la posibilidad del inmediato envío/ recepción de datos, además del Data Warehouse (DW) para consultas y análisis de datos presentes, y de la Minería de Datos (MD) para datos históricos, las cuales crean un enlace entre la estadística y la informática. MD es el análisis que resume los datos en formas nuevas, útiles y comprensibles, y que además encuentra relaciones inesperadas; pues, para las esperadas existe el DW/ procesamiento analítico en línea («On Line Analytical Processing», OLAP), aunque al final ambas dan un conjunto de respuestas.

La MD, hasta ahora ha sido muy utilizada en áreas tales como: educación, telecomunicaciones, comercio y marketing, sector financiero, seguros, Internet, etc.; donde se analiza el patrón de conducta de los usuarios o consumidores. El presente trabajo, orienta la aplicación de la MD hacia el campo agronómico, para analizar el patrón de conducta de organismos procariotes y eucariotes, bióticos y abióticos.

El objetivo principal de esta investigación es proveer "información" al propietario de los datos, reduciendo al mínimo el tiempo para su obtención; a través de la creación de un DW; una aplicación WAP para la captura y envío de los datos; y aplicar técnicas de MD a los datos históricos.

Para el desarrollo de la propuesta se tomó como referencia el proceso de captura de datos en campo utilizado en el Centro de Investigaciones Biotecnológicas del Ecuador (CIBE). Además se aplicaron los modelos por ellos obtenidos, de estudios realizados en la costa ecuatoriana, sobre variables agronómicas y fitosanitarias de diferentes variedades de banano y en varias zonas geográficas, obtenidas desde el 2004 hasta el 2006. Manteniendo el esquema, la presente propuesta podría ampliarse hacia otros modelos, ámbitos, áreas, etc.

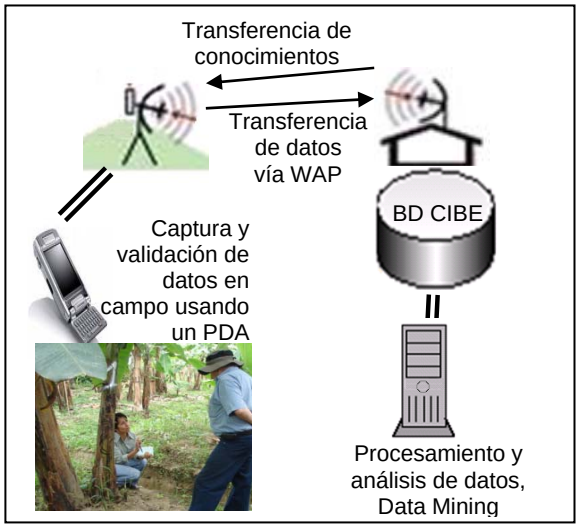

Figura 1. Esquema conceptual del proceso: captura de datos en campo, envío a la BD-CIBE, análisis de datos y devolución de información a campo.

El esquema conceptual de la captura y análisis de datos propuesta se observa en la Figura 1; este esquema es un estándar que busca automatizar el proceso, eliminando tiempo de traslado, digitalización de datos y análisis estadístico básico; haciendo una validación de los datos al momento de su ingreso, disminuyendo errores involuntarios. La disponibilidad de los datos dependería únicamente de la destreza del manejo del equipo por parte del usuario; con ello, los datos estarían listos para su respectivo análisis, ahorrándose tiempo del pre-procesamiento el cual se desarrolla de manera automática y compartida entre el PDA y la BD. Este esquema también ofrece al usuario, una pronta retroalimentación; realizando un análisis estadístico descriptivo desde el sitio donde se aloja el sistema, a través de una consulta en el DW, utilizando una tabla dinámica, la cual genera automáticamente valores tabulados y gráficas con medidas resumen de los datos ingresados. Los resultados estadísticos podrán ser transferidos al campo en minutos.

2. Desarrollo

El proceso actual de captura de datos agronómicos en las hacienda, se lo realiza utilizando bitácoras, en las cuales se registran los valores observados de las variables evaluadas. Una vez finalizada la captura, las bitácoras son trasladadas al centro para que los datos sean ingresados en hojas electrónicas, que sirven como repositorio, estos datos deben ser pre-procesados antes de ser copiados al software estadístico para su posterior análisis; estos resultados pueden ser transferidos al campo en una siguiente visita.

