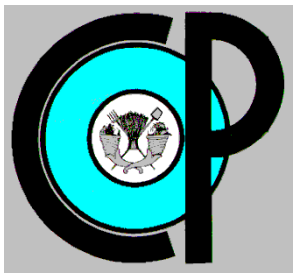


COLEGIO DE POSTGRADUADOS

INSTITUCIÓN DE ENSEÑANZA E INVESTIGACIÓN EN CIENCIAS AGRÍCOLAS

CAMPUS MONTECILLO

POSTGRADO DE SOCIOECONOMÍA, ESTADÍSTICA E INFORMÁTICA
COMPUTO APLICADO

Data Warehouse y minería de datos como alternativas al análisis de datos forestales.

JOSÉ ANTONIO FLORES CRUZ

TESIS

PRESENTADA COMO REQUISITO PARCIAL

PARA OBTENER EL GRADO DE

MAESTRO EN CIENCIAS

MONTECILLO, TEXCOCO, EDO. DE MEXICO

2014

La presente tesis titulada Data **Warehouse y minería de datos como alternativas al análisis de datos forestales.** Realizada por el alumno **José Antonio Flores Cruz,** bajo la dirección del consejo particular indicado, ha sido aprobada por el mismo y aceptada como requisito parcial para obtener el grado de

MAESTRO EN CIENCIAS

SOCIOECONOMÍA ESTADÍSTICA E INFORMÁTICA

COMPUTO APLICADO

CONSEJO PARTICULAR

CONSEJERO

Dr. José Luis García Cué

ASESORA

Dra. Yolanda Margarita Fernández Ordoñez

ASESORA

Dra. Reyna Carolina Medina Ramírez

ASESOR

Dr. Víctor Manuel Cetina Alcalá

Montecillo, Texcoco, Estado de México, Noviembre de 2014.

DATA WAREHOUSE Y MINERÍA DE DATOS COMO ATERNATIVAS AL ANÁLISIS DE DATOS FORESTALES

José Antonio Flores Cruz, MC.

Colegio de Postgraduados, 2014.

Resumen

El presente trabajo surge de la necesidad de analizar grandes volúmenes de datos originados durante el Inventario Nacional Forestal y de Suelos (INFyS) 2004-2009. El objetivo fue "Diseñar un Data Warehouse y la aplicación de modelos de minería de datos como alternativas para el análisis de información forestal". La metodología de la investigación fue cualitativa, de diseño y evaluación de software. Se tomó la base de datos del INFyS 2004-2009 y se reconstruyó su diagrama entidad-relación. Usando la información de la base de datos del INFyS 2004-2009 como insumo, se compararon cuatro modelos de minería de datos para la clasificación del género arbóreo *Quercus*, y se seleccionó el mejor modelo mediante los criterios de gráfica de elevación y de precisión. A partir de esta base de datos, también se diseñó un almacén de datos o "Data Warehouse" para la construcción de cubos de análisis para volumen de madera, biomasa y carbono. El software utilizado fue SQL Server 2008 que contiene el entorno para Desarrollo de Inteligencia de Negocios, el cual cuenta con los proyectos de Servicio de Análisis y Servicio de Integración, utilizados para desarrollar los paquetes que se obtuvieron como productos de la presente investigación. Además, se programaron interfaces para la visualización y análisis de la información, de las cuales dos fueron desarrolladas con las herramientas de Visual Studio 2010 para interactuar con la información del cubo de análisis y dos forman parte de las herramientas de SQL Server. La principal conclusión a la que se llegó es que a partir de la base de datos del Inventario Nacional Forestal y de Suelos 2004-2009 si es posible aplicar modelos de minería de datos para la clasificación del género arbóreo Quercus y también es posible diseñar un Data Warehouse para el análisis del volumen de madera, biomasa y carbono.

PALABRAS CLAVES: Almacén de datos, Base de Datos, Forestal, Minería de Datos, SQL Server 2008

DATA WAREHOUSE AND DATAMINING LIKE ALTERNATIVES TO FOREST DATA ANALYSIS

José Antonio Flores Cruz, MC.

Colegio de Postgraduados, 2014.

Abstract

This work emerge from the necessity to analyze large amounts of data from the National Forest and Soil Inventory (INFyS) 2004-2009. The goal is "Designing a Data Warehouse and the application of data mining models like alternatives to analyze forest information". The methodology for the research was qualitative and designing and evaluation of software. An entity-relationship diagram was building using the INFyS 2004-2009 database. Four data mining models was comparing for the tree genus *Quercus* classification, using INFyS 2004-2009 database. The best model was selected using elevation graph and precision criterial. A Data Warehouse was building to analyze wood volume, biomass and carbon using a multidimensional cube. The software used was SQL Server 2008 that contains the Business Intelligence Development Studio environment; this environment has projects like Analysis Services and Integration Services used to develop the packets obtained of this research. Different interfaces was used in visualization and analysis information, two of this was developed with Visual Studio 2010 tools for interact with the analysis cube information. The main conclusion obtained was that from the National Forest and Soil Inventory 2004-2009, is possible applied data mining models for the tree genus *Quercus* classification, and to design a Data Warehouse to analyze wood volume, biomass and carbon.

KEYWORDS: Data Warehouse, Database, Forest, Data Mining, SQL Server 2008

AGRADECIMIENTOS

Al Consejo Nacional de ciencia y Tecnología (CONACyT) por otorgarme el apoyo económico para continuar con mi formación académica.

Al Colegio de Postgraduados por su gran calidad educativa y brindarme la oportunidad de formar parte de su selecta lista de alumnos y ahora egresado.

Al Postgrado de Cómputo Aplicado, a sus excelente profesores y personal administrativos, que ayudaron a solventar todas mis dudas.

A mi profesor consejero Dr. José Luis García Cué por dedicar gran parte de su tiempo para guiarme durante todo el proceso de mi formación como Maestro en Ciencias. Por su paciencia, consejos y motivación para que siguiera siempre adelante.

A mi asesora Dra. Yolanda Fernández Ordoñez por su valiosa aportación y enseñanza en base de datos.

A mi asesora Dra. Reyna Carolina Medina Ramírez, que a pesar de estar en otra institución siempre me proporcionó su tiempo para atender todas mis dudas.

A mi asesor Dr. Víctor Manuel Cetina por valiosa aportación de conocimientos en el ámbito forestal.

Al MC. Ángel Leyva que me proporcionó mucha ayuda, asesoría y consejos, durante la presente investigación.

A Sofía de la Cruz Candelas por su brindarme su tiempo, paciencia y valiosa asesoría en todos los trámites académicos que se me presentaron.

A mi linda novia Karina Carmona Sosa por todo el tiempo que no pasé con ella y siempre comprendió el gran compromiso que tenía por delante.

A mi amigo Felipe Valentín López Figueroa por aceptarme como huésped en su casa durante todo el tiempo que duró mi Maestría.

A mis abuelos porque ellos son siempre mi inspiración en el trabajo.

A mis padres que siempre me brindaron su apoyo para seguir logrando mis objetivos.

A mis compañeros de generación con los cuales compartí gratos momentos en cada curso que tomamos juntos.

CONTENIDO

LISTA DE CUADROS

LISTA DE FIGURAS

SIGLAS

BD	Base de datos
BIDS	Business Intelligence Development Studio
CMNUCC	Convención Marco de la Naciones Unidas sobre el Cambio Climático
CP	Colegio de Postgraduados
CONAFOR	Comisión Nacional Forestal
COP	Conferencia de las Partes
DBA	Administrador de Base de datos
DW	Data Warehouse
ETL	Extracción, Transformación y Carga
ERF	Evaluación de los Recursos Forestales
FAO	Organización de las Naciones Unidas para la Agricultura y la Alimentación
IF	Inventario Forestal
IFM	Inventario Forestal Mundial
INE	Instituto Nacional de Ecología
INFyS	Inventario Nacional Forestal y de Suelos
INIFAP	Instituto Nacional de Investigaciones Forestales, Agrícolas y Pecuarias
MD	Minería de datos
MD	Minería de Datos
PSEI	Postgrado de Socioeconomía Estadística e Informática
OLAP	Procesamiento Analítico en Línea
OLTP	Procesamiento de Transacciones en Línea
RNA	Redes neuronales artificiales
SEMARNAT	Secretaría de Medio Ambiente y Recursos Naturales
SGBD	Sistema Gestor de Bases de datos
SSAS	SQL Server Analysis Services
SSIS	SQL Server Integration Services
SSMS	SQL Server Management Studio
SSRS	SQL Server Reporting Services
T-SQL	Transact-SQL

I MARCO INTRODUCTORIO

1.1 Introducción

En el Colegio de Postgraduados (CP), el Programa de Cómputo Aplicado tiene sus orígenes en 1991, cuando se creó el Programa de Maestría en Cómputo Estadístico, en el entonces Centro de Estadística y Cálculo, Colegio de Postgraduados de la Escuela Nacional de Agricultura (ahora la Universidad Autónoma Chapingo). Desde su creación, el cómputo sirvió como apoyo de la enseñanza de la estadística en México y como vinculación de ésta con el desarrollo rural, (Santizo Rincón, 2001). Actualmente el Programa de Cómputo Aplicado, del actual Postgrado de Socioeconomía Estadística e Informática (PSEI) del CP, desarrolla trabajos siguiendo seis líneas de investigación, las cuales son: Informática en agricultura, Bases de datos y Sistemas de información y Geomática, y adicionalmente en éste año, 2014, se han incluido la Bioinformática, la Minería de datos y la Comunicación por computadora (http://www.cm.colpos.mx/portal/-postgrados/computacion-aplicada). La presente investigación está contemplada dentro de dos líneas de investigación, Bases de datos y Sistemas de Información, como línea principal, y Minería de datos, como segunda línea.

El Programa de Cómputo Aplicado del CP trabaja en la vinculación con el sector rural del país, mediante el apoyo que brinda a instituciones gubernamentales y de investigación en los sectores agrícolas y forestales. Una de las formas en que el programa se hace presente en la sociedad es mediante la aplicación de nuevas tecnologías (desarrollo de aplicaciones y bases de datos) y la generación de nuevos métodos para la resolución de problemas particulares dentro de los sectores anteriormente mencionados. En este sentido, la presente investigación pretende introducir nuevas metodologías y herramientas para el análisis de datos forestales, especialmente para su aprovechamiento en el análisis de datos provenientes de los inventarios forestales. La metodología Data Warehouse y minería de datos se ha implementado con mucho éxito en sector de los negocios, la distribución y el marketing; en investigaciones se ha implementado en menor escala y todavía mucho menor es sector agronómico y forestal, donde su potencial ha pasado desapercibido por el uso de técnicas y softwares tradicionales, como SAS y R, para el análisis de datos.

El análisis de una base de datos, hasta hace unos cuantos años, se realizaba usando lenguajes de consultas, como SQL, sobre una base de datos operacional. En la actualidad se utiliza una nueva arquitectura, conocida como Data Warehouse (DW), que sumado al avance tecnológico en los medios de almacenamientos de información ha permitido que las instituciones conserven un registro detallado de sus operaciones (Nagabhushana, 2006).

Krzysztof *et al.* (2007) explican la principal diferencia entre una base de datos y un DW, comentando que las empresas cuentan con dos bases de datos, una para llevar a cabo sus operaciones cotidianas (depósitos, retiros, ventas, etc.) y otra para llevar a cabo un análisis de negocio mediante herramientas de procesamiento analítico en línea (OLAP), a la segunda base de datos se le llama Data Warehouse y consiste en información agregada a partir de los datos operaciones, como por ejemplo totales y promedios que se actualizan periódicamente a partir de la base de dato operacional.

Ruíz Torres (2007) define un DW como un repositorio de datos centralizado para apoyo de las actividades de análisis, que permite almacenar datos operacionales y eliminar inconsistencias entre los diferentes formatos existentes en los sistemas fuente, que además de integrar los datos de interés, permite incorporal información adicional validada por un experto.

El incremento en la cantidad de datos almacenados, también trae como consecuencia que se incremente de forma continua la diferencia entre la cantidad de datos para análisis y el conocimiento extraído de los mismos. La creciente necesidad de analizar grandes volúmenes de información ha dado lugar al análisis multidimensional y la minería de datos (Hernández Orallo *et al.*, 2004).

La integración de las técnicas de minería de datos en las actividades del día a día se está convirtiendo en algo habitual. La minería de datos está ligada con el manejo de grandes volúmenes de información y se ha utilizado en la astronomía, la medicina, la biología, la genética y la bioingeniería, entre otras (Larose, 2005 y Olson y Delen, 2008). Como ejemplos de aplicaciones se mencionan, la clasificación de cuerpos celestes, el análisis de secuencias de proteínas y el análisis de secuencias de genes (Fayyad *et al.,* 1996). A través de la minería de datos se mezclan diferentes disciplinas como la estadística, los sistemas de información, las

bases de datos y la inteligencia artificial entre otras para agilizar los procesos de análisis de información, siendo los negocios, el sistema bancario y la publicidad las áreas en las que más se ha empleado (Witten y Frank, 2005 y Krzysztof *et al.*, 2007).

Berry y Linoff (2004) consideran que muchos de los problemas que tienen que ver con la economía y los negocios pueden ser expresados en seis tipos de tareas: clasificación, estimación, predicción, agrupamiento, segmentación y descripción por perfiles. Hernández Orallo *et al.* (2004) y Olson y Delen (2008) agrupan las distintas tareas de la minería de datos en predictivas o descriptivas, que a su vez se dividen en tareas más específicas, por ejemplo, en las predictivas se considera a la clasificación y la regresión, mientras que en las descriptivas se considera el Clustering y las reglas de asociación; para ejecutar éstas tareas se recurre una serie de algoritmos matemáticos entre los que destacan los árboles de decisión, reglas de clasificación, regresión lineal, regresión logística, series de tiempo, k vecinos más próximos y redes neuronales.

En años recientes, la minería de datos ha tomado importancia en investigaciones agronómicas, por ejemplo Untaru M. *et al*. (2012) recopilaron información de diversas investigaciones que relaciona las técnicas de minería datos con los agronegocios; Savin *et al*. (2007) construyeron una red neuronal que predijo satisfactoriamente el rendimiento de los cultivos de trigo de invierno en algunas regiones del sur de Rusia. También Kumar (2011) hace una investigación donde se proponen modelos de redes neuronales para predecir rendimiento en arroz en el norte de la India. Asimismo, Stastny *et al*. (2011) predicen rendimiento de cebolla y tomate mediante la técnica anterior.

Una de las cualidades de los actuales softwares para minería de datos, es que se pueden implementar en plataformas existentes; además de contar con la capacidad de conectarse con nuevos productos y sistemas de recolección en línea (Reinosa *et al.*, 2012).

Elmasri y Navathe (2007) consideran que los DW, el procesamiento analítico en línea (OLAP) y la minería de datos ofrecen todas las funcionalidades necesarias para explotar las grandes bases de datos; principalmente, hace el señalamiento de que el DW es una de las tecnologías más importantes para obtención de información, al hacerlo sin sacrificar el rendimiento de las aplicaciones operacionales.

La tecnologías para la gestión de base de datos ha avanzado a la par de los requerimientos computacionales de tal modo que algunos sistemas gestores contemplan la administración de Data Warehouse y la implementación de algoritmos de minería de datos, como caso particular se menciona el Sistema Gestor de Bases de Datos SQL Server 2008, que desde su aparición en el mercado en 1993 y hasta la más reciente (SQL Server 2014) ha ido incluyendo una serie de herramientas que mejoraron sustanciosamente la forma de acceder y analizar la información, como por ejemplo, análisis multidimensional y minería de datos (Colledge, 2010).

En México, la Comisión Nacional Forestal (CONAFOR) es un Organismo Público Descentralizado dependiente de la Secretaría de Medio Ambiente y Recursos Naturales (SEMARNAT) que fue creada en 2001, con los objetivos de desarrollar, favorecer e impulsar las actividades productivas, de conservación y restauración en materia forestal, así como participar en la formulación de los planes, programas, y en la aplicación de la política de desarrollo forestal sustentable (http://www.conafor.gob.mx/web/nosotros/que-es-conafor/). La CONAFOR, entre sus obligaciones, tiene el de informar a la población mexicana sobre la cuantía, ubicación y condiciones de los recursos forestales del país, uno de los instrumentos utilizado para cumplir con estas obligaciones son los Inventarios Forestales, de los cuales el más reciente es el Inventario Nacional Forestal y de Suelos (INFyS) 2004-2009, fundamentado en la Ley General de Desarrollo Forestal Sustentable publicada en 2003.

La Gerencia de Inventario Forestal y Geomática, perteneciente a la CONAFOR, reporta que la base de datos utilizada, para almacenar la información de los inventarios forestales, está basada en un modelo conceptual Entidad-Relación. El sistema gestor de base de datos utilizado es Microsoft Access, bajo la justificación de que de esta manera se brinda una mayor flexibilidad para el manejo de los datos permitiéndoles mantener una independencia lógica y física de los datos, para evitar la redundancia de información, resguardar la integridad y calidad de los datos así como realizar consultas complejas optimizadas.

En la presente investigación se utiliza SQL Server 2008 como el Sistema Gestor de Bases de Datos (SGBD) encargado de la administración de un Data Warehouse, diseñado a partir de la base de datos del Inventario Nacional Forestal y de Suelos 2004-2009, para el análisis de

volumen de madera, biomasa y carbono. Este mismo software es utilizado para encontrar un modelo de minería de datos para la clasificación del género arbóreo *Quercus*. Como una aportación adicional, se desarrollaron interfaces que facilitan el análisis de la información generada en éste trabajo.

1.2 Justificación

La investigación que se propone pretende dar una solución informática y computacional al análisis de datos forestales mediante la implementación de un Data Warehouse y minería de datos, para el mejoramiento en el almacenamiento y el análisis de grandes volúmenes de datos provenientes de los inventarios forestales.

El principal motivo para realizar la presente investigación se relaciona con la experiencia del autor como analista de datos en instituciones de gobierno relacionadas con el medio ambiente y los recursos naturales, en donde pudo detectar algunas deficiencias al momento de analizar cantidades considerables de datos, más de 300 000 registros. Las principales deficiencias se relacionaban con la ausencia de datos, la presencia de datos atípicos, y el uso de diferentes formatos para un mismo tipo de dato; éstos problemas eran fácilmente solucionados cuando la cantidad de datos era pequeña y se volvía cada vez más difícil de detectar cuando la cantidad de los mismos se incrementaba, hasta el punto de ser una tarea que consumía demasiado tiempo y por consiguiente recursos humanos.

Otro motivo se deriva del diagnóstico que se hizo para Programa Estratégico Forestal para México 2025, el cual reveló que a raíz del uso de diferentes metodologías en cada uno de los inventarios anteriores al 2004-2009, los resultados obtenidos no son aptos para poder realizar una comparación espacio temporal de los recursos forestales del país. A partir del inventario Nacional Forestal 2000, en México se ha pretendido efectuar evaluaciones continuas y periódicas de los recursos forestales, con la finalidad de evaluar de manera consistente la dinámica de cambio de los bosques nacionales, pero que hasta el momento los inventarios no han aportado elementos para apoyar la planeación y orientación de acciones de producción y

restauración, y no han generado información sobre la dinámica de los ecosistemas forestales. El resultado de no contar con una metodología apropiada para llevar a cabo un seguimiento adecuado, por lo menos hasta antes del 2004, ha propiciado que en el país no se cuente con información confiable a cerca de las superficies forestales, ni con datos periódicos actualizados que constituyan una base sólida para la planeación sectorial. Esta falta de información tiene impacto negativos en la elaboración de reportes y evaluaciones nacionales, ya que las estimaciones de los principales indicadores se ha tenido que inferir de trabajos parciales en cuanto a cobertura, o bien, se han tomado de cartas de cobertura vegetal del país. Los resultados negativos obtenidos hasta el momento, se pretende contra restar adoptando una perspectiva para los Inventarios Nacionales Forestales acorde con la Ley General de Desarrollo Forestal Sustentable (CONAFOR, 2012).

La perspectiva que se tiene para el Inventario Nacional Forestal y de Suelos, es la de establecer una metodología que permita las comparaciones temporales entre inventarios, y que pueda ser aplicada tanto a nivel nacional como a nivel de entidad federativa para que estas generen su propio inventario estatal forestal, con la expectativa de que la integración de éstos den origen al Inventario Nacional Forestal (http://www.cnf.gob.mx:8080/snif/portal/infys).

Los trabajos de promoción de inventarios estatales, acorde con las perspectivas del inventario, comenzaron en el año 2008 y para el año 2012 ya se había realizado promoción en 29 de las 32 entidades federativas; además de haberse realizado algunos inventarios estatales (Aldana, 2012).

La revisión de los métodos empleados para el procesamiento y análisis de los datos de los inventarios forestales, tanto para investigación como para reportes, dio como resultado que éstos se concentran en la utilización de consultas SQL predefinidas, el manejo de hojas de cálculo y software para análisis estadísticos, como por ejemplo R y SAS. La manera en que se procesa y analiza la información de los inventario es muy similar a como lo hacían las grandes empresas hace algunos años. Las limitaciones, encontradas en la literatura, sobre el lenguaje de consulta SQL y de los paquetes estadísticos utilizados para el análisis de los datos forestales, hace necesario proponer el uso de nuevas herramientas de integración y análisis. La CONAFOR hizo público que el Inventario Nacional Forestal y de Suelos es un proceso que se mantiene en mejora

continua para garantizar la optimización de recursos y la utilización de técnicas y tecnologías de vanguardia que se transformen en mejores resultados, aprovechando las ideas y aportaciones de expertos y de los usuarios en general.

El Programa Forestal del Colegio de Postgraduados (CP) de Ciencias Agrícolas de México ha trabajado junto con la CONAFOR en diversas investigaciones donde se ha utilizado las bases de datos de los inventarios, detectándose que el volumen de datos es cada vez mayor, de un inventario a otro, y seguirá incrementándose especialmente ahora que se promueve la puesta en marcha de los inventarios estatales. Este incremento en la cantidad de datos obliga las instituciones, como el CP y la CONAFOR, a encontrar opciones computacionales para validar y analizar la información, en menor tiempo y sin requisitos adicionales de hardware, antes de que la cantidad de información sea intratable con los métodos de análisis actuales.

Uno de los trabajos de investigación realizados en el CP destaca algunos problemas de la actual base del Inventario Nacional Forestal y de Suelos 2004-2009. Méndez y De los Santos (2011) especifican que para calcular el volumen de madera en pie y la biomasa forestal, primero fue necesario auditar la base de datos del INFyS 2004-2009. El proceso de análisis de los datos se realizó mediante el software para análisis estadístico R Project, con el cual se programaron tantas líneas de código para los diferentes géneros arbóreos y regiones en las que se dividió al país para realizar las estimaciones de los indicadores mencionados. Sin duda alguna el proceso de limpieza y depuración es una de las tareas más complicadas y tardadas en todo análisis, y si a esto se le suma el hecho de tener que programar muchas líneas de código para obtener un resultado y luego modificar estas líneas para obtener estos resultados, se vuelve una tarea todavía más tardada y tediosa.

Otros trabajos, como el de Carmona Mota (2006) titulado “Minería de Datos usando SAS Enterprise Miner: Una aplicación en datos forestales” pretenden introducir nuevos métodos y herramientas para el análisis de datos forestales. En la investigación mencionada se ejemplificó el potencial uso de la minería de datos en la clasificación de la cobertura vegetal, mediante herramientas de cómputo para resolver problemas forestales particulares relacionados con el manejo de grandes volúmenes de información y se dio a conocer algunos algoritmos de minería

de datos usando la metodología SEMMA desarrollada por SAS Institute para la generación de conocimientos a partir de grandes bases de datos.

En el ámbito internacional se puede mencionar el trabajo de investigación "Disminución de Tiempo y costo en la Obtención de Información Biológica de Campo con Valoración Estadística" de Ruíz *et al.* (2008) llevado a cabo en el Centro de Investigaciones Biotecnológicas del Ecuador, en 2008, donde se implementaron dos aplicaciones informáticas desarrolladas en distintas plataformas, las cuales permitieron reducir el tiempo de ingreso de los datos colectados en campo, así como el almacenamiento y análisis para la obtención de información estadísticamente validada en periodos cortos de tiempo. Se utilizaron diferentes tecnologías de programación para crear los formularios pero el corazón del proyecto fue la construcción de un Data Warehouse administrado por SQL Server. Este es uno de los primeros proyectos agronómicos, que hace uso del potencial de los actuales gestores de bases de datos, con el fin de facilitar el levantamiento de datos de campo.

De todo lo anterior surge la siguiente pregunta, ¿Se puede diseñar un Data Warehouse y aplicar minería de datos como alternativas para el análisis de datos forestales? Para contestar la pregunta anterior, en la presente investigación, se utilizará la base de datos del Inventario Nacional Forestal y de Suelos 2004-2009, sobre la cual se realizarán dos tareas principales con ayuda del Software Microsoft SQL Server 2008 versión Enterprise. La primera tarea es aplicar algunos modelos de minería de datos con el objetivo de clasificar algún género arbóreo, seleccionar el mejor modelo y ejemplificar su potencial uso en el análisis de datos forestales. La segunda tarea consiste en diseñar un Data Warehouse, con la finalidad de hacer más eficiente el proceso de análisis de algunos indicadores forestales.

La CONAFOR y el CP cuentan con licencia para el uso de Microsoft SQL SERVER, por lo cual se convierte en el software ideal para la ejemplificación de las nuevas metodologías y técnicas para el análisis de la información de los inventarios forestales. La apertura de la CONAFOR, a explotar nuevas ideas, brinda la oportunidad para que en la presente investigación, se ejemplifique el uso de nuevas tecnologías (Microsoft SQL Server) combinadas con nuevos métodos de almacenamiento de datos (Data Warehouse) y técnicas para el análisis de

información (Análisis dimensional y minería de datos) para aplicarlo al análisis de datos de los inventarios forestales.

Los principales indicadores que se pretenden analizar usando la información del Data Warehouse son volumen de madera, biomasa aérea y carbono, los cuales son utilizados tanto para reportes nacionales como internacionales, a partir de que México se uniera al grupo de países firmantes del Protocolo de Kioto, en 2005. El Protocolo de Kioto es un acuerdo internacional con el objetivo de reducir las emisiones de seis gases de efectos invernadero que causan el calentamiento global surgido en la Convención Marco de la Naciones Unidad sobre el Cambio Climático (CMNUCC) celebrado en Kioto, Japón, el 11 de diciembre de 1997. El protocolo de Kioto fue el punto de partida para que muchos países volcarán la vista hacia el cuidado y protección del medio ambiente y emprendieran una serie de programas y evaluaciones para cumplir con los resultados propuestos. En México se emprendió una serie de acciones buscando contribuir a mitigar el cambio climático a través de políticas públicas.

La selección del Estado de México para ejemplificar el potencial de procesamiento de datos con que cuenta SQL Server, radica en la existencia de ecuaciones de volumen para los géneros maderables de esta entidad federativa, -propuesta por Méndez y De los Santos (2011)- que comparada con la cantidad de ecuaciones de volumen para otros estados, es relativamente baja lo que permite utilizarlas y no ver afectada la investigación en cuanto a su tiempo de conclusión.

El aporte social que brindará este trabajo se relaciona con las actividades de análisis que realizan los investigadores, en el sector forestal, al ejemplificar el uso de metodologías y herramientas para la integración y manipulación de grandes volúmenes de datos provenientes del Inventario Nacional Forestal y de Suelos 2004-2009 y que es aplicable también para los inventarios estatales; consecuentemente, la CONAFOR se puede beneficiar al implementar, la metodología descrita, en cada uno de sus procesos de generación de reportes y resultados.

1.3 Objetivos

1.3.1 Objetivo General

Diseñar un Data Warehouse y la aplicación de modelos de minería de datos como alternativas para el análisis de información forestal.

1.3.2 Objetivos Específicos

1. Identificar la estructura de la base de datos del Inventario Nacional Forestal y de Suelos 2004-2009.
2. Utilizar modelos de minería de datos, para la clasificación del género arbóreo *Quercus,* a partir de la base de datos del Inventario Nacional Forestal y de Suelos 2004-2009.
3. Elaborar un Data Warehouse a partir la base de datos del Inventario Nacional Forestal y de Suelos 2004-2009 para el análisis de información y la elaboración de cubos para reportes.
4. Analizar el volumen de madera, biomasa y carbono, para el Estado de México, mediante el procesamiento de cubos multidimensionales, a partir del Data de un Data Warehouse.
5. Desarrollar interfaces, para la visualización y el análisis de datos procesados mediante un cubo multidimensional generado a partir de un Data Warehouse.

1.4. Hipótesis

1.4.1 Hipótesis general

A partir de la base de datos del Inventario Nacional Forestal y de Suelos 2004-2009, se pueden aplicar modelos de minería de datos y diseñar un Data Warehouse para el análisis de información forestal específica.

1.4.2 Hipótesis particulares

1. La reconstrucción del diagrama de la base de datos del Inventario Nacional y de Suelos 2004-2009 permite la identificación de su diseño.
2. La información de la base de datos del Inventario Nacional y de Suelos 2004-2009 se utiliza para elegir modelos de minería de datos para la clasificación del género arbóreo *Quercus*.
3. La elaboración de un Data Warehouse a partir de la base de datos del Inventario Nacional Forestal y de Suelos 2004-200,9 permite el análisis de datos y la generación de reportes.
4. Mediante la tecnología de cubos dimensionales a partir de un Data Warehouse se analizan datos de volumen de madera, biomasa y carbono.
5. La plataforma de desarrollo Visual Studio 2010 proporciona las herramientas necesarias para el desarrollo de aplicaciones para la visualización y análisis de datos procesados mediante un cubo multidimensional generado a partir de un Data Warehouse.

1.5 Metodología de la investigación

1.5.1 Tipo de investigación

La presente investigación se ubica en la categoría cualitativa (Hernández Sampieri *et al.,* 2008) y de diseño de software.

1.5.2 Población y muestra

Población: La población comprende todos los registros de la base de datos del Inventario Forestal Nacional y de Suelos 2004-2009.

Muestra: La muestra comprende todos los registros que contengan información sobre los árboles de géneros maderables, vivos y muertos en pie, cuyo diámetro normal sea mayor o igual a 7.5 centímetros y menor o igual a 132.5 centímetros y una altura mayor o igual a 5 metros y menor o igual a 47.5 metros, cuya información haya sido levantada en los conglomerados y sitios de medición pertenecientes al Estado de México. Además, se utiliza una muestra adicional del 10% de la información dasométrica de los géneros maderables *Quercus, Pinus, Bursera, Lysiloma y Piscidia*, a nivel nacional, para el entrenamiento de modelos de minería de datos en la clasificación del género arbóreo *Quercus.*

Instrumentos de recogida de datos cualitativos: Se diseña una entrevista semiestructurada dirigida a los usuarios de los datos del Inventario Nacional Forestal del Programa Forestal del Colegio de Postgraduados Campus Montecillo. También, se elabora un cuadro de rúbricas para evaluar a las interfaces de análisis, desarrolladas para interactuar con los paquetes de SQL Server 2008, utilizados para manipular, procesar y analizar los datos del Inventario Nacional Forestal y de Suelos 2004-2009.

Recolección de datos cualitativos: Se tiene programado entrevistar a profesores investigadores del Programa Forestal en el Colegio de Postgraduados Campus Montecillo, que hacen uso de la información de la base de datos del Inventario Nacional Forestal en el mes de febrero de 2014, acudiendo a la oficina de cada uno de ellos. La rúbrica se emplea para evaluar el software al final del diseño aproximadamente en el mes de septiembre de 2014.

Análisis de datos cualitativos: Análisis de la información recabada en las entrevistas a través de análisis de discurso y de las rúbricas a través de tendencias en las respuestas.

Diseño de Software: Se crea paquetes independientes que operan con SQL Server 2008 y versiones posteriores, y el desarrollo de diferentes interfaces para visualizar y analizar la información procesada por los paquetes; la fuente principal de datos es la base de datos del Inventario Nacional Forestal y de Suelos 2004-2009, en formato de Microsoft Access y ecuaciones para el cálculo de volumen para el género arbóreo *Quercus* en el estado de México, almacenadas en un archivo de Microsoft Excel. El paquete para la extracción, transformación y carga de datos hacia el Data Warehouse, es desarrollado usando las herramientas del servicio de

integración de SQL Server 2008. El paquete para el procesamiento y análisis de la información mediante cubos multidimensionales es desarrollado usando las herramientas del servicio de análisis de SQL Server. Las interfaces para visualización y análisis son desarrolladas usando Visual Studio 2010.

1.5.3 Fases de la investigación

Fase 1. Recopilación de información teórica.

Fase 2. Análisis de la base de datos del Inventario Nacional Forestal y de Suelos 2004-2009.

Fase 3. Desarrollo del paquete de análisis de información a través de modelos de minería de datos.

Fase 4. Diseño del Data Warehouse.

Fase 5. Desarrollo del paquete para Extracción, transformación y carga de datos (ETL).

Fase 6. Desarrollo del paquete de análisis multidimensional.

Fase 7. Desarrollo de interfaces de consulta.

Fase 8. Redacción de resultados.

Fase 9. Elaboración del documento final.

Cada una de estas fases se explica con mayor detalle en el marco empírico.

II MARCO TEÓRICO

2. BASES DE DATOS Y MODELOS DE DATOS

En la presente sección se describe brevemente la historia relacionada con las bases de datos, se introduce los conceptos más importantes de una base de datos y se explica a detalles tres de los modelos de datos más utilizados actualmente a nivel mundial.

Silberschatz *et al.* (2002) comparten una breve reseña historia de las bases de datos, la cual se resume en el Cuadro 1.

Cuadro 1. Historia de las bases de datos

Periodo	**Acontecimientos**
Principios del siglo XX.	• Hollerith inventa las tarjetas perforadas, las cuales fueron usadas para el registro de la información del censo de los Estados Unidos de América.
Década de 1950 y principios de la década de 1960.	• Se desarrollaron las cintas magnéticas para el almacenamiento de datos. • Las tareas que involucraban procesamiento de datos fueron automatizados con los datos almacenado en cintas. • Las cintas y los paquetes de tarjetas perforadas se leían en forma secuencial, la cantidad de datos superaba la capacidad de la memoria principal por lo que los programas de procesamiento procesaban los datos siguiendo un orden determinado.
Finales de la década de 1960 y la década de 1970.	• A finales de la década de 1960 se intensificó el uso de discos fijos que permitieron un acceso directo a los datos. • Los discos permitieron el desarrollo de las bases de datos de red y jerárquica, dando lugar a las estructuras de datos como listas y árboles pudieran ser almacenadas. • En 1970 Codd definió el modelo relacional y formas no procedimentales de consultar los datos, lo que representó el origen de las bases de datos relacionales.

Década de 1980.	• El prototipo de System R condujo al primer producto de bases de datos relacionales IBM: SQL/DL capaz de competir en rendimiento con las bases de datos jerárquicas y de red. • Aparecen las primeras bases de datos relacionales como DB2 de IBM, Oracle, Ingres y Rdb de DEC. • Se realizan investigaciones en las bases de datos paralelas y distribuidas, así como el inicio de los trabajos sobre las bases de datos orientadas a objetos.
Principios de la década de 1990.	• Se comienza a utilizar el lenguaje SQL para las aplicaciones de ayuda a la toma de decisiones. • Se incrementó el número de herramientas para analizar grandes cantidades de datos. • Los vendedores de bases de datos ofrecían productos de bases de datos paralelas y relacionales orientadas a objetos.
Finales de la década de 1990.	• Se da un crecimiento del World Wide Web y con esto las bases de datos evolucionaron desarrollando interfaces web a los datos.

Fuente: Silberschatz *et al.* (2002)

La definición de base de datos (BD), tomada desde la perspectiva de diferentes autores como Ricardo (2004) y Elmasri y Navathe (2007) hacen referencia a un conjunto de datos que guardan una relación entre sí y que son almacenados de manera sistemáticamente, para ser usados en diferentes aplicaciones. El almacenamiento de los datos actualmente se realiza mediante archivos electrónicos, por lo que Date (2001) consideró a las BD como un armario electrónico, que requería de una serie de procedimientos para almacenar, recuperar, modificar, consultar y actualizar la información; además de protegerla contra usuarios no autorizados y caídas de sistemas. Las tareas anteriores dieron paso al desarrollo de Sistemas Gestores de Bases de datos (SGBD).

Ricardo (2004) define a un SGBD como "un paquete de software que configura estructuras de almacenamiento, carga datos, proporciona acceso a programas y usuarios interactivos, formatea datos recuperados, oculta ciertos datos, realiza actualizaciones, controla concurrencia, y efectúa respaldos y recuperación para la base de datos". Elmasri y Navathe (2007) consideran que un sistema gestor debe ser de propósito general utilizado para "facilitar los procesos de definición, construcción, manipulación, y compartición de bases de datos entre varios usuarios y aplicaciones". Reinosa *et al.* (2012) mencionan que la gestión de datos requiere del conocimiento y uso que se le dará a la información, para definir la estructura de

almacenamiento, y proveer los mecanismos para su manipulación, por lo que considera tres tipos de instrucciones principales para comunicarse con una base de datos: el lenguaje de definción de datos, el lenguaje de manipulación de datos y el lenguaje de control de datos.

- El lenguaje de definción de datos es el conjunto de órdenes que permite definir la estructura de una BD, es el utilizado para crear las tablas y las restricciones.
- El lenguaje de manipulación de datos son instrucciones contenidas por lo general en las aplicaciones y que se usan para altera el contenido del archivo de datos, es utilizado para ingresar un nuevo elemento a una tabla o para eliminarlo en caso de que exista.
- El lenguaje de control de datos son órdenes que se utilizan para implementar seguridad en la BD, como por ejemplo los privilegios que tiene cada usuario con respecto a los distintos objetos, los cuales pueden ser: insertar, modificar, eliminar, etc.

Las aplicaciones más usuales de los SGBD son para la gestión de empresas privadas e instituciones de gobierno y en menor medida en entornos científicos. Entre los SGBD más utilizados en la actualidad encontramos, por ejemplo, Oracle, Firebird, Microsoft SQL Server, MySQL, PostgreSQL, Interbase e IBM DB2 (Gómez Pino, 2014).

El cimiento de una BD están representado por el modelo de datos el cual es definido en Morteo *et al.* (2007) como una "representación de la realidad, descrita mediante un determinado formalismo"; además hace referencia a Edgar F. Codd, el padre del modelo relacional, quien lo describe como "una combinación de tres componentes: una colección de estructuras de datos (los bloques constructores de cualquier base de datos que conforman el modelos); un colección de operadores o reglas de inferencia, ……, para consultar o derivar datos de cualquier parte de estas estructuras en cualquier combinación deseada; una colección de reglas generales de integridad, la cuales implícita o explícitamente definen un conjunto de estados consistentes, algunas veces referidas como reglas de actualizar, insertar y borrar". Al respecto Date (2001) comenta que es importante distinguir entre el modelo de datos y su implementación, que consiste en una realización física de los componentes del modelo datos.

El modelo de datos entidad-relación es considerado por Silberschatz *et al.* (2002) como el modelo de datos más usado, ya que proporciona una representación visual de los datos, sus

relaciones y sus restricciones; además de éste modelo, existen otros como son: el modelo relacional, el de datos orientados a objetos, el relacional orientado a objetos, modelos de datos semiestructurados. Más adelante se explica con mayor detalle los modelos de datos entidad-relación, relacional y orientado a objetos.

2.1 Modelo Entidad-Relación

Este tipo de modelos de base se datos contempla tres conceptos importantes, las entidades, las relaciones y los atributos; Korth (2006, en Morteo *et al.*, 2007), lo define como "Una cosa u objeto en el mundo real que es distinguible de todos los demás objetos".

Ricardo (2004) y Ramos Martín *et al.* (2006) mencionan que éste modelo fue desarrollado por P. P. Chen en 1976 con la finalidad de facilitar el diseño de bases de datos y que de manera visual se representan mediante una gráfica generada a partir de la unión de los siguientes objetos: rectángulos que representan a los conjuntos de entidades, elipses que representan a los atributos de los conjuntos de entidades, rombos que representan las relaciones entre conjuntos de entidades, y líneas que son usadas para unir los atributos con los conjuntos de entidades y los conjuntos de entidades con las relaciones. Un ejemplo de diagrama entidad relación, se muestra en la Figura 1.

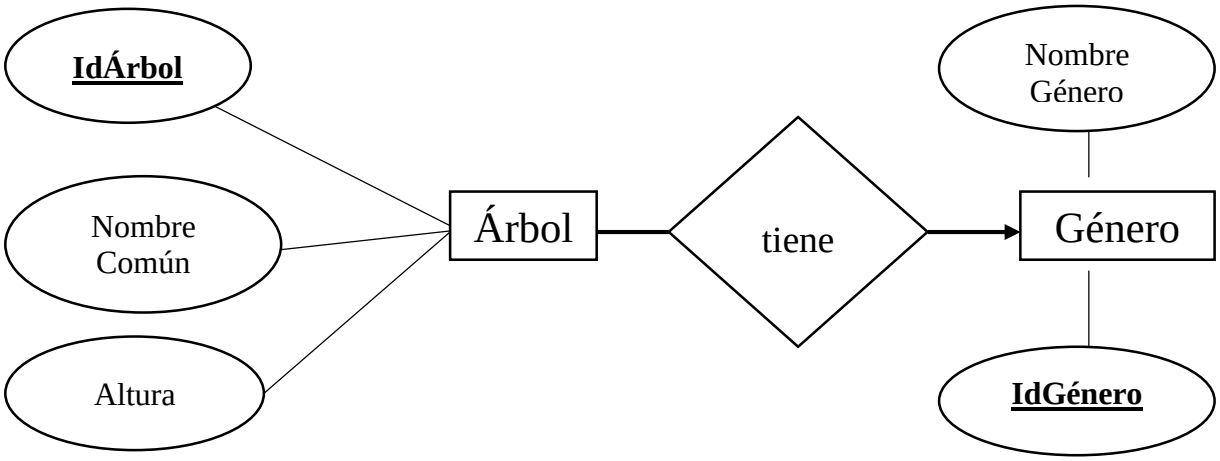


Figura 1. Diagrama Entidad-Relación
Fuente: Elaboración propia para la investigación, a partir de la base de datos del Inventario Nacional Forestal y de Suelos 2004-2009.

En la Figura 1 se puede apreciar dos conjuntos de entidades **Árbol** y **Género** vinculadas mediante la relación **tiene**. El conjunto de entidades **Árbol** está descrito por tres atributos **IdÁrbol, Nombre común** y **Altura,** mientras el conjunto de entidades **Género** está descrito por los atributos **IdGénero** y **Nombre Género**.

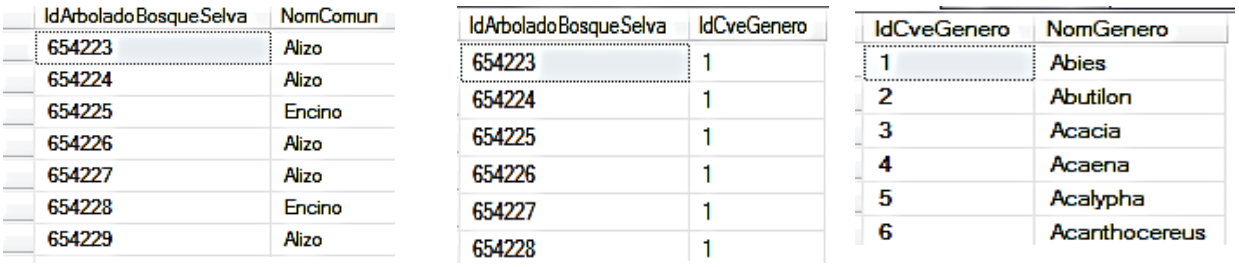
La correspondencia de cardinalidad es una de las restricciones impuestas a las bases de datos relacionales, para garantizar la integridad de los datos, y que expresa el número de entidades con las que otra entidad se puede asociar vía un conjunto de relaciones; además de que va a estar en función de las restricciones implícitas en los procesos que modele la base de datos (De Miguel *et al.*, 2000). Por ejemplo en la Figura 1, la cardinalidad está representada por la flecha que va en dirección de la relación **tiene** hacia el conjunto de entidades **Género**, lo cual se traduce en lo siguiente, un árbol puede pertenecer a un género en particular y no a más de uno, mientras que un mismo género puede estar asociado a varios árboles (la punta de la flecha indica una sola asociación, mientras que el otro lado representa una asociación múltiple).

2.2 Modelo relacional

Date (2001) considera que el modelo relacional, propuesto por Edgar F. Codd en 1970, es el evento más importante en toda la historia de las bases de datos, ya que éste modelo está sólidamente fundamentado en la lógica y en las matemáticas, de hecho considera el término relación como el término matemático para tabla; además comenta que "El usuario de un sistema relacional sólo ve tablas y nada más que tablas, mientras que el usuario de un sistema no relacional ve otras estructuras de datos pudiendo incluir o no tablas".