2.1 Creación de la Base de datos (BD)

La primera etapa de este proyecto es la construcción de la BD, para evitar duplicidad y aumentar la confiabilidad de los datos. Esto se hace siguiendo el esquema lógico planteado en la Figura 2.

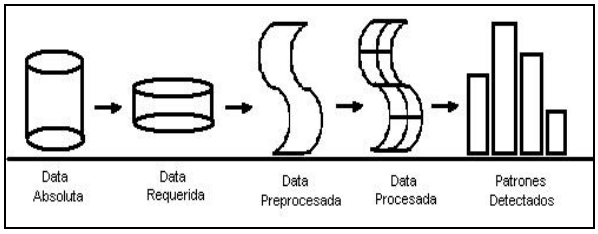

Figura 2. Esquema lógico del proceso de obtención de información, desde el ingreso de datos desde campo, hasta la obtención de información con MD.

Para la creación del almacén de datos, se hizo uso de MS SQL-SERVER por poseer las características necesarias para la propuesta; y de las herramientas de Microsoft Visual Basic y Servicios OLAP. Seguidamente, se elaboró el diseño lógico de la BD desde la perspectiva de la investigación realizada, éste genera una descripción escrita (Script) de BD, que es compatible con varios manejadores de BD (RDBMS) entre ellos MS SQL-Server.

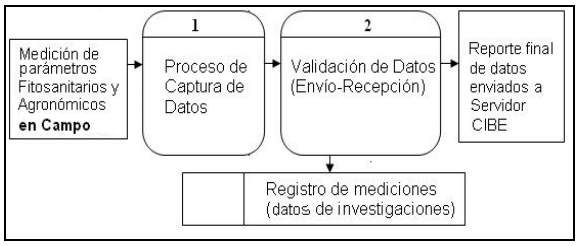

Figura 3. Diagrama de flujo de datos – Nivel 1: Proceso de envío de datos desde campo mediante dispositivos móviles, validación de los datos obtenidos, envío a BD-CIBE, registro en BD y reporte de la recepción de datos recibidos.

La representación gráfica del flujo de información utilizando el diagrama de flujo de datos - DFD se muestra en la Figura 3, el cual sirve como base para la creación de los formularios que se utilizaron en la captura de los datos.

Seguidamente se elaboró el diseño físico de la BD con las respectivas entidades y relaciones presentadas en la Figura 4.

Se determinaron las categorías de los datos o *dimensiones*, las cuales fueron: tiempo, ubicación geográfica, genotipo, etapas de evaluación, nivel del estudio o ensayo, entre otras; algunas no tuvieron más de un nivel en la jerarquía. Por el enfoque inicialmente planteado, el cual era analizar organismos procariotes y eucariotes, además de los factores bióticos y abióticos, y en función de los datos con los que se contó para el estudio, los parámetros agronómicos y fitosanitarios de las plantas de banano, se establecieron como *hechos* o medidas del Data Warehouse, esto son: altura, diámetro, número de hojas, emisión foliar, índice de infección, etc., a las cuales también le llamamos variables, con estricto sentido estadístico.

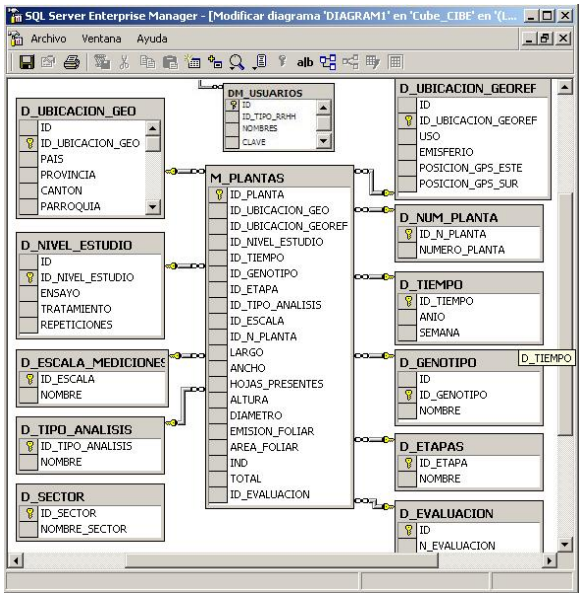

Figura 4. Diagrama Entidad–Relación del Data Warehouse obtenido, indicando cada una de las tablas creadas y sus relaciones, según lo indicado por los expertos agrónomos. En la tabla central M_PLANTAS se observan los *hechos* o medidas y en las tablas que se encuentran al contorno se observan las *dimensiones* o categorización de datos.