En el modelo relacional se utiliza un grupo de tablas para representar los datos y las relaciones entre ellos; cada tabla está compuesta por columnas, y cada columna se diferencia de otra mediante un nombre único (Silberschatz *et al.* 2002 y Morteo *et al.* 2007).

En la Figura 2 se muestran el conjunto de entidades **Árbol** y G**énero** con sus respectivos atributos y la relación **tiene**, presentes en la Figura 1, sólo que en ésta se representa, el conjunto de entidades y las relaciones, mediante tablas



IdArboladoBosqueSelva	NomComun
654223	Alizo
654224	Alizo
654225	Encino
654226	Alizo
654227	Alizo
654228	Encino
654229	Alizo

a) Árbol

IdArboladoBosqueSelva	IdCveGenero
654223	1
654224	1
654225	1
654226	1
654227	1
654228	1

b) tiene

IdCveGenero	NomGenero
1	Abies
2	Abutilon
3	Acacia
4	Acaena
5	Acalypha
6	Acanthocereus

c) Género

Figura 2: Estructura de datos en un modelo relacional.
Fuente: Elaboración propia a partir de la base de datos del Inventario Nacional Forestal y de Suelos 2004-2009.

Hansen y Hansen (2002) menciona que las restricciones como la correspondencia de cardinalidad, puede ser un poco más difícil de visualizar comparado con el esquema entidad relación y se refleja en la tabla que representa la relación, ya que ésta debe tener el identificador del conjunto de entidades que relaciona. De Miguel *et al.* (2000) consideran que cada entidad debe tener un identificador único, comúnmente llamada clave primaria, y que la columna que hace referencia a la relación muchos permanece y la que hace referencia a la relación uno duplica sus valores, como se puede observar en la tabla **tiene** de la Figura 2.

2.3 Modelo de datos orientado a objetos

Bertino y Martino (1995) explican que estos modelos no están orientados al área comercial o administrativa, "su orientación es hacia las ingenierías, sistemas multidimensionales y sistemas de gestión de imágenes", ya que el modelo de datos orientado a objetos permite el almacenamiento de estructuras de los datos complejas y las operaciones se definen en función de las necesidades de las aplicaciones. Mannino (2007) menciona que lo que impulsa la demanda de esta tecnología es la necesidad de almacenar grandes cantidades de datos complejos y la integración de estos con datos simples.

Ramos Martín *et al.* (2006) se refieren a un objeto como una combinación de datos y procedimientos, asociados con un conjunto de variables o atributos que contienen la información del objeto, un conjunto de mensajes a los que responde y conjunto de métodos que son bloques de código que responden los mensajes con un valor, adicionalmente Reinosa menciona que los

objetos son definidos mediante una estructura (OID, Constructor, Estado), donde el OID es un identificador del objeto creado por el sistema.

Reinosa *et al.* (2012) agrupan en clases a los objetos que responden a los mismos mensajes, por estar compuestos por los mismos atributos y contener los mismos métodos, y cada objeto representa una instancia de esa clase (el concepto de clase en el modelo orientado a objetos corresponde al concepto de entidad en el modelo Entidad-Relación aplicado en las bases de datos relacionales).

Morteo *et al.* (2007) explican que este modelo está basado en el paradigma de la programación orientada a objetos, lo que permite utilizar algunos principios de éste lenguaje de programación tales como: el encapsulamiento, la herencia y el polimorfismo.

El encapsulamiento significa que sólo se puede acceder a los objetos mediante sus interfaces, lo que implica que la definición de las variables y la implementación de los métodos no son accesibles, y el polimorfismo permite que un método tenga múltiples implementaciones (Mannino, 2007).

Reinosa *et al.* (2012) consideran que los objetos que se parecen en algunas cosas, pero difieren en otras hace posible aplicar el concepto de herencia, que permite que una clase pueda heredar de otra de mayor nivel parte de su composición (se reutiliza el código evitando codificar todos los elementos de la clase de menor jerarquía, ya que esos son heredados de la de mayor jerarquía).

2.4 Resumen

En el presente capítulo se describió brevemente la historia detrás de las bases de datos, se introdujo los conceptos más importantes relacionados con ésta y se explicó tres de los modelos de datos más utilizados actualmente a nivel mundial, un resumen de las principales características de estos modelos se presentan en los siguientes párrafos.

La definición de base de datos (BD), tomada desde la perspectiva de diferentes autores como Ricardo (2004) y Elmasri y Navathe (2007) hacen referencia a un conjunto de datos que guardan una relación entre sí y que son almacenados de manera sistemáticamente, para ser usados en diferentes aplicaciones. Las tareas relacionadas con almacenar, recuperar, modificar, consultar y actualizar la información dieron paso al desarrollo de Sistemas Gestores de Bases de datos (SGBD). Ricardo (2004) define a un SGBD como "un paquete de software que configura estructuras de almacenamiento, carga datos, proporciona acceso a programas y usuarios interactivos, formatea datos recuperados, oculta ciertos datos, realiza actualizaciones, controla concurrencia, y efectúa respaldos y recuperación para la base de datos".

Lo más destacable del modelo de datos Entidad-Relación lo presentan Ricardo (2004) y Ramos Martín *et al.* (2006) quienes explican que éste modelo fue desarrollado por P. P. Chen en 1976 con la finalidad de facilitar el diseño de bases de datos, que consiste en la presentación visual del modelo mediante una gráfica generada a partir de la unión de los siguientes objetos: rectángulos que representan a los conjuntos de entidades, elipses que representan a los atributos de los conjuntos de entidades, rombos que representan las relaciones entre conjuntos de entidades, y líneas que son usadas para unir los atributos con los conjuntos de entidades y los conjuntos de entidades con las relaciones.

Date (2001) considera que el modelo relacional, es el evento más importante en toda la historia de las bases de datos, ya que éste modelo está sólidamente fundamentado en la lógica y en las matemáticas, de hecho considera el término **relación** como el término matemático para tabla. En el modelo relacional se utiliza un grupo de tablas para representar los datos y las relaciones entre ellos; cada tabla está compuesta por columnas, y cada columna se diferencia de otra mediante un nombre único.

Bertino y Martino (1995) explican que los modelos de datos orientados a objetos no fueron desarrollados para al área comercial o administrativa, "su orientación es hacia las ingenierías, sistemas multidimensionales y sistemas de gestión de imágenes", ya que el modelo de datos orientado a objetos permite el almacenamiento de estructuras de los datos complejas y las operaciones se definen en función de las necesidades de las aplicaciones.

3. DATA WAREHOUSE

En la presente sección se presentan varias definiciones de Data Warehouse (DW) que nos acercan a una definición apropiada para el desarrollo de la presente investigación. Para comenzar, Date (2001) y Sumathi y Sivanandam (2006) explican que las grandes empresas se encuentran siempre a la vanguardia de los desarrollos tecnológicos con recursos suficientes para adquirir equipos de alto rendimiento y capacitar a sus operadores. El uso que le dan a las bases de datos también evoluciona y cada vez se extrae mayor información de ellas y como ejemplo mencionan que ahora son utilizadas como apoyo para la toma de decisiones.

Berry y Linoff (2004) y Jiawei Han *et al.* (2012) coinciden con Date (2001) y Sumathi y Sivanandam (2006) al mencionar que el avance tecnológico, principalmente en los medios de almacenamiento, ha permitido que las instituciones del sector privado y de gobierno acumulen una gran cantidad de datos correspondientes a las actividades operacionales diarias, éste tipo de datos son conocidos, en la teoría de base de datos, como datos operacionales.

Date (2001) define los datos operacionales como datos persistentes, considerando que el término "operacionales", refleja el uso que se les da a los sistemas de bases de datos, en las aplicaciones operacionales o de producción. La principal tarea de los sistemas operacionales son las transacciones en línea, como por ejemplo los depósitos y retiro de efectivo en un sistema bancario; éstas tareas, por la forma en se llevan a cabo se conocen en la literatura como Procesamiento de Transacciones en Línea (OLTP). En el ámbito empresarial, se requiere que los datos recolectados mediante OLTP sean analizados para la toma de decisiones. Algunos sistemas proporcionan las herramientas para la elaboración de reportes que los ejecutivos necesitan para la toma de decisiones, cálculos complejos, consultas no programadas, entre otras utilidades; estos sistemas son conocidos como sistemas de Procesamiento Analítico en Línea (OLAP).

Fayyad *et al.* (1996), Jiawei Han y Micheline Kamber (2006), y Elmasri y Navathe (2007) hacen referencia a Edgard Codd como el primero en usar el término OLAP en 1993.

Reinosa *et al.* (2012) agrega que la característica fundamental de los sistemas OLAP consiste en ofrecer opciones para el modelado analítico, que incluye un motor de cálculo para la obtención de proporciones, desviaciones, promedios, entre otros, con el objetivo de elaborar resúmenes y adiciones, conocidos como consolidaciones. OLAP es ampliamente superior, en cuanto al procesamiento de consultas multitablas, con respecto a la bases de datos relacionales, al utilizar estructuras multidimensionales como almacenamiento.

Date (2001) menciona que las empresas cuentan con dos bases de datos, una para llevar a cabo sus operaciones cotidianas (depósitos, retiros, ventas, etc.) y otra para llevar a cabo un análisis de negocio mediante herramientas OLAP, a ésta base de datos se le llama Data Warehouse y consiste en información agregada a partir de los datos operaciones, como por ejemplo totales y promedios que se actualizan periódicamente a partir de la base de dato operacional.

El párrafo anterior permite que Ruíz Torres (2007) defina un Data Warehouse (DW) como un repositorio de datos centralizados para poyo de las actividades de análisis, que permite almacenar datos operacionales y eliminar inconsistencias entre los diferentes formatos existentes en los sistemas fuente. Además de integrar los datos de interés, permite incorporal información adicional integrada por un experto.

Inmon (1992, en Reinosa *et al.* 2012) e Inmon (2002, en Ricardo, 2004) define los DW como "una colección de datos orientados al sujeto, no volátiles, integrados y variantes en el tiempo, usados principalmente en la toma de decisiones organizacionales".

Ricardo (2004), Ruíz Torres (2007), Elmasri y Navathe (2007), mencionan a William H. Inmon, como el primero en introducir la terminología Data Warehouse, en 1992, lo que le hizo ganar el título de "padre del Data Warehouse".

Inmon (2005) presenta una descripción de cada una de las características que forma la definición original.

- Orientado a temas.- Los datos se organizan en áreas específicas de manera que todos los elementos de datos relativos al mismo evento u objeto del mundo real queden unidos entre sí.

- Variante en el tiempo.- Los cambios producidos en los datos a lo largo del tiempo permanecen para poder ser utilizados en diferentes análisis (comparaciones, tendencias, pronósticos).
- No volátil.- Una vez que los datos son almacenados, no se pueden realizar cambios ni modificaciones a lo largo del tiempo, permanecen de la misma forma a pesar de la evolución de los mismos, y puede ser accedidos como sólo lectura.
- Integrado.- Los datos se integra a partir de diferentes fuentes, de todos los niveles operacionales y mediante un proceso de depuración son eliminadas las inconsistencias presentes, como redundancia y dominio, para finalmente integrarse en un formato único.

Kimball y Ross (2002), definen a los DW como "una copia de las transacciones de datos específicamente estructurada para la consulta y el análisis". Kimball al igual que Inmon fue de los primeros personajes relacionados a la teoría DW y fue él, quién primeramente consideró a los DW como la unión de todos los Data Marts (DM) de una entidad; este enfoque es también conocido como Data Warehouse Bus.

Elmasri y Navathe (2007) consideran que los Data Warehouse (DW), el procesamiento analítico en línea y la minería de datos ofrecen todas las funcionalidades necesarias para explotar las grandes bases de datos; principalmente, hace el señalamiento de que el DW es una de las tecnologías más importantes para obtención de información, al hacerlo sin sacrificar el rendimiento de las aplicaciones operacionales.

Susan Osterfeldt (2003, en Reinosa *et al.* 2012) proporciona una definición más acorde a los propósitos de la presente tesis, al considerar al DW como algo que provee dos beneficios organizacionales reales: Integración y acceso a datos. Ella hace énfasis en que los DW eliminan una gran cantidad de datos inútiles y no deseados, así como que también provee el procesamiento desde el ambiente operacional clásico.

El gran número de definiciones que dan los diferentes autores, a los Data Warehouse, varía de acuerdo a los objetivos que persiguen con su implementación. De esta manera, un DW, se define en esta tesis, como una base de datos cuya principal característica es la integración, el filtrado, agrupamiento o resumen de la información, desde una o varias fuentes que pueden esta en

múltiples formatos, para su análisis con diferentes herramientas y con gran velocidad de respuesta. En la siguiente sección se detalla las similitudes y diferencias encontradas en torno a los dos principales enfoques de la teoría Data Warehouse desarrollada por William H. Inmon y Ralph Kimball.

3.1 Arquitectura de un Data Warehouse

Jiawei Han y Micheline Kamber (2006), Elmasri y Navathe (2007) y Jiawei Han *et al.* (2012) concuerdan en que desde el punto de vista de la arquitectura, existen tres modelos posibles para diseñar un Data Warehouse (DW): el Data Warehouse Empresarial, los Data Marts y el Data Warehouse Virtual.

- El Data Warehouse Empresarial consiste en una colección de toda la información de la organización.
- Los Data Marts (DM) consisten subconjuntos de información agregada a partir de los requerimientos específicos de un grupo o departamento de la organización; este puede ser categorizado como dependiente o independiente en función de su fuente de datos.
- El Data Warehouse Virtual consiste en un conjunto de vistas generadas a partir de la base de datos operacional para hacer más eficiente los procesos de consulta.

Inmon (2005) presenta las principales características a considerar en el diseño e implementación de un Data Warehouse (DW) empresarial. Éste enfoque también es conocido como "diseño de arriba hacia abajo" considera a los DW como el lugar una empresa logra la integración de su información. Primeramente se diseña una base de datos normalizada que corresponde al DW, y posteriormente se crean los Data Marts (DM), a partir del DW, que contiene toda la información requerida para un proceso de negocio o departamento específico, es decir, la información es almacenada al más bajo nivel de detalle. Para Inmon un DW consiste en tener una fuente de datos unificada que contenga toda la información posible y establece que éste debe ser la fuente de información para todos los DM. En la Figura 3 se muestra la estructura de un DW desde el enfoque de Inmon.

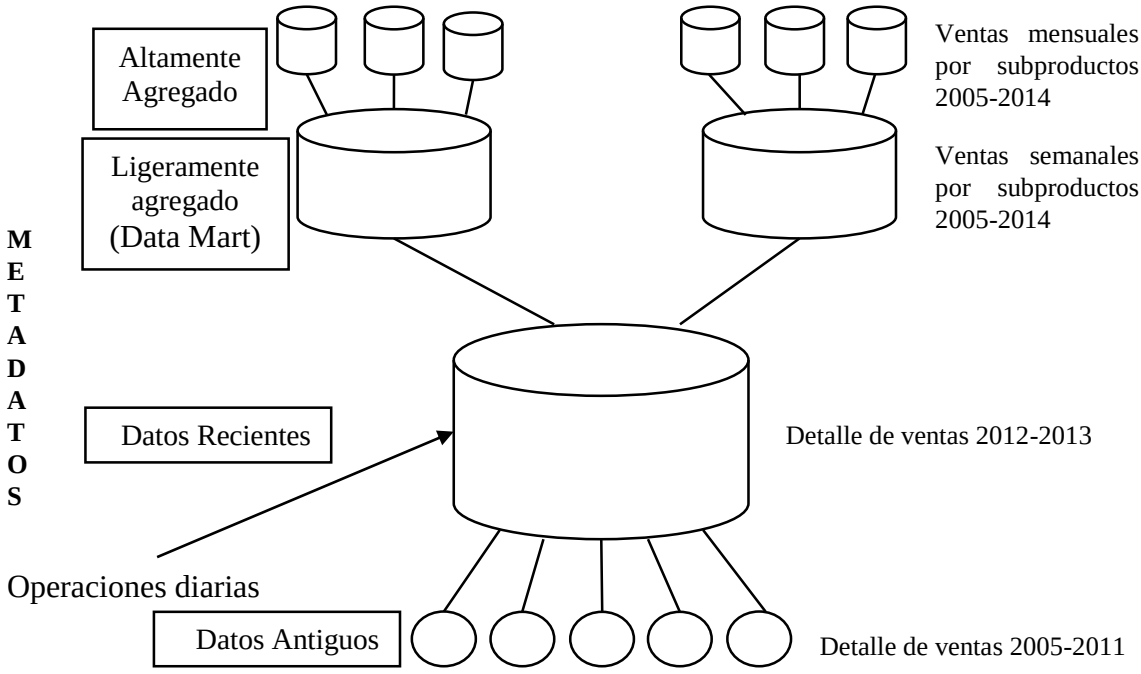


Figura 3: Estructura del Data Warehouse, desde el enfoque de Inmon.
Fuente: Modificado del original de Inmon (2005)

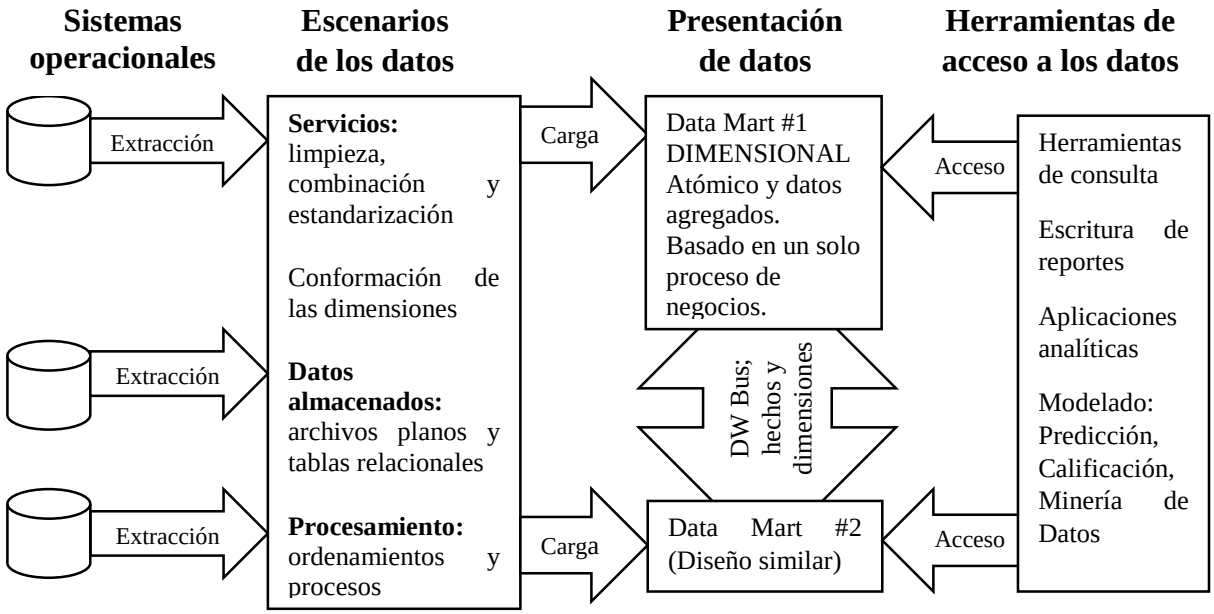


Figura 4: Elementos básicos de un Data Warehouse, desde el enfoque de Kimball.
Fuente: Modificado del original de Kimball y Ross (2002)

Kimball y Ross (2002) detallan las principales características del enfoque conocido como diseño de abajo hacia arriba o Data Warehouse Bus. En éste enfoque los DM que facilitan los reportes

y análisis de información son creados primero, y luego son integrados en un DW que comparte las dimensiones y medidas; es decir, Kimball describe DW como un concepto virtual creado como un bus en donde se conectan todos los DM. Para Kimball un DW tienen como objetivo hacer que la información se encuentre disponible para un análisis lo más rápido y eficiente posible. En la Figura 4 se logra apreciar los elementos básicos que conforman un Data Warehouse, desde el enfoque de Kimball.

Las principales características tomadas de los dos enfoques se presentan en el Cuadro 2.

Cuadro 2. Principales características de los enfoque de Inmon y Kimball

Característica	Inmon	Kimball
Diseño del DW	Toma mucho tiempo	Toma menos tiempo
Tiempo para poner en marcha el proyecto	Un largo periodo	No requiere mucho tiempo
Costo de implementación	Costos iniciales muy altos, subsecuente al desarrollo del proyecto los costos serán más bajos	Bajos costos iniciales y en cada etapa subsecuente, los costos al menos serán los mismos.
Integración de datos	Toda la empresa	Áreas específicas del negocio
Mantenimiento	Fácil	Dificultoso, siempre es redundante y sujeto a revisiones

Fuente: Elaboración propia a partir de Kimball y Ross (2002) e Inmon (2005).

En la Figura 5 se muestran de manera gráfica los tres niveles que Jiawei Han y Micheline Kamber (2006), y Krzysztof *et al*. (2007) mencionan como importantes en el diseño de un DW para una solución de negocios. El nivel inferior, o primer nivel, está integrado por el servidor del DW que contiene información resumida, los metadatos y los Data Marts. En el segundo nivel se encuentra el servidor de procesos analíticos en línea (servidores OLAP) que incluye las herramientas multidimensionales como los cubos OLAP para agilizar las consultas. En el último nivel, tercer nivel, se encuentran las interfaces orientadas a los usuarios, que extraen la información del servidor OLAP para la toma de decisiones, (hojas de cálculo, modelo de minería de datos y otros medios de visualización).

Otro elemento importante en el diseño de un DW, y en toda base de datos, son los metadatos ya que de acuerdo con Jiawei Han *et al.* (2012) y Reinosa *et al.* (2012) en éstos se encuentra la información de los dominios, las reglas de validación, la derivación y la conversión de los datos extraídos.

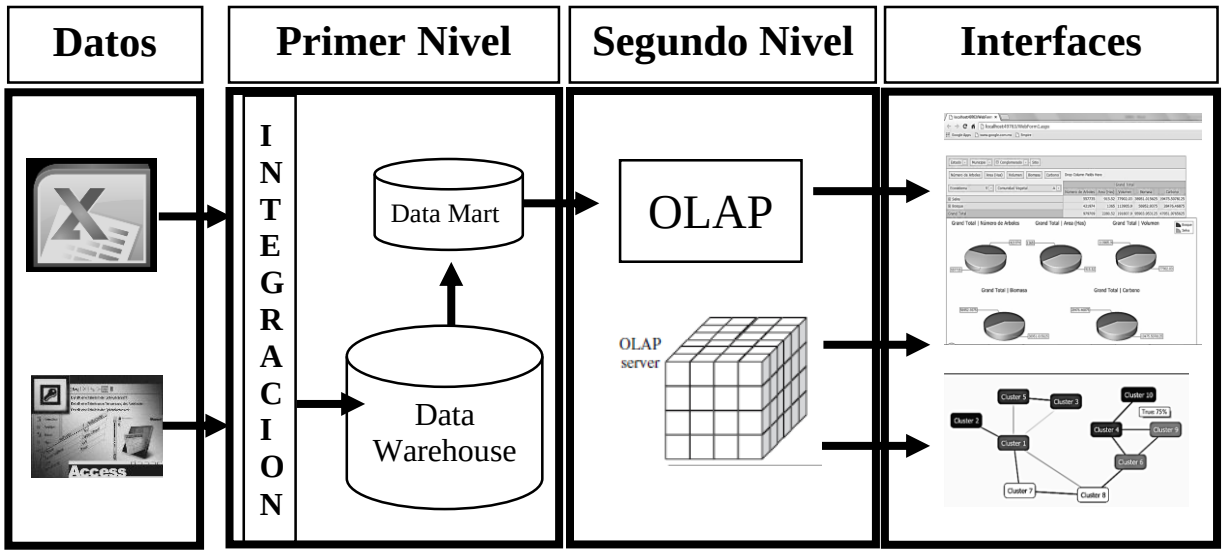


Figura 5: Niveles en la arquitectura Data Warehouse
Fuente: Elaboración propia para la investigación

La estructura interna de los DW se define usando un esquema de almacenamiento, de forma similar al de las bases de datos relacionales (Berry y Linoff, 2004). Este esquema puede ser alguno de los usados para bases de datos multidimensionales como son: los esquemas de estrella, copo de nieve o de constelaciones, dependiendo del o los procesos que se deseen modelar; información más detallada sobre estos modelos se abordan en la siguiente sección dentro de esquemas para la generación de modelos ROLAP.

3.2 Modelo de datos para un Data Warehouse

Inmon (2005) menciona que una de las tecnologías que siempre se discute dentro del contexto de los Data Warehouse son los procesos llevados a cabo mediante bases de datos multidimensionales, o Data Marts. Los datos almacenados de forma detallada en un Data Warehouse son considerados como una fuente robusta de datos, muy conveniente para los sistemas de bases de datos multidimensionales. En la Figura 6 se muestra la relación propuesta por Inmon entre el Data Warehouse y los Data Marts los cuales están representados por base de datos multidimensionales.

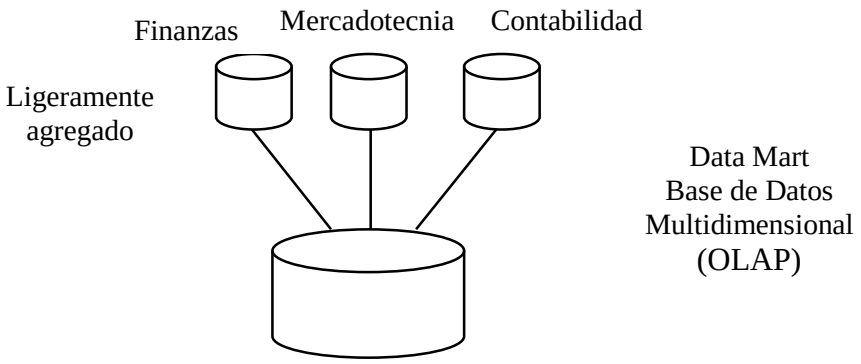


Figura 6: Estructura clásica del Data Warehouse.
Fuente: Modificado del original de Inmon (2005).

Carpani (2000) explica que en una base de datos multidimensional, la información se presenta como matrices multidimensionales, cuadros de múltiples entradas o funciones de varias variables sobre conjuntos finitos, cada una de estas matrices se conoce como cubo, cuanto más dimensiones posea una base de datos multidimensional, mayor será el número de datos o celdas. El cubo consta de tres dimensiones, cuando son más de tres se les conoce como hipercubos.

Un cubo OLAP es una estructura de datos que supera las limitaciones de las bases de datos relacionales y proporciona un análisis rápido de datos. Los cubos pueden mostrar y sumar grandes cantidades de datos, y cuentan con la capacidad para resumir y reorganizarlos según sea necesario (Berry y Linoff, 2004).

El esquema de un cubo queda determinado al conocer sus ejes y sus respectivas estructuras, tomando en cuenta que los datos en todas las celdas deben ser uniformes. A los ejes se les llama *dimensiones* y al dato que se presenta en la matriz se le llama *medida*, mientras que a los elementos del producto cartesiano se le conoce como *coordenada* (Jiawei Han *et al.*, 2012).

En la Figura 7 se muestra el potencial de uso de los cubos OLAP, para contestar preguntas específicas, en un área como la de negocios. El cubo del ejemplo consta de tres dimensiones: clientes, productos y tiempo; que a través de diferentes operaciones, que se verán más adelante, se accede a la información que podría contestar una gran número de preguntas, tanto generales como particulares, por ejemplo:

- ¿Cuántas muñecas fueron vendidas en el primer semestre?

- ¿Cuántas muñecas fueron vendidas en el tercer trimestre?
- ¿Cuántas muñecas fueron vendidas en el tercer trimestre a las jugueterías?

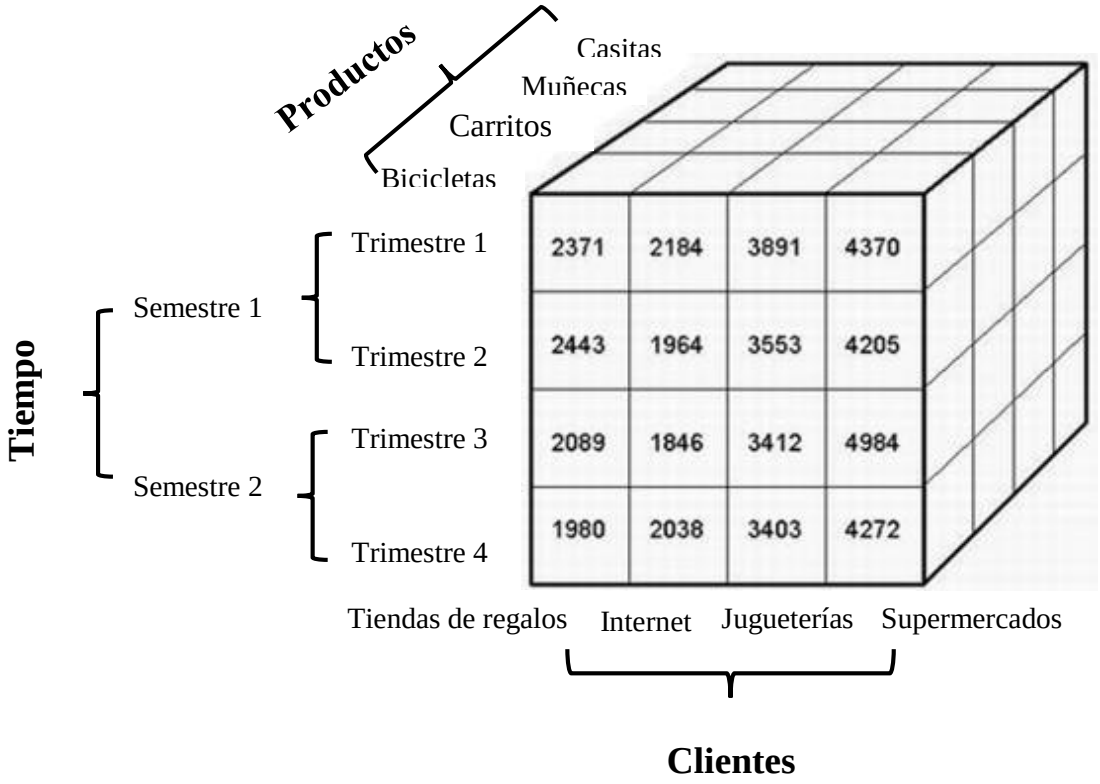


Figura 7. Lógica de almacenamiento de información en bases de datos multidimensionales.
Fuente: Elaboración propia para la investigación

3.2.2 Dimensiones y jerarquías

Cada una de las dimensiones de un cubo OLAP puede resumirse mediante una jerarquía. La Figura 8, es una representación gráfica de una dimensión que es utilizada para ejemplificar el concepto de jerarquía. Por ejemplo la fecha "Mayo de 1998" puede incluirse en una categoría mayor denominada "Segundo Trimestre de 1998", que a su vez se incluye en "Año 1998". Sumathi y Sivanandam (2006) consideran que el conjunto de todos estos niveles de categorías definen una dimensión, para el ejemplo la dimensión tiempo, con diferentes niveles de jerarquía (cada nivel de jerarquía se conoce como nivel de agregación).

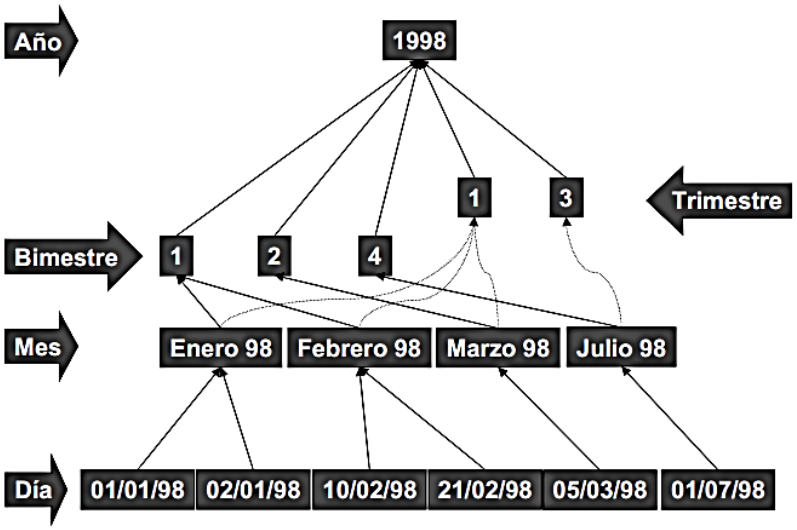


Figura 8. Nivel de jerarquía para una dimensión tiempo.
Fuente: Carpani (2000)

Una dimensión puede tener más de una jerarquía; por ejemplo, considerando nuevamente la dimensión tiempo, los niveles, bimestre y trimestre no están relacionados entre sí, como se muestra en la Figura 8, aunque estén relacionados con mes y año. Cuando se presenta ésta situación, la relación que une datos de un nivel con datos de otro nivel superior debe ser confluente, es decir, todos los caminos que parten de un elemento *e* del nivel E, llegan al mismo elemento *d* del nivel *D*, superior al nivel *E* (Carpani, 2000).

Reinosa *et al.* (2012) menciona que las "dimensiones se pueden estructurar jerárquicamente para construir caminos de consolidación que analice la información, desde lo más general hasta lo más específico, mediante operaciones como **drill dow** o viceversa **roll up**". En la siguiente sección se amplía en la descripción de las operaciones que se pueden llevar a cabo con un cubo multidimensional.

3.2.3 Operaciones multidimensionales.

Las operaciones que se pueden realizar en un cubo multidimensional se describen a detalle en Berry y Linoff (2004), Sumathi y Sivanandam (2006) y Krzysztof *et al.* (2007). Un bosquejo de los que se puede realizar con éstas operaciones se resumen a continuación:

- **Roll up** se refiere al incremento en el nivel de agregación de los datos.
- **Drill down** se refiere al incremento en el nivel de detalle, opuesto a Roll Up.
- **Slice** se refiere a la reducción de la dimensionalidad de los datos mediante selección (selecciona dimensiones de trabajo a partir de un cubo mayor, permitiendo enfocarse en una sola parte de los datos del cubo).
- **Dice** se refiere a la reducción de la dimensionalidad de los datos mediante proyección (selecciona secciones del cubo en función de valores de las dimensiones, puede ser considerado como un slice más pequeño).
- **Pivotaje o rotación** se refiere a la reorientación de la visión multidimensional de los datos (permite presentar diferentes planos de un cubo).

La Figura 9 muestra las operaciones más comunes en un cubo multidimensional.

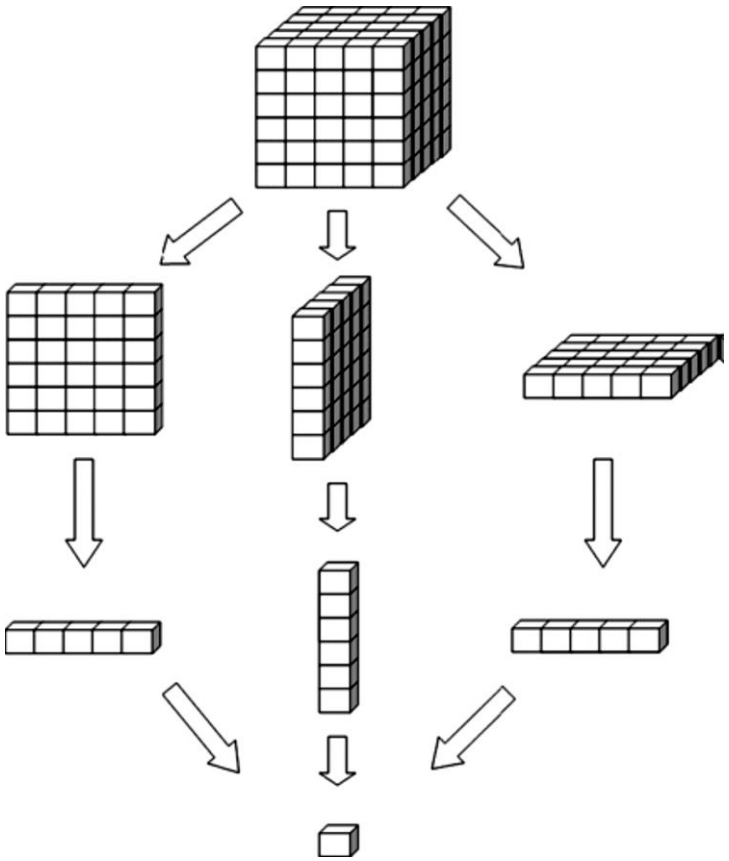


Figura 9: Posibles selecciones y proyecciones para un cubo de datos tridimensional.
Fuente: Krzysztof *et al.* (2007)

Carpani (2000) agrupa a estas operaciones y otras más, dentro de tres categorías: selección y visualización, agregación y relacionamiento. En la categoría de selección y visualización se encuentran las operaciones Slice, Dice y pivotaje. En la categoría de agregación se incluye aquellas operaciones que resultan de realizar movimientos en las jerarquías de un cubo, dentro de estas se encuentra el Roll up y el Drill down. Finalmente, en la categoría de relacionamiento, se incluyen a aquellas operaciones que permiten acceder a datos no contenidos en el cubo; si estos datos están en otro cubo la operación se conoce como Drill Across y si están en el Data Warehouse o base en la base de datos operacional entonces es conocida como Drill Through. La Figura 10 muestra algunas operaciones de los tres grupos mencionados en este párrafo.

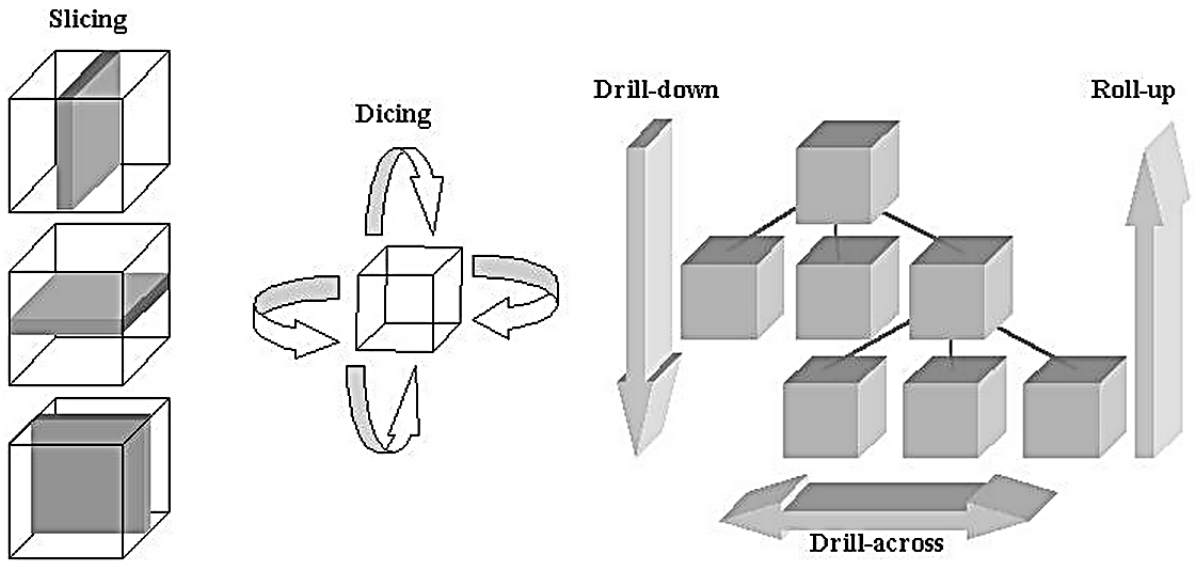


Figura 10. Operaciones comunes en un cubo multidimensional.
Fuente: http://wappreview.de/tino/projekt_handel_III/datawrhs/slice.htm

3.2.4 Diferenciación de modelos OLAP

Reinosa *et al.* (2012) hacen una diferenciación de los modelos OLAP de acuerdo a la forma en que se almacenan los datos. En un OLAP relacional o ROLAP, los datos se almacenan físicamente en un Sistema Gestor de Bases de Datos Relacional; mientras que un OLAP multidimensional o MOLAP se almacenan en una base de datos multidimensional. Un híbrido generado a partir de estos dos modelos son los HOLAP, en los que se mezclan ambas

arquitecturas, bases de datos multidimensionales y de datos relacionales. En internet pueden encontrarse otros acrónimos que no son muy reconocidos, por lo menos no tanto como los anteriores, por ejemplo:

- WOLAP o Web OLAP: OLAP basado u orientado para la web.
- DOLAP o Desktop OLAP: OLAP de escritorio
- RTOLAP o Real Time OLAP: OLAP en tiempo real
- SOLAP o Spatial OLAP: OLAP espacial

3.2.5 Esquemas para la generación de modelos ROLAP

Imhoff *et al*. (2003) sugieren que la mejor técnica para modelar un Data Warehouse se debe basar en el diseño original de una base de datos relacional, proponen el uso de un diagrama Entidad-Relación (ER) desarrollado por C. J. Date y E. F. Codd.

Algunos autores -como Berry y Linoff (2004), Sumathi y Sivanandam (2006), Jiawei Han *et al.* (2012) y Reinosa *et al.* (2012)- consideran que un modelo OLAP ofrece una visión multidimensional lógica de los datos, la cual permite realizar consulta de manera rápida sin interrumpir el proceso de análisis. La eficiencia en las consultas se debe a que durante el armado del modelo se efectúan los cálculos previos para la integración de la información. Los cálculos que se deben realizar está determinado por el diseño técnico de la base de datos, que puede ser de acuerdo a alguna de las siguientes estructuras: el esquema de estrella, el esquema de copo de nieve o el esquema de constelaciones. En las siguientes secciones se describe a detalle cada una de las estructuras mencionadas.

Esquema en estrella: Berry y Linoff (2004) mencionan que el esquema de estrella es una técnica de modelado de datos, desarrollada por Ralph Kimball. En el trabajo de Kimball y Ross (2002) éste esquema es utilizado para hacer corresponder un modelo multidimensional sobre una base de datos relacional; este modelo consta de cuatro componentes: tabla de hechos, dimensiones, atributos y jerarquías de atributos. La tabla de hechos, que contiene datos para análisis se encuentra al centro y se encuentra rodeada por tablas de dimensiones, semejando una

estrella. La tabla de hechos se vincula con cada dimensión mediante una relación uno a muchos (por medio de claves foráneas). En la Figura 11 se presenta un esquema de estrella donde la tabla de hechos se encuentra al centro y las dimensiones a su alrededor.

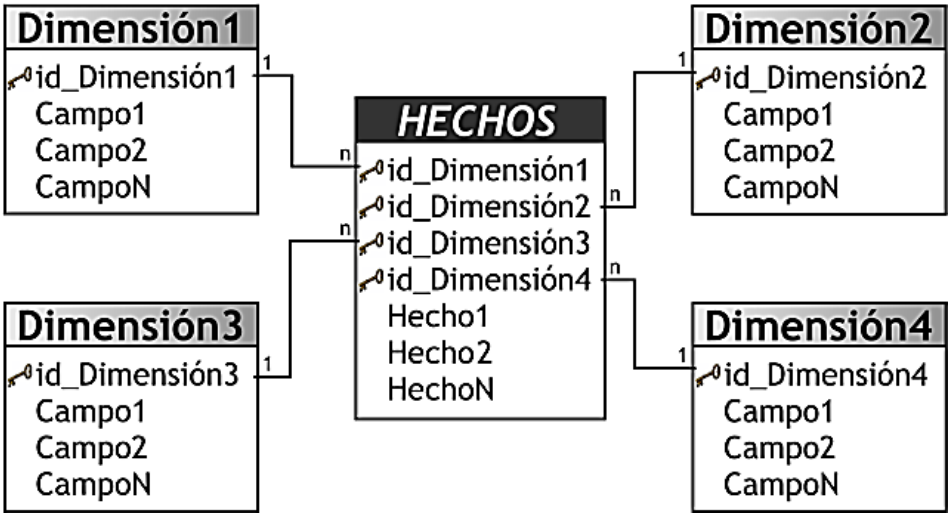


Figura 11. Esquema en estrella para modelar las ventas de una empresa.
Fuente: http://www.dataprix.com/data-warehousing-y-metodologia-hefesto/arquitectura-del-data-warehouse/34-datawarehouse-manager

Esquema Copo de Nieve: Elmasri y Navathe (2007) dice que un esquema de copo de nieve es una es una variación del esquema en estrella. En cada dimensión se almacenan las jerarquías de atributos o se dividen en otra entidad (más dimensiones y más relaciones entre tablas), con la finalidad de normalizar las tablas y así reducir el espacio de almacenamiento al eliminar la redundancia de datos; como resultado de ésta división, el rendimiento decrece al ejecutar consultas que necesiten acceder a múltiples tablas.