2.2 Desarrollo de Formularios para la aplicación WAP

Se utilizaron diferentes tecnologías de programación WAP (WML, WML Script, XHTML) para crear los formularios requeridos. Analizando la información necesaria para identificar a que investigación pertenecen los datos que van a ser colectados, se crearon cuatro ventanas de diálogo:

1. Formulario de inicio de sesión (reconoce al usuario);
2. Formulario de selección de hacienda o sitio del ensayo;
3. Formulario de selección del ensayo (información del diseño de experimento); y
4. Formulario de ingreso de datos de las unidades de observación.

Pruebas del sistema

Con respecto a la publicación de la aplicación WAP, ésta fue desarrollada localmente en una estación de trabajo con el simulador "Palm OS Cobalt Simulator". Para su posterior puesta en marcha debieron cargarse o publicarse en el Web SERVER, utilizando protocolos de transferencias, los archivos se enviaron hasta el servidor ESPOL, el cual aloja la BD creada. Este servidor tiene las mismas características que un servidor web normal; es decir, puede responder a solicitudes de clientes remotos por medio de archivo HTML, además del soporte para procesar páginas WAP que puedan ser leídas desde un PDA. En cuanto al hardware, las características básicas para un óptimo rendimiento son: Servidor HP ML500 Proliant, DELL PowerEdge o similar, capacidad de almacenamiento de al menos 100 GB, procesador Intel Pentium Dual core 2 duo, 2GB de memoria RAM, y sistema operativo WIN2003 server.

Para la configuración y conexión de la BD CIBE y BD WAP; la información que se envía desde el campo por medio de los PDA es almacenada en el servidor de la ESPOL. Para que esta información se vea reflejada en el servidor del CIBE es necesario ejecutar un proceso de migración entre ambos servidores; esto lo realiza un programa rastreador que detecta modificaciones en el servidor de la ESPOL y manda a ejecutar un programa de transferencia de datos para almacenarlos desde la BD ESPOL a la BD CIBE.

Se configuró el servicio WAP del PDA Sony-Ericson P910; se debieron habilitar los módulos necesarios (usuario y clave WAP asignado por proveedor de servicio de comunicación WAP, dirección IP del servidor del proveedor del servicio de telefonía móvil) para que estos puedan acceder a la aplicación WAP desarrollada.

Debido a que los datos viajan a manera de trama (cadena de datos), se debieron realizar pruebas de envío/ recepción de datos desde el campo hasta el servidor Web, para verificar su correcto almacenamiento en la BD CIBE.

2.3 Desarrollo del Data Warehouse y la aplicación de Minería de Datos DM

Con la BD multidimensional, se procedió a elaborar el diseño del Cubo de Datos (DC), mediante un esquema en estrella. Se utilizaron como dimensiones las variables de procesamiento mostradas en la Figura 5. Para dar facilidades de manipulación supervisada de los datos, hacia los usuarios de la aplicación, se creó un enlace entre el "Analysis Services" y la herramienta de escritorio, Microsoft Excel, de tal manera que el DC pueda ser administrado desde una tabla dinámica.

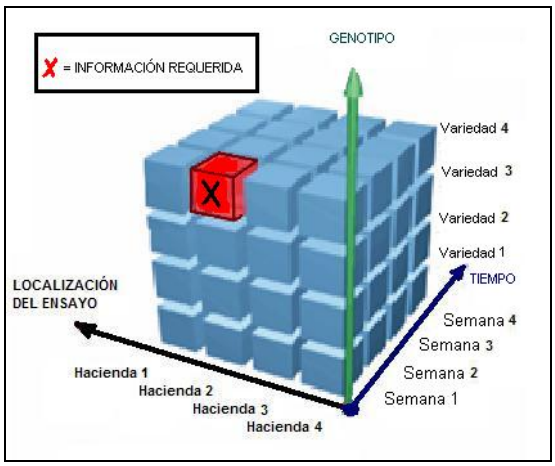

Figura 5. Consulta al Cubo de Datos, considerando como dimensiones las variables cualitativas: tiempo, localización del ensayo y variedades estudiadas; ofreciendo información de las variables cuantitativas (*hechos*) obtenidas de las mediciones realizadas en campo.

Recopilación y transferencia de datos históricos

Se recopilaron datos históricos de parámetros agronómicos obtenidos en campo de las investigaciones realizadas por el CIBE desde el 2004 hasta el 2006, los que se encontraban almacenados en hojas electrónicas de Excel, y debieron pasar por un largo proceso de depuración y validación hasta dejarlos listos y ser depositadas en la nueva BD CIBE, para lo cual se diseñó y desarrolló un Sistema para Transferencia de Datos (DTS), el cual extrae datos de las hojas electrónicas de Excel y la envía a la BD en MS SQL-Server.

Algo muy importante antes de ejecutar el DTS, fue verificar que todos los archivos de Excel por migrar tengan la misma estructura y secuencia para que la aplicación reconozca automáticamente en que registro y tabla deben ingresarse los datos de la fila y columna de la hoja electrónica.

Pre-procesamiento o Preparación de los datos.

La primera etapa de la preparación de los datos, fueron las validaciones que se realizaron con los expertos agrónomos, dieron pautas acerca de los comportamientos de las variables, valores máximos y mínimos permisibles, si son crecientes en el tiempo o son series temporales, etc. Con estas restricciones, se procedió a la transferencia de los datos. En la segunda etapa, se examinó la presencia de sesgos debidos a valores extraños o especiales; ya que es recomendable no tratarlos dentro del grupo general sino de manera separada, pues pueden ser fuente de información de la presencia de un factor que inicialmente pudo no ser considerado, pero que se expresa al momento de realizar los análisis estadísticos; estos factores podrían ser

variables exógenas que contaminaron los resultados, por errores ajenos al diseño, como la presencia de fuertes corrientes de viento o un temporal que afectó a las plantaciones, o podrían ser variables endógenas de no fueron detectadas o consideradas al momento establecer el diseño, como la presencia de un río cercano o las condiciones naturales del suelo, o simplemente son valores que fueron ingresados erróneamente por el digitador al pasar los datos desde la bitácora hasta la hoja electrónica.

Además, se debió decidir qué hacer con los valores perdidos (*missing values*), cual sería la regla que se utilice para inferir dichos valores para completar los datos faltantes; para ello existen diferentes metodologías en dependencia del tipo de datos y su comportamiento, se utilizaron las siguientes:

- si la variable es monótona creciente o monótona decreciente en el tiempo, se aplicó interpolación cruzada o regresión lineal;
- si la variable oscila en el tiempo, se utilizaron modelos de series temporales;
- si sus valores giran alrededor de un valor central, se utilizó la media de los valores conocidos.

Por la naturaleza de las variables analizadas, no fueron necesarios métodos numéricos de interpolación no lineal (Lagrange, Newton, etc.).

Se presentó la necesidad de transformar los valores originales de algunas variables; las utilizadas en este trabajo fueron: la transformación logarítmica, la estandarización de los datos, la discretización por medio de escalas y la aplicación de una función matemática para obtener el "área bajo la curva" para analizar la evolución (longitudinal) en el tiempo que describen características especiales de las plantas, como su altura o diámetro, desde la siembra hasta la cosecha, las que requieren de un único análisis que refleje el progreso que tuvieron durante el tiempo de evaluación; este concepto también se aplicó para lagunas *métricas* como Volumen = Altura * Diámetro, ó área foliar = largo * ancho (de la hoja 4 de la planta).

Selección de metodologías estadísticas básicas y los algoritmos de MD.

Para la realización del análisis exploratorio de los datos, fueron indispensables las técnicas estadísticas clásicas de análisis descriptivo, tales como análisis de Frecuencias, Sesgos, medidas de tendencia central y dispersión; a estos se les agregó estadística inferencial a través de contrastes de hipótesis para medias, varianzas, proporciones y pruebas de bondad de ajuste. La siguiente etapa fue la estadística predictiva a través de la obtención de modelos matemáticos que describan el comportamiento estadístico de los datos, para ello se aplicó análisis de regresión o multivariado.