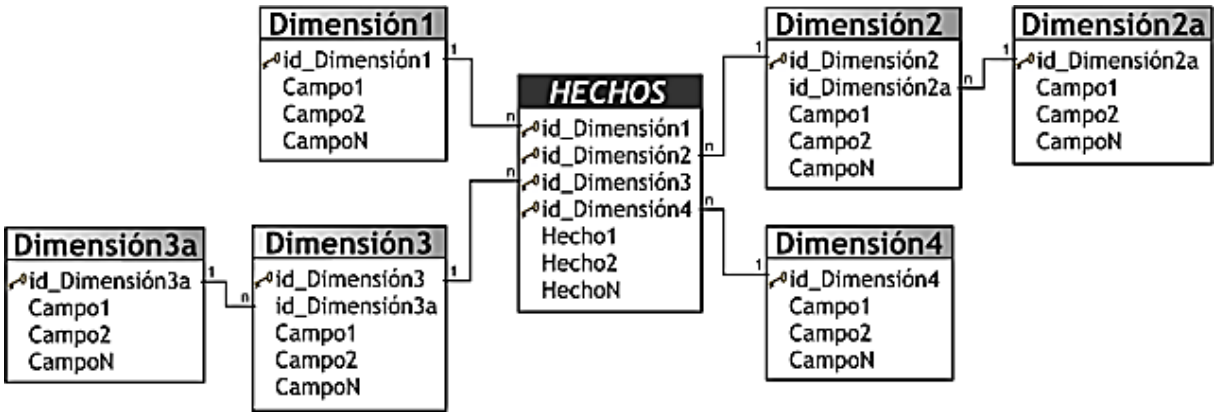


Figura 12. Esquema de copo de nieve para modelar las ventas de una empresa.
Fuente: http://www.dataprix.com/data-warehousing-y-metodologia-hefesto/arquitectura-del-data-warehouse/34-datawarehouse-manager

Un ejemplo del esquema de copo de nieve, se muestra en la Figura 12, en la cual, las dimensiones 2 y 3 son divididas en tablas relacionadas con las tablas 2a y 3a, respectivamente.

Esquema de constelaciones: Este esquema es más complejo que las otras arquitecturas debido a que contiene múltiples tablas de hechos (es una combinación de un esquema de estrella y un esquema de copo de nieve). Jiawei Han y Micheline Kamber (2006) y Jiawei Han *et al.*, (2012) mencionan que en los esquemas de constelaciones las tablas de dimensiones pueden estar compartidas entre más de una tabla de hechos. El objetivo de los esquemas de constelación radica en aprovechar las ventajas de los esquemas de estrella y de copo de nieve, por ejemplo, las jerarquías en los esquemas en estrella están desnormalizadas, mientras que en los esquemas de copo de nieve están normalizadas; los esquemas en constelación están normalizados para eliminar las redundancias de las dimensiones, mediante jerarquías dimensionales compartidas, IBM (http://pic.dhe.ibm.com/infocenter/idm/docv3/index.jsp? topic=%2Fcom.ibm.datatools.di

mensional.ui.doc%2Ftopics%2Fc_dm_schema_starflake.html). Sin embargo, el problema es que cuando el número de las tablas vinculadas aumenta, la arquitectura puede llegar a ser muy compleja y difícil para mantener. En la Figura 13 se muestra un esquema de constelación con dos *tablas de hechos* compartiendo la *dimensión2*.

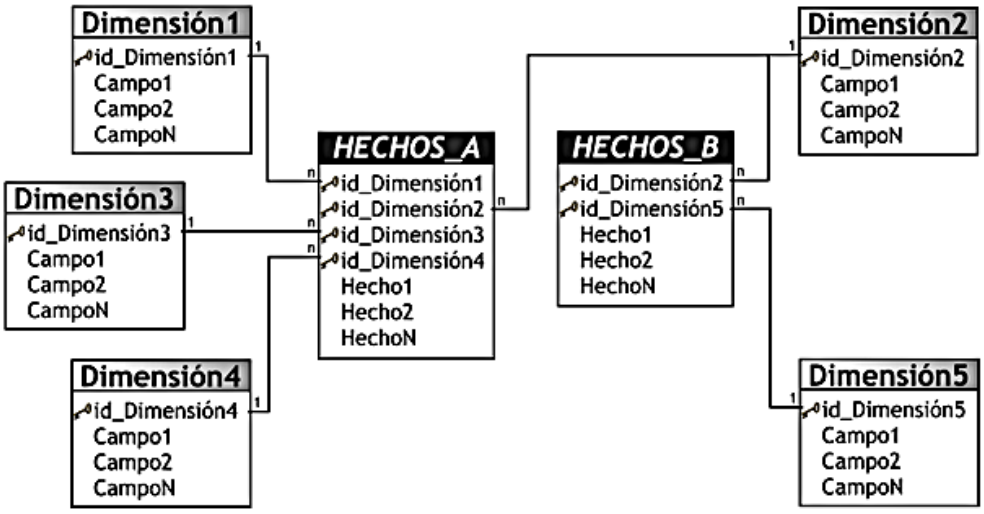


Figura 13. Esquema de constelación.
Fuente: http://www.dataprix.com/data-warehousing-y-metodologia-hefesto/arquitectura-del-data-warehouse/34-datawarehouse-manager

3.3 Modelado dimensional

El modelo dimensional captura las medidas de importancia y los parámetros, a través de los cuales dichas medidas son descompuestas, con se vio en la sección de bases de datos multidimensionales.

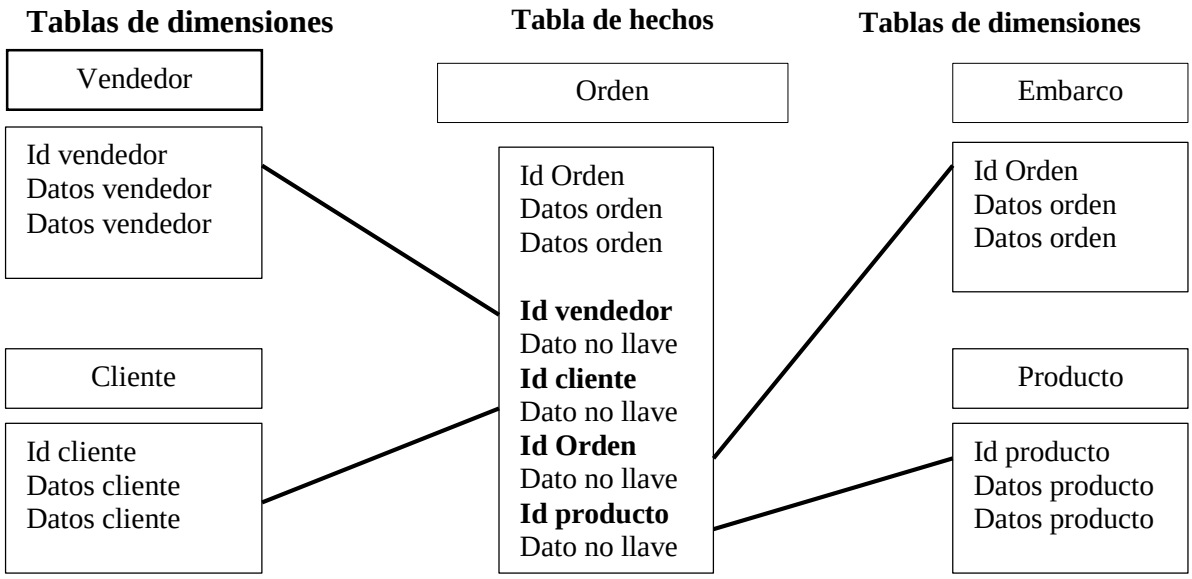


Figura 14: Estructura de las tablas de hechos y las tablas de dimensiones bajo el enfoque de Inmon.
Fuente: Modificado del original de Inmon (2005).

Las medidas son referidas como hechos o métricas, mientras que los parámetros usados para analizar una métrica son referidos como dimensiones. En la Figura 14 se muestra la estructura de las tablas de hechos y las tablas de dimensiones bajo el enfoque de Inmon.

Kimball y Ross (2002), Inmon (2005) y Ruíz Torres (2007) proporciona una serie de consejos, que se detallan en las siguientes subsecciones, para definir los diferentes niveles de granularidad o nivel de detalle, las tablas de dimensiones y las tablas de hechos.

3.3.1 Definición del nivel de Granularidad

Inmon (2005) explica que la granularidad se refiere al nivel de detalle o agregación de las unidades de datos dentro del Data Warehouse (DW), es decir, que los datos en un DW tienen

una estructura diferente debido a los distintos niveles de agrupación y resumen a los que son sometidos, lo que hace importante definir un adecuado nivel de detalle. En la mayoría de los casos los datos ingresados al DW tienen un alto grado de granularidad y sólo ocasionalmente se da el caso en que los datos son agregados o filtrados para ser integrados.

Kimball y Ross (2002) argumentan que la importancia del nivel de detalle radica principalmente en los objetivos de análisis, por lo que se debe garantizar la correcta integración de los datos, utilizando sólo la información requerida en los procesos de análisis y no más.

Algunas de las sugerencias para definir el nivel de granularidad de los datos se listan a continuación:

- Los hechos más recientes son de mayor interés.
- El volumen de almacenamiento aumenta con forme disminuye el nivel de granularidad.
- La unidad de tiempo, sobre la cual se va a resumir la información, debe ser considerada.

Los registros en las tablas de hechos se deben almacenar exactamente con el mismo nivel de detalle, es decir, la información relevante con un nivel de granularidad diferente debe guardarse en una segunda tabla de hechos. Reinosa *et al.,* (2012) coinciden con Ruíz Torres (2007) en que la información resumida se almacena de tal manera que se obtenga un acceso rápido a ella, puede ser en línea o en discos; la información de detalle con mayor antigüedad se guarda con algún proceso de almacenamiento masivo pero más económico, ya que su acceso no es tan frecuente; y los históricos se guardan en otros medios o son eliminados en dado caso de que ya no sean útiles para la toma de decisiones.

3.3.2 Definición de la Tabla de hechos

Kimball y Ross (2002) dicen que ésta tabla es la principal en un modelo multidimensional y se define por lo general después de identificar la granularidad. La tabla de hechos contiene métricas o hechos del negocio, representado por columnas numéricas que muestran la granularidad de los hechos almacenados en la tabla, además, incluye una llave única y varias llaves foráneas que la

conectan con las dimensiones, además de que incorpora un gran número de registros. El término hecho es usado para representar las medidas del negocio, es decir, que cada fila en la tabla de hechos corresponde a una medición. Lo hechos que son de mayor interés son los numéricos que presentan una propiedad aditiva.

Inmon (2005), menciona que el nivel de granularidad definido para la tabla de hechos, determinara no sólo el nivel de detalle en el que se puede consultar la información, sino también el número de cálculos que deben procesarse para proveer la respuesta a una consulta; mientras mayor detalle se pueda consultar, aumenta el tamaño de la tabla de hechos. En la mayoría de los casos, la tabla de hechos es poblada sólo por datos del tipo numérico y llaves foráneas.

3.3.3 Definición de las Dimensiones

Kimball y Ross (2002) explican que las dimensiones por lo general se definen después de que se ha definido la granularidad de la tabla de hechos; el objetivo de ésta es reflejar un conjunto de detalles en torno al proceso de negocios. Las tablas de dimensiones contienen descriptores textuales del negocio, en forma de columnas o atributos que son usados para describir las filas de la tabla de hechos. Los atributos de las dimensiones sirven como la fuente principal para las consultas, agrupaciones y etiquetas en los reportes. El comentario más importante que hacen es que la potencia del Data Warehouse es directamente proporcional a la calidad y profundidad de los atributos en las dimensiones.

Ruíz Torres (2007) afirma que las relaciones entre los atributos existentes es una dimensión no necesitan ser estrechas, por lo que es aceptable incluir un conjunto de tributos directamente relacionados con otros, como es el caso cuando se utiliza un esquema de copo de nieve o constelación en las que las dimensiones son normalizadas y compartidas. La llave primaria de una dimensión siempre es un atributo único definido por el sistema, debido a que el uso de llaves concatenadas degrada el desempeño y también evita la dependencia hacia los sistemas operacionales al asignar una nueva clave propia. Las características que le atribuye a las tablas de dimensión, son: contiene información textual descriptiva (nombres, dirección, estatus, etc.),

se utiliza como una fuente para las consultas, su relación es uno a muchos con la tabla de hechos, incluye un número limitado de registros que se incrementa lentamente con el tiempo.

3.4 Extracción, transformación y carga (ETL)

La limpieza de los datos es un aspecto importante para la creación de un Data Warehouse eficiente porque elimina la mayoría de las inconsistencias presentadas en los datos operacionales (Ruíz Torres, 2007 y Jiawei Han *et al.*, 2012). El Data Warehouse extrae datos desde una variedad de bases de datos heterogéneas y sistemas operacionales en intervalos regulares y la etapa de limpieza tiene que remover la duplicación y conciliar diferencias entre los distintos estilos de almacenamientos de datos, como pueden ser diferencias en dominios y formatos; después del proceso anterior, los datos son integrados al Data Warehouse. A todo el proceso anterior, de extracción, transformación y carga de datos se le conoce como ETL (Sumathi y Sivanandam, 2006 y Berry y Linoff, 2004). Por lo general la ejecución de estas funciones se efectúa de forma automática (Ruíz *et al.*, 2008).

3.5 Resumen

En el presente capítulo se propusieron varias definiciones y enfoques de la tecnología Data Warehouse (DW) siendo los de mayor peso los enfoques de William Inmon y Ralph Kimball; además se analizaron cada uno de estos enfoques comparando sus similitudes y diferencias. Se introdujo los conceptos más importantes de la teoría DW y su relación con las bases de datos multidimensionales. Todo lo anterior proporcionó las herramientas para dar una definición de DW como una base de datos cuya principal característica es la integración, el filtrado, agrupamiento o resumen de la información, desde una o varias fuentes que pueden esta en múltiples formatos, para su análisis con diferentes herramientas y con gran velocidad de respuesta.

Jiawei Han y Micheline Kamber (2006), y Krzysztof *et al*. (2007) mencionan tres niveles importantes en el diseño de un DW para una solución de negocios. El nivel inferior, o primer nivel, está integrado por el servidor del DW que contiene información resumida, los metadatos y los Data Marts. En el segundo nivel se encuentra el servidor de procesos analíticos en línea (servidores OLAP) que incluye las herramientas multidimensionales como los cubos OLAP para agilizar las consultas. En el último nivel, tercer nivel, se encuentran las interfaces orientadas a los usuarios, que extraen la información del servidor OLAP para la toma de decisiones, (hojas de cálculo, modelo de minería de datos y otros medios de visualización).

La estructura interna de los DW se define usando un esquema de almacenamiento, de forma similar al de las bases de datos relacionales (Berry y Linoff, 2004). Este esquema puede ser alguno de los usados para bases de datos multidimensionales como son: los esquemas de estrella, copo de nieve o de constelaciones.

4. MINERÍA DE DATOS

4.1 Definición de Minería de Datos

La ciencia hace uso del método clásico, hipótesis y deducción, para la generación de nuevos conocimientos; esto se traduce en que a partir de un conjunto de observaciones y de conocimientos previos, la intuición del investigador lo conduce a formular una hipótesis. Esta intuición puede no ser de mucha utilidad cuando se trata millones de datos almacenados en algún medio electrónico. Frawley *et al.* (1992) remarca que las técnicas de análisis estadístico permiten obtener información útil, pero no inducir relaciones cualitativas generales, o leyes, previamente desconocidas y que para poder llevar a cabo esto se requieren técnicas de análisis inteligente.

En capítulos anteriores, se mencionó que el avance tecnológico en los medios de almacenamientos de información ha permitido que las instituciones conserven en diferentes medios un registro detallado de sus operaciones, como resultado de esto también se incrementa de forma continua la diferencia existente entre la cantidad de datos disponibles y el conocimiento extraído de los mismos (Reinosa *et al.*, 2012). Hernández Orallo *et al.* (2004) asegura que a consecuencia de la necesidad anterior, se desarrollaron nuevos métodos matemáticos y softwares con la capacidad de realizar un análisis inteligente de los datos, dándose paso, de ésta manera, a la Minería de Datos (MD).

Algunas definiciones que se pueden encontrar en la literatura sobre MD hacen referencia a la definición que se dio en Frawley *et al.* (1992) "La minería de datos puede definirse como la extracción no trivial de información implícita, previamente desconocida y potencialmente útil, a partir de los datos".

El paso de tiempo ha ido agregando conceptos a la definición anterior, relacionándola con los medios de almacenamiento de información, como la de Witten y Frank (2000, en Clark y Boswell, 2000) quienes definen a la minería de datos como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, a partir de grandes cantidades de datos almacenados en distintos formatos.

En la actualidad, la información de las corporaciones se guardan mayormente en grandes bases de datos por la cual la definición de MD también ha acuñado éste concepto, como por ejemplo en Zhu (2009) se menciona que la minería de datos puede ser conocida como la extracción de información implícita, previamente desconocida y potencialmente útil a partir de grandes bases de datos.

Al considerar las definiciones anteriores, la minería de datos debe superar la problemática asociada a trabajar con grandes volúmenes de datos como son: datos perdidos, datos atípicos, volatilidad de los datos; además de tener que utilizar las técnicas adecuadas para analizarlos.

La integración de las técnicas de MD en las actividades del día a día se está convirtiendo en algo habitual. El éxito de la minería de datos ha permitido aplicarla en investigaciones en ciencias relacionadas con el manejo de grandes volúmenes de información como la astronomía, la medicina, la biología, la genética y la bioingeniería, entre otras (Larose, 2005 y Olson y Delen, 2008). Como ejemplos de aplicación se mencionan, la clasificación de cuerpos celestes, el análisis de secuencias de proteínas y el análisis de secuencias de genes (Fayyad *et al.*, 1996). A través de la minería de datos se mezclan diferentes disciplinas como la estadística, los sistemas de información, las bases de datos y la inteligencia artificial entre otras para agilizar los procesos de análisis de información; pero (Berry y Linoff, 2004; Witten y Frank, 2005 y Krzysztof *et al.*, 2007) dicen que los negocios, el sistema bancario y la publicidad han sido las áreas en las que más se ha empleado.

En la actualidad la minería de datos ha tomado importancia en investigaciones agronómicas, como ejemplo podemos citar a Untaru *et al*., (2012) donde se recopila información de diversas investigaciones que relaciona las técnicas de minería datos con los agronegocios. Otra aplicación es propuesta por Savin *et al*. (2007) que construyeron una red neuronal que predijo el rendimiento de los cultivos de trigo de invierno en algunas regiones del sur de Rusia. También hay otras investigaciones como Kumar (2011) y Stastny *et al*. (2011) donde proponen otros modelos de redes neuronales para predecir rendimiento de cultivos agrícolas.

4.2 Tareas de la minería de datos

Berry y Linoff (2004) consideran que muchos de los problemas que tienen que ver con la economía y los negocios pueden ser expresados en seis tipos de tareas: clasificación, estimación, predicción, agrupamiento, segmentación y descripción por perfiles; mientras que Hernández Orallo *et al.* (2004) y Olson y Delen (2008) agrupan las distintas tareas de la minería de datos en predictivas o descriptivas, que a su vez se dividen en tareas más específicas, por ejemplo, en las predictivas se considera a la clasificación y la regresión, mientras que en las descriptivas se considera el Clustering y las reglas de asociación. En las siguientes subsecciones se proporciona un resumen de las tareas mencionadas.

- **Clasificación o discriminación:** La clasificación o discriminación es considerada como una tarea predictiva que de acuerdo con Pajares y De la Cruz (2010) pretende usar la información aportada por un conjunto de datos dados para generalizarlo a nuevas muestras. Para esta tarea las variables de entradas se denominan patrones, casos, entradas, instancias u observaciones y las variables de salida se denominan etiquetas, objetivos, salidas y a veces observaciones. Hernández Orallo *et al.* (2004) dicen que "El algoritmo utilizado en ésta tarea maximiza la razón de precisión de la clasificación de las nuevas instancias, la cual se calcula como el cociente entre las predicciones correctas y el número total de predicciones, correctas e incorrectas". Jiawei Han y Micheline Kamber (2006) consideran que los modelos resultantes de esta tarea pueden ser representados de varias formas, como reglas de clasificación (Si-entonces), arboles de decisión, fórmulas matemáticas o redes neuronales.

- **Regresión:** La regresión es una tarea predictiva que de acuerdo con Martín y Rosario de Paz Santana (2007) tiene que ver con el análisis simultáneo de dos o más variables numéricas que permite determinar el grado de dependencia entre ellas; es decir, se utiliza éste tipo de tareas para poner a prueba la relación de causalidad entre las variables. El algoritmo utilizado trata de minimizar el error cuadrado medio entre los valores predichos y los reales.

- **Análisis Clúster (Clustering):** El análisis clúster, también conocido con el nombre de segmentación, es una tarea descriptiva que de acuerdo con Macías Rodríguez (2008) identifica grupos de datos que son "similares", usando funciones de distancia especificadas por los usuarios o por expertos. Witten y Frank (2005) dicen que algunos algoritmos de clustering permiten a una instancia pertenecer a uno a más clústeres, mientras que, otros algoritmos asocian instancias a clústeres de manera probabilística, de manera que existe una probabilidad asociada o un "grado de pertenencia" asignado a un determinado clúster.

- **Reglas de asociación:** Las reglas de asociación son tareas descriptivas, que de acuerdo con Frawley *et al.* (1992) están basadas en un antecedente o precondición y un consecuente o conclusión por ejemplo "Si el atributo X toma el valor *x* entonces el atributo Y toma el valor *y*, $X \Rightarrow Y$ "; ésta notación no implican una relación causa-efecto, es decir, puede no existir una causa para que los datos estén asociados.

4.3 Técnicas de minería de datos.

La estadística fue la primera ciencia en considerar a los datos como su materia prima, pero las nuevas necesidades y, en particular, las nuevas características de los datos (volumen y tipología) hacen que las disciplinas que integran la minería de datos sean numerosas y heterogéneas. En Wang (2009) se realizó una recopilación de trabajos y algoritmos considerados como técnicas de minería de datos, que fueron clasificados en 10 grupos:

1. **Modelación estadística paramétrica**: modelos de regresión, modelos de regresión sobre componentes no correlacionados, modelos de regresión con variables categóricas, modelos lineales generalizados (regresión logística el más común), análisis discriminante y series de tiempo.
2. **Modelación estadística no paramétrica**: regresión no paramétrica y discriminación no paramétrica.
3. **Reglas de asociación y dependencia**: Reglas de asociación, reglas de dependencia, reglas de asociación multinivel y reglas de asociación secuenciales.