Seguidamente se seleccionaron los métodos o técnicas de clasificación para la MD, sobre la base de la aplicación de métodos estadísticos robustos y comprobados; debido a que la idea principal del presente estudio es proveer información ágil con soporte estadístico pero de fácil interpretación para los investigadores, se consideraron las metodologías de menor complejidad al momento de analizar los resultados que estas proveen, por tal motivo como técnica de agrupamiento se utilizó el "clustering", considerándose una buena alternativa el No supervisado – jerárquico, porque ofrece la posibilidad de mostrar a través de Dendogramas el agrupamiento de los casos o registros, no así el clustering No supervisado – No jerárquico, pues no ofrece la posibilidad de mostrar de forma visual el agrupamiento de los casos. La Correlación, también fue considerada como método para realizar agrupamiento por afinidad.

El análisis de regresión fue considerado para realizar predicciones cuando se detectó correlación entre uno o más pares de variables de interés. El análisis discriminante como método de clasificación, pues este asigna un nuevo caso o registro a uno de los diferentes grupos previamente definidos en base a la información histórica.

Otra técnica considerada es la "Regla de inducción", que no es mas que la extracción de reglas if-then de datos basados en significado estadístico, ésta trata de identificar elementos de las poblaciones estudiadas que pudiesen responder de manera similar ante eventos específicos.

Aunque el sistema inicialmente no fue desarrollado para que realice comprobación de los supuestos estadísticos tales como bondad de ajuste, homogeneidad de varianzas o independencia de las observaciones; estas pueden fácilmente ser agregadas utilizando la misma estructura y herramientas de análisis.

Los algoritmos genéticos, no fueron considerados en el presente estudio.

Para los análisis se utilizó el Software Estadístico R, por su amplio contenido de librerías útiles para alcanzar el objetivo inicialmente planteado y además es libre.

Las librerías de R utilizadas son: RODBC y MASS, estas permiten que la máquina cliente o servidor defina una conexión a un origen de datos (BD en Access, SQL Server, MySQL, etc.), luego que la conexión esta lista, se procede a crear un Data Frame sobre el cual se aplican los algoritmos, este no es más que es un tabla compuesta por filas de registros y columnas de variables, lo que está en la memoria que administra R. Las salidas de datos estadísticos son realizadas a través de gráficos en formato jpg o pdf y tablas utilizando la exportación en R con el comando savePlot. Por último se puede utilizar RODBC para grabar las salidas antes mencionadas en la BD del sistema, siendo esta opción muy utilizada porque la aplicación es dinámica.

Finalmente se procedió a la implementación y el acoplamiento de las diferentes herramientas.

2.4 Breve Análisis comparativo entre las dos metodologías de captura y análisis de datos

Considerando los procesos actual y propuesto para la captura de datos y obtención de información, se hace un análisis comparativo de tiempos involucrados en cada uno de ellos.

Análisis de tiempos de respuesta

Obteniendo los tiempos promedios de ejecución en cada etapa del proceso actual, de acuerdo a información recibida de los involucrados, y con las pruebas realizadas en el proceso propuesto, se observan diferencias estadísticas muy significativas (T.test, p-value=0.000) entre los tiempos de respuesta de una y otra metodología, indicadas en la Figura 6.

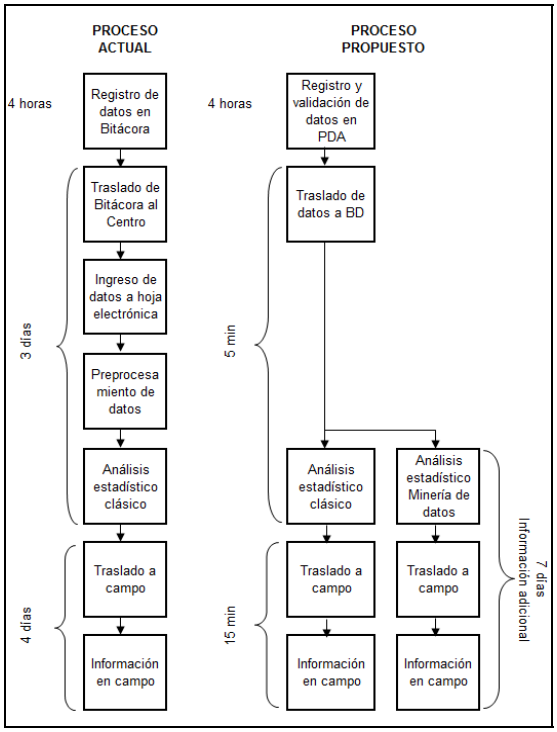

Figura 6. Comparación de los tiempos involucrados en los procesos actual y propuestos, reducción de 7 días a 20 minutos para tener información en campo de los parámetros analizados; si se desea un análisis más profundo de los datos tomará cerca de 7 días adicionales al tiempo del proceso propuesto.

Breve análisis del costo por el uso de esta tecnología

Egresos por el uso de la tecnología

- Costo del equipo = $ 463.56
- Costo anual de consumo aproximado = $66.00 * (12 meses) = $ 792.00 al año
  - Costo de grupo de datos enviados (incluido impuesto) ≤ $ 0.50;
  - Envío de datos por mes (aproximadamente 6 paquetes de datos diarios) = (22 días laborables al mes)*(6 mensajes al día)*($ 0.50 costo del mensaje) = $ 66 al mes

Adicionalmente se considera el costo del Diseño y Desarrollo de las aplicaciones, el cual está dividido por etapas:

- Levantamiento de información: $ 1000.00
- Análisis y requerimientos del sistema: $2000.00
- Diseño y desarrollo de base de datos: $3500.00
- Configuración de Servidores y Base de datos: $800.00
- Diseño y desarrollo de algoritmos en lenguaje R: $2300.00
- Diseño y desarrollo de las aplicaciones WAP y Web Interna: $3800.00
- Puesta en Marcha la aplicación: $600.00

Estos rubros fueron obtenidos a través de fondos competitivos.

3. RESULTADOS

Se obtuvo una aplicación informática que puede ser utilizada desde un PDA, la cual presenta ventanas de diálogo de fácil manejo y que en cuatro pasos está lista para enviar los datos de campo a la BD de CIBE y que en menos de 5 minutos tiene listos los datos para su análisis estadístico, haciendo uso de tecnología WAP, mostrada en la Figura 7.

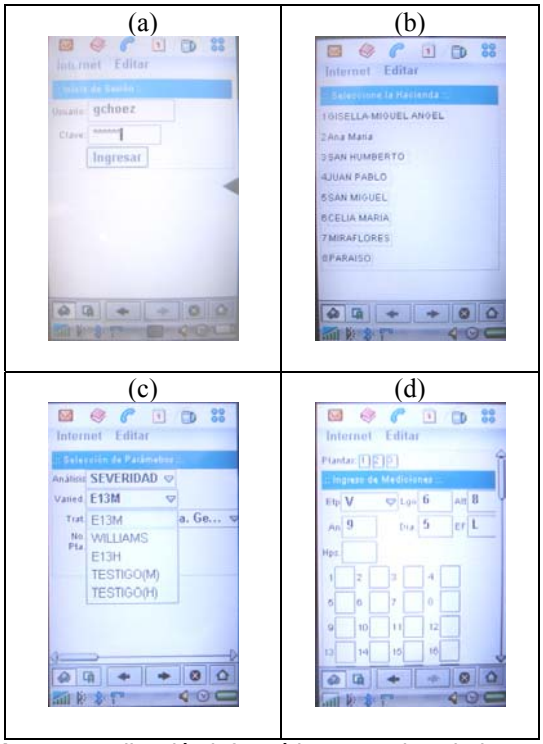

Figura 7. Aplicación informática WAP, instalada en el PDA; los cuatro formularios muestran los pasos a seguir para la captura de datos, a)Inicio de sesión, b) Selección de localidad, c)Selección del ensay, y d) Ingreso de dato.

Se obtuvo una aplicación informática amigable al usuario, bajo ambiente Excel, que utiliza tablas dinámicas de fácil manejo y que además da respuestas inmediatas, con el respectivo sustento estadístico, utilizando librerías del software estadístico R, ver Figura 8.