4. **Métodos bayesianos**: Teorema de Bayes e hipótesis MAP, Naïve Bayes, Redes bayesianas y clasificadores basados en redes bayesianas.
5. **Árboles de decisión y sistemas de reglas**: árboles de decisión para clasificación, sistema de aprendizaje de reglas por cobertura, poda y reestructuración, árboles de decisión para regresión y agrupamiento o estimación de probabilidades.
6. **Métodos relacionales y estructurales**: programación lógica y base de datos, y programación lógica inductiva.
7. **Redes neuronales artificiales**: redes neuronales con aprendizaje supervisado y redes neuronales con aprendizaje no supervisado.
8. **Máquinas de vectores soporte**: máquinas de vectores soportes para clasificación binaria.
9. **Extracción de conocimientos con algoritmos evolutivos y reglas difusas**: Computación evolutiva y lógica difusa.
10. **Métodos basados en casos y en vecindad**: técnicas para agrupamiento (mapas auto-organizativos de Kohonen, k medias, agrupamiento jerárquico), técnicas para clasificación (estimación bayesiana de funciones de densidad, K vecinos más cercanos, redes de cuantificación vectorial), métodos de vecindad con técnicas evolutivas (clasificación por vecindad mediante algoritmos genéticos, algoritmos evolutivos de estimación de distribuciones, aprendizaje incremental basado en poblaciones, algoritmo genético compacto), y razonamiento basado en casos.

La considerable cantidad de algoritmos desarrollados para realizar minería de datos, obligan a que en la presente investigación sólo se haga referencia a las técnicas presentes en la herramienta de análisis de SQL Server 2008, las cuales son:

- Árboles de decisión
- Regresión lineal
- Regresión logística
- Naïves Bayes
- Reglas de asociación
- Clustering o segmentación

- Redes neuronales artificiales

En las siguientes secciones se aplica con mayor detalle cada una de las técnicas anteriores.

Árboles de decisión: Los árboles de decisión constituyen un método de aprendizaje y clasificación muy utilizado, debido a la facilidad de organización y comprensión del conocimiento que proponen Larose (2005) y Witten y Frank (2005) menciona que éste método de aprendizaje ha sido utilizado satisfactoriamente en un amplio conjunto de aplicaciones: desde sistemas de diagnósticos médicos hasta sistemas de asesoramiento sobre el riesgo de concesión de créditos para préstamos.

Berry y Linoff (2004), MacLennan *et al.* (2009) y Krzysztof *et al.* (2007), asocian un árbol de decisión a un conjunto de restricciones o condiciones que se organizan de forma jerárquica, y que se aplica sucesivamente desde una raíz hasta llegar a un nodo terminal u hoja del árbol. Los árboles de decisión con conocimiento adquirido se pueden representar de maneras alternativas como un conjunto de reglas SI-ENTONCES para mejorar su compresión por parte del humano.

Pajares y De la Cruz (2010) dicen que los ejemplos a partir de los cuales se desarrolla la regla de clasificación se conoce a partir de los valores de un conjunto de propiedades o atributos, y por tanto, los árboles de decisión se expresan en función de ellos; los ejemplos pueden subministrarse a partir de una base de datos que contenga históricos de observaciones.

Regresión lineal: Pajares y De la Cruz (2010), asocian la tarea de regresión con los problemas de clasificación que determina una función con recorrido en los números reales. Martín y Rosario de Paz Santana (2007), amplían la definición considerando que el análisis de regresión tiene que ver con el análisis simultáneo de dos o más variables, y que principalmente se intenta describir la dependencia de una variable Y (llamada variable dependiente o de respuesta, que es una variable cuantitativa) en relación a una variable independiente X (también llamada variable predictora o explicativa). Por medio de este análisis se explica los cambios en la variable dependiente en relación a una combinación lineal de la variable independiente. Uno de los supuestos primordiales de este análisis es la existencia de una relación de causalidad entre las variables analizadas. El modelo resultante del análisis de regresión permite predecir el valor de

la variable dependiente estimando qué se obtendría para un valor de la variable independiente que no se encuentre en el conjunto de datos.

Regresión logística: De la Fuente Fernández (2011) describe la regresión logísticas como parte de un conjunto de métodos estadísticos y la define como "la variante que corresponde al caso en que se valora la contribución de diferentes factores en la ocurrencia de un evento simple". La regresión logística es útil cuando se trata de predecir el valor de una variable respuesta politómica (que admite varias categorías de respuesta, como por ejemplo excelente, muy bien, bien, regular, malo), pero es de especial utilidad cuando la variable es dicotómica (una respuesta del tipo binario 0 o 1, ausente o presente, sano o enfermo, etc.). Al igual que la regresión lineal depende de otras variables (variables explicativas).

Clasificador Naïve Bayes: Witten y Frank (2005) dicen que éste algoritmo tiene sus bases en las Reglas de Bayes y su fundamento principal es la suposición de que todos los atributos son independientes. Las reglas de Bayes, sobre probabilidad condicional, dicen que si se tiene una hipótesis H y la evidencia E, entonces:

$$\Pr[H\,!\,E] = \frac{P(E!\,H)Pr(H)}{\Pr(E)}$$

Donde $\Pr(E)$ denota la probabilidad del evento E y $P(E!\,H)$ denota la probabilidad de E condicionado a la ocurrencia de un evento H.

Duda y Hart (1973, en Hernández Orallo *et al.* 2004) mencionan que pesar de asumir la suposición de independencia de atributos, este clasificador es sin duda bastante fuerte y poco realista en la mayoría de los casos.

Reglas de asociación: Hernández Orallo *et al.* (2004) explican que en el ámbito de la minería de datos, una regla de asociación es una proposición probabilística sobre la ocurrencia de ciertos eventos en una base de datos, que puede ser representada como reglas de la forma SI α ENTONCES β ($\alpha \Rightarrow \beta$) donde α y β son dos conjuntos de atributos disjuntos (el conjunto α es conocido como el predecesor y β como el sucesor o consecuente). En las reglas de

asociación, a diferencia de las reglas de clasificación, en la parte derecha de la ecuación puede aparecer uno o más atributos.

MacLennan *et al.* (2009) dicen que éste algoritmo expresa patrones de comportamiento entre los datos, en función de la aparición conjunta de valores de dos o más atributos, es decir, que expresan las combinaciones de los valores de los atributos que suceden más frecuentemente.

Análisis Clúster (Clustering): Jain y Dubes (1988, en Fayyad *et al.*, 1996) se refieren a ésta técnica como una tarea descriptiva para identificar un conjunto finito de categorías para describir los datos. Jiawei Han y Micheline Kamber (2006) mencionan que ésta técnica usa el principio de maximizar la similitud entre los elementos de un grupo minimizando la similitud entre los distintos grupos. Es decir, se forman grupos tales que los objetos de un mismo grupo son muy similares entre sí y, al mismo tiempo, son muy diferentes a los objetos de otros grupos. Al agrupamiento también se le suele llamar segmentación, ya que parte o segmenta los datos en grupos que pueden ser o no disjuntos. El agrupamiento está muy relacionado con la agregación, que algunos autores consideran una tarea en sí misma, en la que cada grupo formado se considera como un resumen de los elementos que lo forman para así describir de una manera concisa los datos.

Redes Neuronales Artificiales: Hilera y Martínez (1995) definen a las redes neuronales artificiales (RNA) como programas de aprendizaje y procesamiento automático inspirados en la forma en que funciona el sistema nervioso central. Éstas están integradas por elementos simples de procesamiento llamados nodos o neuronas, organizadas en capaz o niveles. Cada nivel es un conjunto de neuronas cuyas entradas de información provienen de la misma fuente (que puede ser otra capa de neuronas) y cuyas salidas de información se dirigen hacia el mismo destino (que puede ser otra capa de neuronas). En este sentido se distinguen tres tipos de capas: la capa de entrada recibe la información del exterior, la o las capas ocultas son aquellas cuyas entradas y salidas se encuentran dentro del sistema y, por tanto, no tienen contacto con el exterior; por último, la capa de salida envía la respuesta de la red al exterior. Cada neurona está conectada con otra neurona mediante enlaces de comunicación, cada uno de los cuales tiene asociado un peso. Los pesos representan la información que será usada por la red neuronal para resolver un

problema determinado. Palmer y Montaño (1999) dicen que las RNA son sistemas adaptativos que aprenden de la experiencia, esto es, aprenden a llevar a cabo tareas mediante un entrenamiento o etapa de aprendizaje con ejemplos ilustrativos. El tipo de representación de la información que manejan las RNA tanto en los pesos de las conexiones como en las entradas y salidas de la información es numérica.

Yelitza y Talavera (2007) señalan que un dato de entrada puede consistir en un valor real continuo como la edad de una persona o puede ser un valor numérico discreto o binario como el sexo de una persona codificada de forma binaria, por ejemplo, mediante: hombre=1, mujer=0.

4.4 El proceso de extraer conocimientos de bases de datos

La minería de datos presenta una gran ventaja de análisis de información en comparación con los métodos tradicionales, pero también tiene la desventaja de que la calidad de los nuevos conocimientos generados mediante su utilización solo va a depender de la calidad de los datos con que se trabaje, por lo tanto es necesario la unificación de los datos en un formato acorde al análisis que se pretenda realizar (Krzysztof *et al.,* 2007). Antes de poder realizar minería de datos es necesario contar con información confiable y adecuada para los modelos que se pretendan utilizar, de esta manera se abre paso a un proceso iterativo que considera a la minería de datos como una de sus fases; este proceso es conocido como descubrimiento de conocimiento en bases de datos, (Knowledge Discovery in Data bases, KDD), que muchas veces se hace referencia a éste con el nombre de minería de datos, siendo ésta última sólo una fase de este proceso (Fayyad *et al.*, 1996). En McLennan (2009) mencionan que la minería de datos también se ha considerado haciendo referencia a otros términos como máquinas de aprendizaje y analítica predictiva.

El proceso KDD en sus primeras fases intenta superar los retos que comúnmente se presentan con los datos de análisis. Hernández Orallo *et al.* (2004) hacen una importante contribución al tema de la extracción de conocimientos en bases de datos al dividirlo en cinco fases que a continuación se enlistan:

1. Integración y recopilación de datos.
2. Selección, limpieza y transformación
3. Minería de datos.
4. Evaluación e interpretación.
5. Difusión y uso.

Para llevar a cabo estas tareas se recurre a las técnicas de cómputo y de software especializados que han evolucionado a la par de las exigencias de procesamiento.

Las primeras cuatro fases se contemplan dentro de la presente investigación y la aplicación de la quinta dependerá del impacto que tenga la nueva metodología en el análisis de datos forestales. En las siguientes secciones se explicará más a detalle en que consiste cada una de las etapas.

4.4.1 Fase de integración y recopilación

Las bases de datos tienen las cualidades suficientes para llevar a cabo las operaciones diarias de una organización como puede ser compra, venta, facturación, etc. (Date, 2001). Hernández Orallo *et al.,* (2004) explican que al momento de aplicar funciones más complejas como el análisis, la planificación y la predicción, para la toma de decisiones, a veces es necesario el uso de la información generada por otras organizaciones e incluso recurrir a la información en bases de datos públicas lo que "representa un reto debido a que cada fuente de datos están estructurada de diferentes formas, es decir, diferentes formatos de registro, diferentes grados de agregación de los datos, diferentes claves primarias, etc."

La integración y recopilación involucra una combinación de técnicas y tecnologías con las cuales se pretende eliminar las inconsistencias y redundancias de los datos que provienen de diferentes fuentes (datos homogéneos) para integrarlos a una base de datos central.

4.4.2 Fase de selección limpieza y transformación

El siguiente paso del proceso KDD, posterior a la integración y recopilación, es la selección y preparación de los datos que se van a minar. Fayyad *et al.* (1996) opinan que la limpieza de los datos es necesaria debido a que no toda la información recopilada es objeto de estudio, por lo cual son considerados como irrelevantes. Hernández Orallo *et al.* (2004) consideran que existen datos que no se ajustan al comportamiento general de los datos, conocidos como datos atípicos, y que pueden influir en el modelo de minería; pero que también se puede dar el caso de que estos datos sean importantes para el proceso, considerando el objetivo del análisis, por ejemplo en la detección de fraudes con tarjetas de crédito. Otro ejemplo puede ser la presencia de datos perdidos que también pueden afectar la confiabilidad de los modelos; o que alguno de estos necesite de un tipo de dato particular. La transformación de los datos, de acuerdo con MacLennan *et al.* (2009) se relaciona con la definición de las variables de entrada de los modelos de minería, algunos de ellos sólo pueden ser procesados usando datos continuos y otros usando datos discretos. Las tareas de transformación más utilizadas en la tarea de transformación son: transformaciones numéricas, agrupaciones y agregaciones.

4.4.3 Fase de Minería de datos.

Fayyad *et al.* (1996) y Berry y Linoff (2004) dicen que ésta fase es la más característica del proceso KDD y lo más sobresaliente es la exploración y el análisis de grandes volúmenes de datos con la finalidad de generar uno o varios modelos que puedan ser usados por los usuarios finales, es decir que sean de alguna manera útiles.

4.4.4 Fase de evaluación e interpretación

Hernández Orallo *et al.* (2004) menciona que la en esta fase se llevan a cabo las tarea de entrenamiento y prueba de los modelo de minería, mediante la creación de dos conjuntos de datos, generados a partir del conjunto original, conocidos como datos de entrenamiento y datos

de prueba. Esta separación se lleva a cabo para garantizar que la validación de la precisión del modelo sea una medida independiente, de otra manera, la precisión del modelo estaría sobrestimada. Como aporte adicional, también menciona que la tarea de minería condiciona la medida de evaluación de los modelos, de acuerdo con el resumen siguiente.

- En las tareas de clasificación lo normal es evaluar la calidad de los patrones encontrados con respecto a su precisión predictiva, la cual se calcula como el número de instancias del conjunto de prueba clasificadas correctamente, dividido por el número de instancias totales de ese mismo conjunto.
- Para las tareas de reglas de asociación se evalúan de forma separadas cada una de las reglas con el objeto de restringirse a aquellas que pueden aplicarse a un mayor número de instancias y que tienen una precisión relativamente alta sobre estas instancias. Esto se hace en base a dos conceptos, la cobertura y la confianza. La cobertura o soporte, se define como el número de instancias que la regla predice correctamente y la confianza o precisión que mide el porcentaje de veces que la regla se cumple cuando se puede aplicar, es decir, la cobertura dividida por el número de instancias a las que se puede aplicar la regla.
- En las tareas de regresión se utiliza el error cuadrático medio del valor predicho con respecto al valor que se utiliza como validación.
- Para las tareas de agrupamiento, las medidas de evaluación suelen depender del método utilizado, aunque suelen ser función de la cohesión de cada grupo y de la separación de los grupos. La cohesión y separación de los grupos se puede formalizar, por ejemplo, utilizando la distancia media al centro del grupo de los miembros de un grupo y la distancia media entre grupos, respectivamente. El concepto de distancia y de densidad son dos aspectos cruciales tanto en la construcción de modelos de agrupamiento como en su evaluación. Existen otras medidas subjetivas como el interés, la novedad, la simplicidad o la comprensibilidad que pueden ser revisadas con mayor detalle en la referencia dada.

4.4.5 Fase de difusión, uso y monitorización

Fayyad *et al.* (1996) dicen que en esta fase se usa el conocimiento adquirido incorporándolos a otros sistemas para llevar a cabo un conjunto de acciones o simplemente documentarlo o reportarlo a las partes interesadas. Hernández Orallo *et al.* (2004) apoyan la idea anterior agregando que "una vez construido y validado el modelo puede ser usado principalmente con dos finalidades, para que un analista recomiende acciones basándose en el modelo y en sus resultados, o bien para aplicar el modelo a diferentes conjunto de datos". En el caso de una aplicación manual o automática del modelo, recomienda realizar su difusión, y monitorear la evolución del modelo, para detectar posibles cambios en los patrones.

4.5 Softwares para minería de datos

El método tradicional de convertir los datos en conocimientos consiste en un análisis e interpretación realizada de forma manual; esta forma de actuar es lenta, cara, y altamente subjetiva. El análisis manual es impracticable en dominios donde el volumen de los datos crece exponencialmente: la enorme abundancia de los datos desborda la capacidad humana de comprenderlos sin ayuda de herramientas potentes (Hernández Orallo *et al.* 2004). Algunos paquetes de software, ampliamente utilizados actualmente, para el desarrollo de modelos de minería de datos sobre las cuales se puede encontrar mucha información en diversas páginas web y artículos electrónicos en internet se muestran en el Cuadro 3.

Cuadro 3. Softwares más utilizados para minería de datos

Software Gratuitos		**Software Comerciales**	
• KEEL • OpenNN • R • RapidMiner • Weka	• JHepWork • KNIME • Orange	• Oracle Darwin • SAS Enterprise Miner • SPSS Clementine • SQL Server Analysis Services • STATISTICA Data Miner	• dVelox • KXEN • Powerhouse • Quiterian • Neural Designer

Fuente: Elaboración propia

Cuadro 4. Análisis de características de las principales herramientas de minería de datos

PRODUCTO	Redes neuronales	Arboles de decisión	Criterio de Bayes	Empleo de k_medias	Técnicas estadísticas	Predicción	Series de tiempo	Agrupación	Asociación	Comp. Windows	Comp. Unix	Escalabilidad paralela	Extensiones SQL
Knowl. Seeker		X				X				X	X		
Knowl. Studio	X	X		X		X		X		X	X	X	X
BusinesMiner		X								X			
4Thought	X					X	X			X			
Scenario		X								X			
Marksman	X					X		X		X		X	
Red Brick			X			X				X	X		X
Intelligent Miner	X	X				X	X	X	X	X	X	X	
Dec. Series	X	X	X			X		X	X		X	X	
Neural SIM	X					X				X			
Darwin	X	X				X					X	X	
CART		X				X				X	X		
Enterprise Miner	X	X			X	X	X	X	X	X	X		
Answer tree		X			X	X				X	X		
Clementine	X	X				X	X	X	X	X	X		
Neural Connection	X				X	X				X	X		
Pattern recog. Workbench	X			X	X	X	X	X		X			

Fuente: Modificado de Sánchez Cañizares *e. al.* (2005)

Sánchez Cañizares *et al.* (2005) presentaron una evaluación de las principales características de 17 aplicaciones relacionadas con la minería de datos. Las características que se evaluaron se relacionaron con el empleo o no de ciertas herramientas o lenguajes estadísticos y su compatibilidad con ciertas plataformas informáticas; esencialmente se analizó el empleo de algunos algoritmos de minería de datos como: redes neuronales, árboles de decisión, k-medias, uso del criterio de Bayes, técnicas estadísticas tradicionales como la obtención de los principales estadísticos descriptivos, realización de predicciones, el uso de series de tiempo, la formulación de agrupaciones, la detección de asociaciones, la compatibilidad con Windows 95/98/NT y UNIX, la escalabilidad paralela y el uso de extensiones SQL; los resultados obtenidos en ésta evaluación se muestran en el Cuadro 4. Importante es mencionar que, a la fecha, los softwares de minería de datos han evolucionado de tal manera que muchas de las características que no presentaban durante esta evaluación ya han sido incorporadas dentro de las herramientas de análisis actuales.

4.6 Resumen

En este capítulo se introdujo el concepto de minería de datos, el proceso por el cual se genera la minería de datos y se describieron algunos de los algoritmos de minería datos más utilizados en la actualidad.

Algunas definiciones que se pueden encontrar en la literatura sobre MD hacen referencia a la definición que se dio en Frawley *et al.* (1992) “La minería de datos puede definirse como la extracción no trivial de información implícita, previamente desconocida y potencialmente útil, a partir de los datos”.

El paso de tiempo ha ido agregando conceptos a la definición anterior, relacionándola con los medios de almacenamiento de información, como la de Witten y Frank (2000, en Clark y Boswell, 2000) quienes definen a la minería de datos como el proceso de extraer conocimiento

útil y comprensible, previamente desconocido, a partir de grandes cantidades de datos almacenados en distintos formatos.

En la actualidad la información de las corporaciones se guardan mayormente en grandes bases de datos por la cual la definición de MD también ha acuñado éste concepto, como por ejemplo en Zhu (2009) se menciona que la minería de datos puede ser conocida como la extracción de información implícita, previamente desconocida y potencialmente útil a partir de grandes bases de datos.

La minería de datos forma parte de un proceso conocido como extracción de conocimiento a partir de bases de datos, o proceso KDD, que en sus primeras fases intenta superar los retos que comúnmente se presentan con los datos de análisis. Hernández Orallo *et al.* (2004), hace una importante contribución al tema de la extracción de conocimientos en bases de datos al dividirlo en cinco fases:

1. Integración y recopilación de datos.
2. Selección, limpieza y transformación
3. Minería de datos.
4. Evaluación e interpretación.
5. Difusión y uso.

La estadística fue la primera ciencia en considerar a los datos como su materia prima, pero las nuevas necesidades y, en particular, las nuevas características de los datos (volumen y tipología) hacen que las disciplinas que integran la minería de datos sean numerosas y heterogéneas. En Wang (2009) se realizó una recopilación de trabajos y algoritmos considerados como técnicas de minería de datos, que fueron clasificados en 10 grupos:

1. **Modelación estadística paramétrica**: modelos de regresión, modelos de regresión sobre componentes no correlacionados, modelos de regresión con variables categóricas, modelos lineales generalizados (regresión logística el más común), análisis discriminante y series de tiempo.

2. **Modelación estadística no paramétrica**: regresión no paramétrica y discriminación no paramétrica.
3. **Reglas de asociación y dependencia**: Reglas de asociación, reglas de dependencia, reglas de asociación multinivel y reglas de asociación secuenciales.
4. **Métodos bayesianos**: Teorema de Bayes e hipótesis MAP, Naïve Bayes, Redes bayesianas y clasificadores basados en redes bayesianas.
5. **Árboles de decisión y sistemas de reglas**: árboles de decisión para clasificación, sistema de aprendizaje de reglas por cobertura, poda y reestructuración, árboles de decisión para regresión y agrupamiento o estimación de probabilidades.
6. **Métodos relacionales y estructurales**: programación lógica y base de datos, y programación lógica inductiva.
7. **Redes neuronales artificiales**: redes neuronales con aprendizaje supervisado y redes neuronales con aprendizaje no supervisado.
8. **Máquinas de vectores soporte**: máquinas de vectores soportes para clasificación binaria.
9. **Extracción de conocimientos con algoritmos evolutivos y reglas difusas**: Computación evolutiva y lógica difusa.
10. **Métodos basados en casos y en vecindad**: técnicas para agrupamiento (mapas auto-organizativos de Kohonen, k medias, agrupamiento jerárquico), técnicas para clasificación (estimación bayesiana de funciones de densidad, K vecinos más cercanos, redes de cuantificación vectorial), métodos de vecindad con técnicas evolutivas (clasificación por vecindad mediante algoritmos genéticos, algoritmos evolutivos de estimación de distribuciones, aprendizaje incremental basado en poblaciones, algoritmo genético compacto), y razonamiento basado en casos.

5. ADMINISTRACIÓN DE DATA WAREHOUSE Y MINERÍA DE DATOS CON SQL SERVER 2008

En el presente capítulo se introducen algunas de las características y herramientas más importantes del Sistema Gestor de Bases de Datos (SGBD) SQL Server 2008. Se hace una breve revisión histórica de éste importante SGBD, se revisan los dos entornos que conforman la versión Enterprise y se describen sus herramientas de integración, análisis y reporte de datos.

La tecnologías para la gestión de base de datos ha avanzado a la par de los requerimientos computacionales de tal modo que algunos sistemas gestores contemplan la administración de Data Warehouse y algoritmos de minería de datos; un ejemplo de éstos es Microsoft (MS) SQL Server. SQL Server apareció por primera vez en el mercado en 1993, bajo la versión SQL Server 4.2 y hasta la versión que se utiliza en la presente investigación, SQL Server 2008, transcurrieron 15 años y 7 versiones de este importante SGBD. En cada versión se incluyeron una serie de nuevas herramientas y características que mejoraron sustanciosamente la forma de acceder y analizar la información (Colledge, 2010). En el Cuadro 5 se muestra la cronología de las versiones de MS SQL Server.

Cuadro 5. Versiones de SQL Server

Versión	Año de lanzamiento	Nombre Código
SQL Server 4.2	1993	SQL Server par Windows NT
SQL Server 6.0	1995	SQL95
SQL Server 6.5	1996	Hydra
SQL Server 7	1998	Sphinx
SQL Server 2000	2000	Shiloh
SQL Server 2005	2005	Yukon
SQL Server 2008	2008	Katmai

Fuente: Modificado de Colledge (2010)

La versión Enterprise de MS SQL Server 2008 cuenta con una serie de herramientas que facilitan el acceso y análisis de la información. SQL Server Management Studio (SSMS) y Business Intelligence Development Studio (BIDS) son dos entornos integrados a SQL Server 2008 Enterprise con los objetivos de administrar los servicios y la creación de objetos comerciales, mediante soluciones y proyectos; éstos entornos, aunque están integrados en SQL Server, están

diseñados de forma independientes adecuado a la programación de aplicaciones empresariales que trabajen con SQL Server, por ejemplo programas de servicio de análisis (Analysis Services-SSAS), Servicio de integración (Integration Services-SSIS) y servicio de reportes (Reporting Services-SSRS). Con estas herramientas no es posible crear aplicaciones personalizadas o crear grandes proyectos de desarrollo pero si excelente paquetes para mover y procesar grandes cantidades de datos (http://msdn.microsoft.com/es-es/library/ms174170(v=sql.100).aspx).

Nielsen *et al.* (2009) considera a SSMS como un potente conjunto de herramientas dentro de una Shell de Visual Studio que permite a los desarrolladores o administradores de base de datos (DBA) crear proyectos de bases de datos y administrarlo mediante una interfaz y código Transact-SQL (T-SQL). Las principales características de SQL Server Management Studio se enuncian a continuación:

- La función principal de éste entorno es obtener acceso a todos los componentes de SQL Server, para configurarlos y administrarlos.
- Este entorno combina un amplio grupo de herramientas gráficas con un editor de texto enriquecido para ofrecer acceso a SQL Server a los programadores y administradores.
- Este entorno combina las funciones del Administrador corporativo y el Analizador de consultas.
- Mediante este entorno se pueden administrar SSAS, SSIS, SSRS y XQuery

Business Intelligence Development Studio es una extensión de Microsoft Visual Studio 2008 con una serie de proyecto adicionales enfocados al BI con SQL Server, en otras palabras, es el entorno de SQL Server utilizado para desarrollar soluciones empresariales que incluye los proyectos de SSAS, SSIS y SSRS. Cada tipo de proyecto proporciona las plantillas necesarias para crear los objetos utilizados en las soluciones de BI y proporciona los medios para trabajar con esos objetos como son: diseñadores, herramientas y asistentes. Por ejemplo, se puede optar por un proyecto de SSAS si el objetivo es crear una base de datos para la obtención de cubos, dimensiones o modelos de minería de datos (http://msdn.microsoft.com/es-es/library/ms174170(v=sql.100).aspx).

5.1 Servicio de Análisis (Analysis Services-SSAS)

El proyecto del servicio de análisis se incluye dentro de Business Intelligence Development Studio y es utilizado principalmente para el desarrollo de procesos OLAP y de minería de datos para aplicaciones de Inteligencia de Negocios. SSAS incluye plantillas para cubos, dimensiones, estructuras de minería de datos, orígenes de datos, vistas de orígenes de datos y roles, y proporciona las herramientas para trabajar con estos objetos. (http://msdn.microsoft.com/es-us/library/ms173709(v=sql.100).aspx). Nielsen *et al.* (2009) menciona que mediante esta herramienta se logra agregar billones de registros en segundos, tarea que le tomaría varios minutos a una base de datos relacional.

5.1.1 SSAS para procesos OLAP

Los cubos dimensionales creados con SSAS presentan una serie de operaciones básicas realizadas sobre agregaciones como como suma, cuenta, máximo o mínimo, éstas pueden definirse como sumas parciales o basarse en algún tipo de cuenta o personalizarse para cumplir con requisitos especiales. Un cubo multidimensional también puede contener miembros calculados, no asociados directamente con el origen de datos pero derivado de éste. Por ejemplo, puede definirse un miembro calculado, como la varianza en los precios de un producto u otras operaciones matemáticas. Adicionalmente, se pueden definir conjuntos de entidades de interés para el usuario; por ejemplo, los 10 clientes principales (por volumen de ventas) o los productos más importantes. Estos conjuntos pueden utilizarse con facilidad para restringir el ámbito de una consulta a un conjunto específico de entidades (http://technet.microsoft.com/es-es/library/ms174783(v=SQL.90).aspx)

SSAS integra las herramientas para realizar operaciones complejas. Ejemplos de cálculos complejos que se pueden realizar mediante cubos dimensionales creados con SSAS, son los siguientes:

- En un periodo de tres meses es posible mostrar la media móvil para cada período.

- Una comparación del crecimiento interanual de este período con el mismo período del año pasado.
- Las ventas que se muestran en la moneda base, se pueden volver a convertir a la moneda original utilizando la tasa de cambio media diaria en el momento de la venta.
- Las ventas presupuestadas se pueden calcular por categoría para el próximo año como un aumento del 10% sobre este año y asignar un presupuesto para cada producto según las ventas medias relativas de los últimos tres años.

El cubo multidimensional creado con SSAS se parece a una hoja de cálculo multidimensional, en la que el valor de una celda puede calcularse a partir de los valores de otras celdas o según el valor que suele haber en dicha celda (se admiten ecuaciones simultáneas); por ejemplo, los beneficios se derivan de los ingresos menos los gastos, pero las bonificaciones, que se incluyen en los gastos, se derivan de los beneficios (http://technet.microsoft.com/es-es/library/ms174783(v=SQL.90).aspx).

Veerman *et al.* (2009) menciona que SSAS también proporcionar el uso del lenguaje MDX (Expresiones multidimensionales), que se ha diseñado específicamente para crear todo tipo de cálculos en cubos multidimensionales que además permiten la integración con Microsoft .NET. Mediante esta integración se pueden escribir funciones y procedimientos almacenados en cualquier lenguaje .NET comprobable, como C#.NET o Visual Basic .NET. La función o el procedimiento almacenado pueden invocarse luego mediante MDX para usarlo en cálculos.

5.1.2 Minería de datos con SSAS

Los paquetes estadísticos actuales como el SAS, SSPS, STAT y R, utilizan estadística paramétrica y no paramétrica para inferir patrones a partir de un conjunto de datos. Hernández Orallo *et al.* (2004) explican que el problema que se presenta, con los paquetes estadísticos, es que no son sencillos de utilizar por cualquier usuario y muchos de ellos no son compatibles con las grandes bases de datos, que contienen cientos de tablas, millones de registros y algunos tipos

de datos como los atributos nominales que tienen muchos valores, datos textuales, multimedia, etc. Además de que no se integran bien a los sistemas de información.

Reinosa *et al.* (2012) afirman que una de las cualidades de la minería de datos es que se puede implementar en plataformas existentes, por software o hardware, y que cuentan con la capacidad de conectarse con nuevos productos y sistemas de recolección en línea; ésta cualidad es ampliamente utilizada por SQL Server para integrarse a casi cualquier sistema.

Veerman *et al.* (2009) y Nielsen *et al.* (2009) dicen que SSAS proporciona una amplia gama de herramientas que puede usar para generar soluciones de minería de datos, tanto con datos de cubo como con datos relacionales. La minería de datos mediante las herramientas de SSAS utiliza el análisis matemático para deducir los patrones y tendencias que existen en los datos (normalmente, estos patrones no se pueden detectar mediante la exploración tradicional de los datos porque las relaciones son demasiado complejas o porque hay demasiado datos). Estos patrones y tendencias se pueden recopilar y definir mediante un modelo de minería de datos que pueden ser aplicados a situaciones empresariales como: predecir ventas, dirigir correo a clientes específicos, determinar los productos que se pueden vender juntos, buscar secuencias en el orden en que los clientes agregan productos a una cesta de compra, etc.

SSAS incluye algoritmos y herramientas de minería de datos que facilitan la generación de una solución completa para diversos proyectos, los algoritmo implementados son árboles de decisión de Microsoft, clústeres de Microsoft, Bayes naïve de Microsoft, reglas de asociación de Microsoft, red neuronal de Microsoft, regresión logística de Microsoft y regresión lineal de Microsoft. A continuación, se brinda un pequeño bosquejo del funcionamiento de los algoritmos de minería implementados en SSAS, de acuerdo a la descripción dada por (MacLennan *et al.*, 2009).

- El algoritmo de árboles de decisión de Microsoft es un híbrido desarrollado por un grupo de investigación de Microsoft que soporta tareas de clasificación, regresión y asociación.
- El algoritmo de clústeres de Microsoft encuentra grupos naturales dentro del conjunto de datos cuando las agrupaciones no son obvias, principalmente cuando se manejan miles de datos; en otras palabras, lo que hace es encontrar la variable que clasifica con

mayor precisión a los datos, una vez encontrada la variable los datos son etiquetados con la varible descubierta.

- El algoritmo de Naïve Bayes de Microsoft es un método sistemático para el aprendizaje basado en evidencias que permite crear modelos predicitivos, de manera muy rápida, y genera un nuevo método para explorar y entender los datos con ayuda del visor integrado en la herramienta de SQL Server.
- El algoritmo de reglas de asociación de Mivrosoft está basado en lo que se conoce como "análisis de la cesta de comparas" que consiste en entender el comportamiento de compra de los clientes mediante una asociación de los productos que se suelen comprar al mismo tiempo.
- El complejo analisis realizado por la red neuronal de Microsoft se deriva de dos factores, el primero es que algunas o todas las entradas pueden estar relacionadas con algunas o todas las salidas y la red lo debe considerar en la etapa de entrenamiento (consecuentemente analiza todas las posibles relaciones). Segundo, las diferentes combinaciones de entradas pueden estar relacionadas directamente con las salidas. Las relaciones detectadas por este algoritmo considera un máximo de dos niveles. En el caso de usar sólo un nivel, las entradas o hechos son conectados directamente con las salidas. Para el caso en que se consideran dos niveles, la combinación de las entradas se convierten en nuevas entradas, las cuales son conectadas a las salidas, el nivel que se encarga de hacer las transformaciones se conoce como capa oculta ya que no es visible en los datos pero es usado por el algoritmo.
- Regresión logística de Microsoft es un caso particular de las redes neuronales (es una red neuronal con una capa o sólo un nivel de relaciones). Es necesario remarcan que el modelo de regresión logística tradicional no está basado en las redes neuronales pero el modelo matemático es identico al modelo matemático de la Red Neuronal de Microsoft sin capa oculta, por lo cual, consume menos recursos en la etapa de entrenamiento.
- El algortimo de regresión lineal de Microsoft es el mismo algoritmo utilizado para realizar un análisis de regresión lineal.

Los modelos de minería pueden implementarse en otro servidor para que los usuarios puedan realizar análisis y predicciones ad hoc mediante los modelos almacenados, se puede tener acceso a los modelos de minería de datos a través de clientes personalizados, incluyendo servicios web, o usando las aplicaciones de Microsoft Office, complementos de minería de datos para Excel (http://technet.microsoft.com/es-mx/library/ms174949(v=sql.100).aspx).

5.2 Servicio de Integración (Integration Services-SSIS)

SSIS se encuentra incluido dentro de BIDS y de acuerdo con Nielsen *et al.* (2009) comúnmente se hace referencia a éste como una herramienta de ETL, asociada a la preparación de datos para el almacenamiento en un Data Warehouse, análisis y reporte. Algunas de sus características principales son las siguientes:

- Sencillez y rapidez para mover grandes cantidades de datos, desde y hacia diferentes fuentes.
- La capacidad para enlazar tareas de manera simultánea, es decir, pueden ser ejecutadas en paralelo.
- Las conexiones para lectura y escritura, para la mayoría de los tipos de datos, son administrados sin alguna programación especial.
- Las tareas de administración de bases de datos y datos comunes son implementadas sin la necesidad de escribir código; adicionalmente algunas tareas permiten la programación en un ambiente .Net.
- Los paquetes resultantes pueden administrarse de acuerdo a como sean requeridas.

SSIS incluye las plantillas para paquetes, orígenes de datos y vistas de orígenes de datos, y proporciona las herramientas para trabajar con estos objetos (http://msdn.microsoft.com/es-us/library/ms174181(v=sql.100).aspx).

5.3 Servicio de Reportes (Reporting Services-SSRS)

BIDS incluye los proyectos, modelos de informe e Informes, ambos para desarrollar soluciones de informes. El tipo de proyecto Modelo de informe incluye las plantillas para modelos de informes, orígenes de datos y vistas de orígenes de datos, y proporciona las herramientas para trabajar con estos objetos. El proyecto Informe incluye las plantillas para trabajar con informes y orígenes de datos compartidos (http://msdn.microsoft.com/es-us/library/ms173745(v=sql.100).aspx).

5.4 Resumen

En éste capítulo se resume las características y herramientas más importantes del Sistema Gestor de Bases de Datos SQL Server 2008 y se revisan los dos entornos que conforman la versión Enterprise proporcionando una descripción de sus herramientas de integración, análisis y reporte de datos.

La versión Enterprise de MS SQL Server 2008 cuenta con una serie de herramientas que facilitan el acceso y análisis de la información. SQL Server Management Studio y Business Intelligence Development Studio son dos entornos integrados a SQL Server 2008 Enterprise con los objetivos de administrar los servicios y la creación de objetos comerciales, mediante soluciones y proyectos.

Nielsen *et al.* (2009) describe a SQL Server Management Studio como un potente conjunto de herramientas dentro de una Shell de Visual Studio que permite a los desarrolladores o administradores de base de datos (DBA) crear proyectos de bases de datos y administrarlo mediante una interfaz y código Transact-SQL (T-SQL).

Business Intelligence Development Studio es una extensión de Microsoft Visual Studio 2008 con una serie de proyectos adicionales enfocados a la inteligencia de negocios con SQL Server, en otras palabras, es el entorno de SQL Server utilizado para desarrollar soluciones

empresariales usando los proyectos: servicio de análisis, servicio de integración y servicio de reporte.

El proyecto de servicio de análisis es utilizado principalmente para el desarrollo de procesos OLAP y de minería de datos para aplicaciones de inteligencia de negocios (http://msdn.microsoft.com/es-us/library/ms173709(v=sql.100).aspx).

El servicio de integración comúnmente es conocido como una herramienta de extracción, transformación y carga, asociada a la preparación de datos para el almacenamiento en un Data Warehouse, análisis y reporte (Nielsen *et al.*, 2009).

El servicio de reportes incluye los proyectos de modelos de informes, para desarrollar soluciones tipo informes (http://msdn.microsoft.com/es-us/library/ms173745(v=sql.100).aspx).

III MARCO CONTEXTUAL

6. LOS INVENTARIOS FORESTALES Y LA IMPORTANCIA DE LA ESTIMACIÓN DE CARBONO.

En el presente capítulo se pretende destacar la importancia de los Inventarios Forestales (IF) mediante las actividades llevadas a cabo por diferentes países, incluyendo México, en cuanto a la evaluación y cuantificación de los recursos presente en bosques y selvas. En primer término se describe los hechos relevantes acontecidos a nivel mundial que dieron lugar a los primeros Inventarios Forestales Mundiales (IFM). Después se da una descripción cronológica de los inventarios forestales llevados a cabo en territorio nacional, las actividades y resultados obtenidos. Posteriormente, se detalla la importancia de calcular el volumen de madera, la biomasa y carbono. Finalmente, se muestran los principales métodos y ecuaciones utilizados para calcular estos indicadores.

6.1 Antecedentes de los Inventarios Forestales

Un bosquejo de la historia de los Inventarios Forestales Mundiales es narrado en Holmgren y Persson (2003), en el cual se describen los hechos más relevantes ocurridos durante un periodo de casi un siglo. En los siguientes párrafos se menciona cada evento relevante ocurrido desde el año 1910 hasta el año 2002.

Los Inventarios Forestales (IF) desde sus inicios y hasta la fecha, han evolucionado a la par de las necesidades mundiales de conocer y cuantificar los recursos forestales. En Zon (1910) se da conocer lo que es considerado como el primer intento de lo que ahora son los inventarios forestales; éste fue llevado a cabo con el objetivo de cuantificar la dimensión de la explotación de los recursos madereros en las principales naciones del mundo. Esta importante tarea se llevó a cabo mediante la compilación de información y finalmente la redacción de un informe sobre los recursos forestales mundiales; dicho trabajo estuvo a cargo del Servicio Forestal de los Estados Unidos. La meta principal era cuantificar y valorar los productos forestales de todos los países considerando aspectos tales como la propiedad, gestión y sostenibilidad. Esta publicación

fue catalogada por Holmgren y Persson (2003) como similar a la Evaluación de los recursos forestales mundiales emprendida por la FAO en el 2000, en cuanto a alcance se refiere.

La FAO, en 1945, durante una conferencia sentó las bases de lo que sería el primer inventario forestal mundial (IFM), el cual se realizó hasta 1947 y tuvo sólo una duración de un año. El siguiente inventario tuvo lugar en 1953 y a partir de allí se dieron dos inventarios con un espaciamiento en tiempo de cinco años, en 1958 y 1963, respectivamente.

A partir de 1960 los inventarios forestales para los países en desarrollo se convirtieron en una forma de asistencia por parte de la FAO. La asistencia se aplicaba una sola vez a los países en desarrollo y no se daba continuidad a los proyectos. Como resultado de ello pocos países en desarrollo tienen un buen conocimiento de sus recursos forestales. La preocupación de las naciones durante este periodo claramente no era de tipo ambiental como ocurre en la actualidad. Los IFM se realizaban con el único objetivo de evaluar la relación entre oferta y demanda, es decir, se cuantificaba la cantidad de madera explotable en cuanto a tamaño y especies comerciales. Estos inventarios no tenían una adecuada planeación ya que todos estaban basados en un mismo esquema, ya que recursos forestales en todos los países se describían de la misma forma que en Europa.

Evaluaciones regionales fueron llevadas a cabo durante los años setenta, pero ningún estudio mundial. En este periodo, se invirtió muchos recursos en una herramienta innovadora conocida como teledetección que amenazaba con sustituir el trabajo de campo. Holgrem y Person (2003) señalan que la teledetección tiene un gran potencial para ciertas mediciones de superficies, pero también resaltan que un mapa de la cubierta vegetal no es una evaluación forestal. Más adelante se verá que las técnicas actuales combinan el uso de mapas y técnicas de muestreo para llevar a cabo los inventarios forestales (Velasco *et al.*, 2003) y todavía más sofisticado es el actual uso de algunas técnicas de minería de datos para el reconocimiento de especies vegetales a partir de imágenes satelitales (Muñoz *et al.*, 2012).

La FAO incluyó por primera vez el tema de la deforestación en la evaluación de los recursos forestales mundiales (ERF) en 1980. A partir de aquí ya se empezaba a dar otro giro al tema de

los recursos forestales cambiando la visión más hacia el cuidado del medio ambiente. Para 1990 la deforestación fue el tema principal de la ERF, junto con algunos otros de ámbito ecológico-ambiental como es la diversidad biológica.

La deforestación siguió siendo un tema importante en el año 2000, pero el interés por los problemas de la conservación empezó a tener mayor auge y en el ERF de este año se incluyeron temas como biodiversidad, zonas protegidas e incendios forestales; además de incluir información del tipo productivo como oferta total de madera, plantaciones forestales, árboles fuera de los bosques y productos forestales no madereros.

FAO (2002) menciona la presentación de informe en donde se mostraba que de 137 países en desarrollo, sólo 22 habían repetido sus inventarios, 54 se basaban en un inventario único, 33 tenían un inventario forestal parcial y 28 no tenían inventario alguno. Pocos países en desarrollo tenían información actualizada sobre sus recursos forestales y todavía para estas fechas incluso los países industrializados no estaban exentos de este mismo problema. De acuerdo con Holmgren y Persson (2003) la falta de fuentes fidedignas de datos, repercute en la fiabilidad de las evaluaciones mundiales.

6.2 Los inventarios forestales en México

La Comisión Nacional Forestal (CONAFOR) tiene, entre muchas obligaciones, el informar a la población mexicana sobre la cuantía, ubicación y condiciones de los recursos forestales del país. Los primeros indicios de registro de información sobre la cantidad y ubicación de los recursos forestales en México se remontan a la época precortesiana, no con un fin ecológico-ambiental, sino con la intensión de cobrar un tributo acorde a las riquezas de los antiguos moradores del territorio INIFAP (1984). En 1976, por primera ocasión, se da a conocer información de los recursos forestales del país, obtenida a través de estudios con diferentes características y niveles de precisión. Dicha información fue ajustada en 1983 a partir de nuevos trabajos y datos adicionados, pero fue hasta la década de los noventa cuando nuevamente se pusieron en marcha actividades para cuantificar el recurso forestal del territorio nacional CONAFOR (2012).

Caballero (1998) y SEMARNAT (2002), hacen una revisión histórica de los inventarios forestales llevados a cabo en el territorio nacional, de donde se puede rescatar que a la fecha se han llevado a cabo cinco inventarios forestales de cobertura nacional:

- Primer Inventario Nacional Forestal 1961 – 1985
- Inventario Nacional Forestal de Gran Visión 1991
- Inventario Nacional Forestal Periódico 1992-1994
- Inventario Forestal Nacional 2000
- Inventario Nacional Forestal y de Suelos 2004-2009

La CONAFOR (2012) señala que sólo el Primer Inventario Nacional Forestal 1961-1985 y el Inventario Nacional Forestal y de Suelos 2004-2009, pueden ser considerados como inventarios completos. En los siguientes párrafos, se menciona los principales acontecimientos ocurridos en cada uno de los inventarios llevados a cabo hasta el momento.