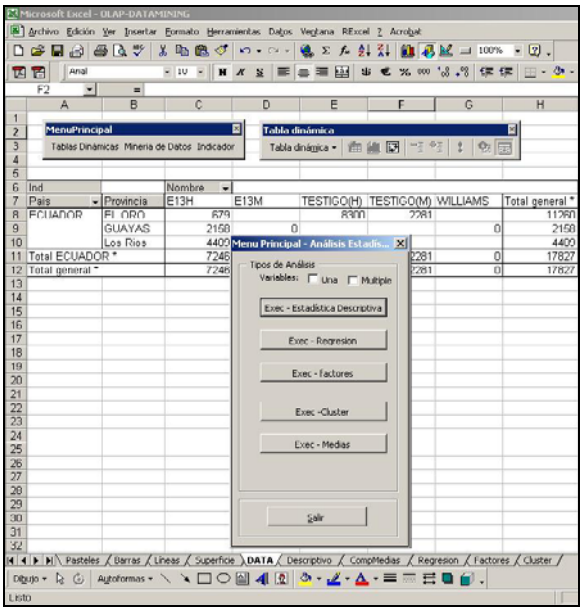

Figura 8. Aplicación informática DW y MD, la cual trabaja con tablas dinámicas en Excel, ofreciendo un entorno amigable al usuario.

La aplicación ofrece, gráficos explicativos y de fácil entendimiento con respecto a las variables seleccionadas, obtenidos inmediatamente ejecutada la consulta en el cubo de datos, lográndose de esa manera imágenes editables, ver Figura 9.

Haciendo uso de la Aplicación, se realizaron pruebas que validaran estadísticamente los resultados obtenidos; con la ayuda de las librerías del R se comprobó de manera numérica las diferencias apreciadas con su respectiva significancia estadística.

4. CONCLUSIONES

La combinación de las aplicaciones WAP y minería de datos, reduce de 7 días a 20 minutos la obtención de respuesta con sustento estadístico, del desarrollo de los parámetros agronómicos estudiados en campo.

Aproximadamente el 70% del tiempo de duración del proyecto, se trabajó en la depuración y preparación de los datos históricos que se encontraban en hojas electrónicas.

La aplicación WAP, ahorra alrededor del 90% del tiempo invertido actualmente en la captura y disponibilidad de los datos, y reduce al menos el 95% de errores en el ingreso de los datos.

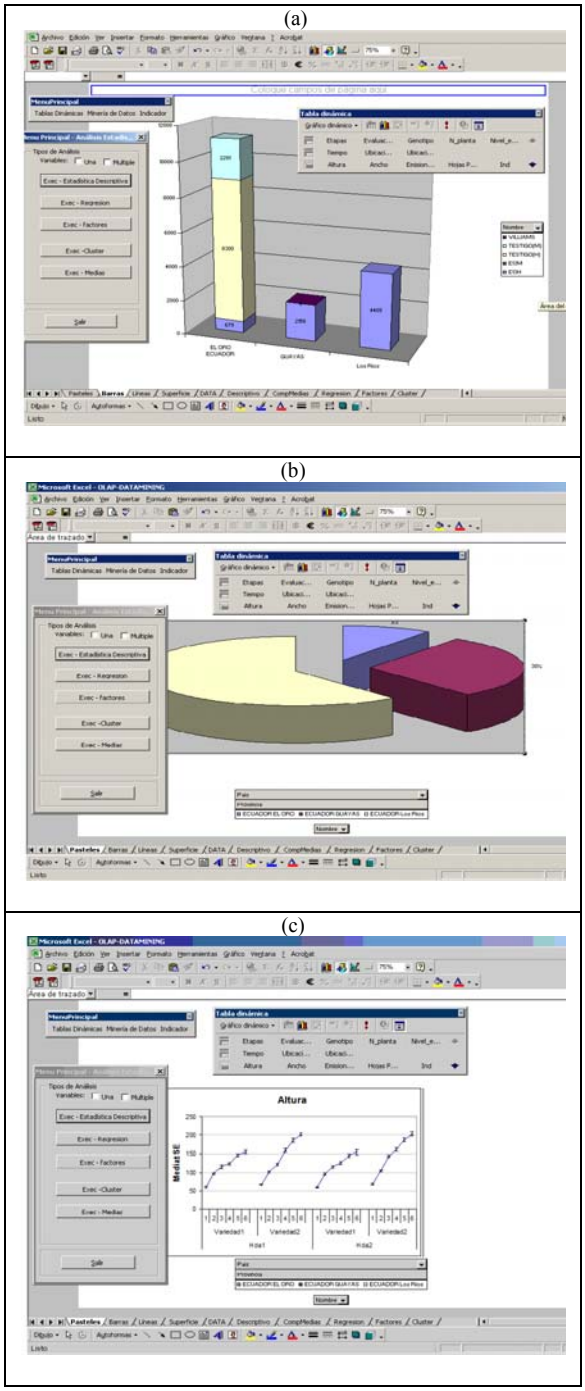

Figura 9. Resultados gráficos de una consulta utilizando el DW. a) gráfico de barras de los ensayos por localidad, b) gráfico de pastel de las variedades estudiadas en una localidad, c) Gráfico de cuatro variedades analizadas en el tiempo, en dos localidades

La base de datos y la aplicación WAP, necesitaron actualizarse a los 2 meses de creada, para agregar

nuevos *hechos* y *dimensiones*, que necesitan ser considerados en diferentes estudios.

Se aplicaron las tres técnicas de minería de datos más utilizadas, Árboles de decisión (56.6%), Agrupamiento (43.9%) y Estadística clásica (43.2%). Debido a la no presencia de variables de respuestas cualitativas, no fue conveniente trabajar con los árboles de decisión.

La aplicación de las reglas de inducción "if-then" ayudó a clasificar a las unidades de observación, encontrando relaciones que antes no habían sido percibidas.

El incorporar datos de clima, darían respuesta a preguntas muy importantes acerca de la relación entre parámetros bióticos y abióticos.

5. BIBLIOGRAFÍA

[1] Grossman R., Kamath C., Kegelmeyer P., Kumar V., Namburu R., Data Mining for Scientific and Engineering Applications, capítulos 2-3, Chicago, 2001.

[2] Hand D., Mannila H., Smyth P., Principles of Data Mining, capítulos 1-3, MIT Press, Cambridge, MA, 2001.

[3] Hamilton B., Blewett R., Programming SQL Server, capítulo 4,5,7,8, 2005.

[4] Adamson C., Venerable M., Data Warehouse Design Solutions, capítulo 1,5, New York ,1998.

[5] Kimball R., Wiley J., The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses, capítulo 1, 4, New York, 1996.

[6] Ibarra A., Lopez J., Cuestiones Éticas en Ciencia y Tecnología en el siglo 21, capítulos 2,4,7, Costa Rica, 2006.

[7] Hand D., Manila H., Principles of Data Mining, capítulo 1-5,10,Volumen 2, Cambridge, 2001.

[8] Minería de Datos; Fundación COTEC para la Innovación tecnológica, 1ra Edición, Nov. 2004, Depósito legal: M. 48.518-2007; ISBN: 84-95336-48-0.

[9] Carreño M., Sandoval J., Torres J., Construcción de una Bodega de Datos para el Proceso de Autorización de Gastos Médicos de la UABCS, 2005.

[10] Metodología bioinformática asociada, metodología, análisis de datos; Análisis de datos, Instituto de Salud Carlos III área Bioinformática.

[11] Guía de recursos bioinformática; Glosario; Instituto de Salud Carlos III área Bioinformática.

[12] Ramírez Bartutis, Rosa Mune Jiménez, Mayra Mansur Manuel et al. Estudio de la estabilidad genética del sistema de expresión de la proteína E del virus dengue 4 en la levadura metilo trófica Pichia pastoris. Rev Cubana Med Trop, sep.-dic. 2005, vol.57, no.3, p.0-0. ISSN 0375-0760.

[13] Minería de datos.(2006, Octubre).Disponible en: http://www.daedalus.es/AreasMDCasos-E.php

[14] Técnicas de Minería de datos.(2006, Noviembre). Disponible en: http://es.wikipedia.org/wiki/Miner%C3%ADa_de_datos#Ejemplos_de_uso_de_la_miner.C3.ADa_de_datos

[15] Algunas áreas donde la minería de datos ha sido exitosa.(2006, Octubre).Disponible en: http://www.answermath.com/data-mining/mineria-de-datos-3-aplicaciones.htm

[16] Bioinformática: Validación de Clusters de Genes basados en el análisis de rutas Metabólicas. (2006, Noviembre).Disponible en: http://www.upo.es/eps/bigs/geneval.html