Primer Inventario Nacional Forestal: La historia de los inventarios forestales nacionales comienza en 1961, cuando México presentó un proyecto ante la FAO para realizar el primer Inventario Nacional que daría inicio en 1962.En México se creó el Departamento de Fotogrametría e Inventario perteneciente al extinto Instituto Nacional de Investigaciones Forestales. La FAO jugó un papel muy importante al brindar soporte técnico becando a profesionales nacionales para capacitarse en Europa y estados Unidos y contratando expertos de talla internacional para que vinieran a asesorar las actividades, el apoyo de la FAO duró sólo dos años culminando en 1964. En esta etapa e establecieron las bases técnicas y administrativas del proyecto del Inventario Nacional Forestal; por primera vez se utilizaron herramientas nuevas como el cómputo electrónico y algunas técnicas estadísticas como el muestreo, además de fotogrametría y fotointerpretación (Caballero, 1998).

El inventario nacional se basó en el análisis de fotografías aéreas (escala 1:50 000) y en un muestreo intensivo en campo aplicado solamente a las principales zonas arboladas con valor comercial (aproximadamente el 52% de la superficie total). El resto de la superficie se cuantifico

mediante imágenes de satélite escala 1:3 000 000. Se llevó a cabo un muestreo sistemático en sitios circulares de aproximadamente un décimo de hectárea. (INIFAP-FAO, 1961-1964).

Holmgren y Persson (2003) señalaron que los inventarios de estas fechas estaban orientados hacia las áreas maderables, a lo cual México no fue la excepción. Los principales resultados que se obtuvieron son: el levantamiento de información dasonómica y de carácter ecológico silvícola, se generaron cartografía escala 1:50 000 y 1:100 000, tablas de volumen con corteza para pino, encino y para trozas de fuste limpio, al igual que documentos estadísticos dasométricos como la memoria nacional y estatal. (INIFAP-FAO, 1961-1964).

Inventario Forestal de gran visión: En 1991 la Subsecretaría Forestal y de Fauna Silvestre de la entonces Secretaría de Agricultura y Recurso Hidráulicos (SARH), llevó a cabo el Inventario Nacional Forestal de Gran Visión, cuyo objetivo principal consistía en realizar una actualización a bajo costo de la delimitación de los recursos forestales del país. La metodología utilizada se limitó a mediciones indirectas a partir de la información de campo del primer Inventario Nacional Forestal, respaldándose con la carta de uso de suelo y vegetación serie I, del Instituto Nacional de Estadística y Geografía (INEGI) y mediante la clasificación digital de imágenes de satélite NOAA-AVHRR, de baja resolución. Cabe mencionar que visión del proyecto fue servir de base para la realización del Inventario Forestal Periódico (SEMARNAT, 2005).

Los productos y resultados obtenidos fueron: mapas (escala 1:1 000 000) de la vegetación forestal del país, clasificada en 10 tipos, además de una memoria nacional con información general, SEMARNAT (2002).

Inventario Nacional Forestal Periódico: En el año 1992 se inició el proyecto denominado Inventario Nacional Forestal Periódico, cuyos resultados se dieron a conocer en 1994. Este inventario tuvo como objetivos actualizar y detallar la información sobre los recursos forestales, zonificar los terrenos forestales y preferentemente forestales de acuerdo a sus aptitudes y funciones y sentar las bases para actualizar la información en forma permanente (SEMARNAT, 2005).

La metodología utilizada se sustentó en la combinación de mapas de uso de suelo y Vegetación del INEGI, Serie I, además, del análisis de imágenes de satélite Landsat TM 5 y de un muestreo de campo de baja intensidad. La SEMARNAT (2005) considera que el inventario tiene características sin precedentes debido a que por primera vez se usaron imágenes de satélite de mediana resolución para elaborar mapas de todo el territorio nacional en escala 1:250 000; el levantamiento de información de campo se dio mediante parcelas de muestreo distribuidas sistemáticamente, se obtuvieron mapas en los cuales se zonificaron los terrenos forestales de acurdo a su aptitud y funciones y se logró el almacenamiento de toda la información en archivos magnéticos para su uso posterior en sistemas de información geográfica.

Como productos y resultado de este inventario se obtuvieron: mapas (escala 1:250 000) de la vegetación forestal del país dividida en seis categorías principales que fueron subdivididas dando un total de 40 categorías de vegetación o uso del suelo. También se obtuvo la zonificación de terrenos forestales en tres clases (conservación, producción y restauración). Finalmente, se elaboró una memoria nacional con las estadísticas dasométricas generadas. (SEMARNAT, 2002).

Inventario Nacional Forestal 2000: En el año 2000, la Secretaria de Medio Ambiente, Recursos Naturales y Pesca (SEMARNAP) comisionó a la Universidad Nacional Autónoma de México la realización de un nuevo inventario nacional forestal que fue truncado durante su primer año.

Como producto sólo se obtuvo una carta de vegetación y uso actual del suelo (escala 1:250 000), con ocho formaciones de vegetación, subdivididos en 17 tipos, 47 comunidades y 28 subcomunidades, para un total de 75 categorías. La carta fue elaborada mediante la interpretación visual de imágenes de satélites Landsat ETM 7y debía servir de base para la realización de un nuevo inventario nacional forestal, sin embargo, la información generada quedó organizada de manera distinta a la que se reportó en el inventario anterior y solamente se publicaron resultados parciales. No hubo muestreo en campo ni la cartografía fue validada en campo SEMARNAT (2002).

6.3 El Inventario Nacional Forestal y de Suelos 2004-2009

En CONAFOR (2012), se menciona que en el 2003 fue aprobada la Ley General de Desarrollo Forestal Sustentable (LGDFS), en el cual quedó establecido que la entidad normativa para el INFyS sería la Secretaría de Medio Ambiente y Recursos Naturales (SEMARNAT), la cual a su vez se autorizó a la CONAFOR, como el organismo encargado de recopilar e integrar la información derivada de dichos trabajos de inventario, acoplándose a las siguiente características:

- El inventario tendría una periodicidad de 5 años.
- La metodología sería homogéneas para permitir la aplicación de los inventarios a nivel nacional, regional y estatal; así como la comparación en el tiempo para la determinación de cambios y tendencias en los recursos forestales nacionales.

El seguimiento de la normatividad marcó el inicio del Inventario Nacional Forestal y de Suelos en el año 2004 y concluyó hasta el año 2009. Posterior al año 2009 dio inicio el segundo ciclo de inventario que consistía en la recopilación de información en sitios de remuestreo, marcado con esta finalidad siguiendo el objetivo de llevar a cabo evaluaciones de la dinámica de los bosques nacionales. El presente inventario nacional se elaboró con base en las cartas de Uso de Suelo y Vegetación que produce el INEGI, y el análisis de imágenes de satélite de mediana resolución (herramienta que empezaron a usarse en los años 70) CONAFOR (2012).

En el Cuadro 6 se proporciona información más detallada de las acciones realizadas durante el período 2004-2012, que abarca desde el inicio con la fase de muestreo hasta la impresión del Informe de Resultados del INFyS 2004-2009.

Cuadro 6. Descripción cronológica de las actividades realizadas para el INFyS 2004-2009

Año	Descripción de actividades realizadas para el INFyS
2004	• Dio inicio la primera parte del INFyS

2005	• Sin información relevante
2006	• Se llevó a cabo la entrega de información levantada en campo y se finiquitaron los contratos plurianuales firmados en 2004. Para finales de este año ya se tenía el 86% de los conglomerados muestreados correspondientes al primer ciclo de muestreo.
2007	• Se levantaron 3,772 conglomerados en campo concentrándose el trabajo en la cuenca de Lerma-Santiago-Pánuco. • En este año concluyó el primer ciclo de muestreo.
2008	• Se llevó a cabo la calibración de criterios y nuevas metodologías a operarse en el segundo ciclo de medición del INFyS, aplicándose de manera experimental la re medición en 2,517 conglomerados. • Se llevó a cabo la promoción en 5 entidades federativas, de la realización de los inventarios forestales y de suelos a nivel estatal.
2009	• Se realizaron las gestiones necesarias a fin de establecer los criterios y sentar las bases técnicas y administrativas para dar inicio con la segunda fase del Inventario Nacional Forestal y de Suelos. • Se inició la fase de remuestreo con el levantamiento de 4,780 conglomerados correspondientes al 24% del total de conglomerados que se programaron levantar para el segundo ciclo de inventario (2009-2014). • Se llevó a cabo la promoción, en 8 Entidades Federativas, para la implementación de los inventarios forestales estatales, además de dar continuidad a los trabajos iniciados en los 5 Estados contactados durante el 2008. • Se recopiló, analizó e integró la información que corresponde a las 17 tablas que constituyen la Evaluación de los Recursos Forestales, Informe México 2010.
2010	• Se incorporaron nuevas variables a los formatos de campo del INFyS sobre el tema de sanidad forestal, con lo que se buscó aportar información tabular y fotográfica al respecto de las plagas detectadas durante la toma de datos del Inventario. • Se incorporó el levantamiento de los pastizales naturales dentro del muestreo, anteriormente considerada en la categoría denominada "sin cubierta vegetal". • Se adecuó y complementó el manual de campo del remuestreo, de acuerdo con las nuevas variables incorporadas. • Se continuó con el remuestreo, de aproximadamente 4,655, el 23% del total programado para el sexenio. • Se realizó promoción en 7 Entidades Federativas para la implementación de los inventarios forestales estatales, comprendiendo el contacto y ejecución de talleres de trabajo con la finalidad de dar apoyo en la comunicación, planeación y diseño de Inventarios Forestales, además de dar continuidad a los trabajos iniciados en 6 Estados contactados durante el 2008 y en 7 contactados en 2009. • Se procesó la base de datos del primer ciclo del INFyS para llevar a cabo el análisis de la información y comenzar con la integración del informe de resultados del inventario 2004-2009.
2011	• Se implementó la revisión automatizada de algunos aspectos en la entrega de productos de los proveedores, con miras a la mejora en el control de calidad de la información del inventario, además de incluir la toma de fotografías hemisféricas en campo. • Se continuó con el remuestreo en 4,652 conglomerados (23% del total programado). • Se realizaron promociones en las entidades que no visitadas en los años anteriores y se dio seguimiento a los estados donde ya se había estado trabajando. • Se integró y analizó la información para generar el informe de resultados del INFyS 2004-2009. • Se realizó un diagnóstico de la base de datos del remuestreo 2009.

	• Se inició la migración del concentrado digital de los remuestreos 2009 y 2010 a la base de datos del servidor central denominado TULE.
2012	• Se llevó a cabo la integración de 9 variables nuevas y la sustitución de 3 en el cuadro de vegetación mayor en formato de bosques y selvas para evaluar salud del bosque. • Con la introducción de nuevas variables, se modificó el cliente de captura para la integración de la información del remuestreo 2012 a la base de datos y se continuó con la migración de la información al servidor central. • Se realizaron ajustes para el levantamiento de la información de campo: eliminación de los formatos de inaccesibilidad; revisión automática de nomenclatura fotográfica y eliminación de las colectas especiales botánicas y las aéreas de difícil acceso. • Se llevó a cabo el remuestreo de 4,362 conglomerados (22% del total programado). • Se realizó el diagnóstico de la base de datos del remuestreo 2010 para integrar el reporte correspondiente, además de llevar a cabo el análisis de datos de campo de los remuestreos 2009 y 2010. • Se generó la versión impresa y digital del informe de resultados del INFyS 2004-2009.

Fuente: Propia con información de Aldana (2012)

Los productos y resultados más importantes obtenidos durante este largo proceso de inventario fue la integración de una base de datos geoespacial, estructurada bajo un modelo relacional orientado hacia el análisis y la obtención de los principales indicadores que se establecen en la LGDFS. Además, se generó el informe de resultados donde se describen los principales indicadores dasonómicos que son utilizados en la presente investigación como punto de comparación y validación de resultados.

Aldana (2012) explica en su informe que en cuanto a las mejoras técnicas, se incluyeron más variables y se transformaron aquellas que lo ameritaban; para la fase de muestreo en campo se exigió a las empresas que prestaron sus servicios a la CONAFOR, la profesionalización del personal responsable del levantamiento de datos de campo y se actualizaron los manuales de campo. Como mejora administrativa resalta el ajuste que en hizo en los tiempos y las formas para los procesos de las licitaciones y la relación con los proveedores. Se implementó criterios y procedimientos de revisión más estrictos y se realizó mejoras tecnológicas para garantizar una mayor calidad de los datos que fueron integrados a la bases de datos del INFyS.

6.4 Importancia de la estimación de carbono

La investigación sobre la importancia de estimar la captura de carbono nos lleva hasta sus orígenes en la Convención Marco de la Naciones Unidas sobre el Cambio Climático

(CMNUCC) celebrado en Kioto, Japón, el 11 de diciembre de 1997 donde fue adoptado el protocolo de Kioto que consiste en un acuerdo internacional con el objetivo de reducir las emisiones de seis gases de efectos invernadero que causan el calentamiento global. En esta convención se estableció el compromiso obligatorio de cumplimiento, cuando los países industrializados responsables de, al menos, un 55 % de las emisiones de CO_2 lo ratificaran, Naciones Unidas (1998).

Los gases que están contemplado dentro de este documento son:

- Dióxido de carbono (CO_2)
- Gas metano(CH_4)
- Óxido nitroso (N_2O)
- Hidrofluorocarbonos (HFC)
- Perfluorocarbonos (PFC)
- Hexafluoruro de azufre (SF_6)

En un principio México no se encontraba en la lista de los países firmantes pero en 1998 México se unió al grupo de los 187 países en ratificar este tratado (Sheinbaum y Masera, 2000).

La primera fase del protocolo de Kioto entró en vigor el 16 de febrero de 2005 y comprendía un periodo de cuatros años, de 2008 a 2012. En la decimosexta Conferencia de las Partes (COP 16) celebrado en Cancún Quintana Roo, México y los más de 190 países asistente adoptaron un acuerdo por el cual se aplazó el segundo periodo de vigencia del protocolo de Kioto que abarca a partir del 1 de enero de 2013 hasta el 31 de diciembre de 2020, el cual fue ratificado hasta la Decimoctava Conferencia de las Partes (COP 18) sobre Cambio Climático, (http://www2.inecc.gob.mx/publicaciones/libros/437/tudela.html).

En el COP 16 también se llegó a un acuerdo con la reserva de Bolivia, donde se decidió crear un Fondo Verde Climático, dentro de la Convención Marco, que contará con un consejo de 24 países miembro y será diseñado por un comité de transición que formarán 40 países. También se llegó al compromiso de proporcionar 30.000 millones de dólares de financiación rápida, aunque se reconoce la necesidad de movilizar 100.000 millones de dólares por año a partir de 2020 para atender a las necesidades de los países en desarrollo. Los acuerdos se complementan con el Plan de Acción de Bali, que identifica cuatro elementos clave: mitigación, adaptación,

finanzas y tecnología. Después del COP 16 la canalización de financiamiento y tecnología de apoyo a países en desarrollo tuvo avances importantes y para el COP 18, los países desarrollados reiteraron su compromiso de continuar el financiamiento a largo plazo, con miras a movilizar 100 mil millones de dólares para adaptación y mitigación hasta el 2020, (http://cc2010.mx/).

El protocolo de Kioto fue el punto de partida para que muchos países volcarán la vista hacia el cuidado y protección del medio ambiente y emprendieran una serie de programas y evaluaciones para cumplir con los resultados propuestos. En México se emprendió una serie de acciones buscando contribuir a mitigar el cambio climático a través de políticas públicas. Alguno documentos que se pueden consultar al respecto son "La Visión de México sobre REDD+", "Propuesta de preparación REDD (R-PP), "Inventario Nacional de Emisiones de Gases Efecto Invernadero 1990 - 2006" y la Estrategia Nacional REDD+.

En investigaciones llevadas a cabo en el Colegio de Postgraduados, relacionadas con la Estrategia Nacional REDD+, se encuentra y que hacen uso directo de la información del INFyS 2004-2009 encontramos a Méndez y De los Santos (2011), donde emplean diferentes ecuaciones de volumen-biomasa, algunas extraídas de compendios estatales, y otras obtenidas mediante el promedio de los coeficiente de las ecuaciones, aplicado sólo para aquellas especies que carecían de ecuaciones de volumen regionales. El objetivo de este trabajo es el de estimar el potencial de captura de carbono de los bosques tropicales y templados de México.

En investigaciones de este tipo es donde surge la importancia de establecer mecanismos para el cálculo automático de algunos parámetros o indicadores que puedan ser recurrentes y que se estén actualizando constantemente con la integración de información nueva. Como caso particular se tiene las estimaciones de volumen de madera, biomasa y carbono que están directamente relacionados de forma jerárquica, ya que para determinar la captura de carbono es necesario tener las estimaciones de biomasa y para estimar la biomasa es necesario calcular el volumen de madera. Uno de los objetivos de esta tesis es dar conocer nuevas herramientas para realizar cálculos más eficientes a partir de grandes volúmenes de información como es la proveniente del INFyS, ejemplificando los pasos a seguir para modelar cualquier proceso de cálculo. Para mostrar lo anterior, en el presente trabajo se procederá al cálculo de volumen de

madera biomasa y carbono, que servirá a los expertos forestales en sus reportes e investigaciones, como las que se realizan como parte de las estrategias de mitigación y adaptación al cambio climático.

6.5 Métodos para estimar volumen de madera, biomasa y carbono.

6.5.1 Método utilizado por la CONAFOR para estimar volumen de madera.

En CONAFOR (2012), se detalla el método utilizado para la estimación de volumen, misma que a continuación se resume.

El volumen de cada árbol se obtiene a partir de una ecuación que incluye como variables el diámetro normal y la altura total. Las ecuaciones empleadas son las mismas que se utilizaron en el Primer Inventario Nacional Forestal (1961-1985) y se aplican por especie, grupos de especies y/o región, acorde a las especificaciones de cada modelo.

Las ecuaciones de volumen empleadas se concentran en el Anexo 7 del INFyS. En los Estados que carecían de una ecuación de volumen se utilizó las ecuaciones de otras entidades que presentaban características similares en cuanto a especies, grupo de especies o condiciones ambientales y para aquellos que presentaban diferentes ecuaciones por región, sólo se consideró la que tuvo mayor número de observaciones, y valores más para R^2 y F. En entidades como Quintana Roo se aplicaron directamente las tablas de volúmenes y de manera similar para Chihuahua y Durango para los géneros *Quercus, Abies, Picea* y *Pseudotsuga*. En Chihuahua, Durango, Sinaloa, Sonora y Zacatecas se utilizó el coeficiente mórfico para especies latifoliadas.

Para la utilización de los modelos se filtró la información de modo que sólo se consideró a aquellos árboles cuyo diámetro normal (medición del diámetro del fuste a 1.30 metros de altura a partir del nivel del suelo) y la altura correspondiera al rango contemplado en ellos. Los filtros utilizados se listan a continuación:

- Altura mayor o igual a cinco metros y menor o igual a 47.5 metros (en algunos casos se utilizó una altura máxima de 42.5 metros).
- Diámetro normal mayor o igual a 7.5 centímetros y menor o igual a 132.5 centímetros (el algunos casos se utilizó un diámetro mínimo de 12.5 centímetros y un máximo de 112.5 centímetros).
- Se toman en cuenta sólo los géneros maderables.
- Se consideran los árboles vivos y muertos en pie

El volumen calculado a nivel de unidad de muestreo secundaria (sitio de 400 metros cuadrados) se obtuvo sumando el volumen de cada árbol presente en él (considerando las restricciones). Para calcular el volumen a nivel de unidad de muestreo primaria, se suma el volumen de las unidades de muestreo secundarias que lo forman. La variable auxiliar (área medida en hectáreas) se calcula contando los sitios que fueron muestreados en cada conglomerado y se multiplica por 0.04.

6.5.2 Estimación del volumen de madera mediante ecuaciones de volumen

Méndez y De los Santos (2012) propusieron dividir al país en 8 regiones y aplicar ecuaciones de volumen por género, de manera similar al realizado por la CONAFOR. Las ecuaciones utilizadas fueron obtenidas de compendios estatales y las que son de interés, dado los alcances de la tesis, son las correspondientes a la región centro del país que contempla a los estados de Hidalgo, México, Morelos, Puebla, Querétaro, Tlaxcala, Distrito Federal). Las ecuaciones utilizadas son las siguientes:

$$VAbies = e^{(-9.380673332+1.72200226*\text{Log(DN)}+1.059321587*\text{Log(H)})} \quad (1)$$

$$VCupressus = e^{(-9.380673332+1.72200226*\text{ Log(DN)}+1.059321587\ *\text{Log(H)})} \quad (2)$$

$$VJuniperus = e^{(-9.380673332+1.72200226*\text{ Log(DN)}+1.059321587\ *\text{Log(H)})} \quad (3)$$

$$VPseudotsuga = e^{(-9.380673332+1.72200226* \text{Log(DN)}+1.059321587 *\text{Log(H)})} \quad (4)$$

$$VQuercus = e^{(-9.739914454+1.9333872* \text{Log(DN)}+0.983925879* \text{Log(H)})} \quad (5)$$

Para el género *Pinus* se considera un promedio de los coeficientes para toda la región

$$VPinus = e^{((-9.640658417+1.855774305* \text{Log(DN)}+1.00058223 *\text{Log(H)})} \quad (6)$$

Para otras especies se considera un promedio de los coeficientes de todas las especies hojosas.

$$VOtras = e^{(-9.635727101+1.854491535* \text{Log(DN)}+1.018174717* \text{Log(H)})} \quad (7)$$

Al final se obtiene el volumen total sumando el volumen de todos los géneros

$$Volumen = VAbies + VCupressus + VJuniperus + VPseudotsuga + VQuercus + VPinus + VOtras \quad (8)$$

Donde

DN: diámetro normal
H: altura

Las ecuaciones de volumen que aquí se presentan forman parte del Anexo 1 en Méndez y De los Santos (2003). Estas ecuaciones están escritas en lenguaje R y utilizan variables indicadoras para correr los análisis (para una mejor visualización se escribieron en formato de ecuaciones).

6.5.3 Estimación de volumen de madera usando factores de forma.

Este método también fue extraído de Méndez y De los Santos (2003), el cual utiliza factores de forma para cada especie. La ecuación utilizada se muestra en la ecuación (9) y los factores se presentan en el Cuadro 7.

$$Volumen = \frac{\pi}{40000} * DN^2 * H * Factor \quad (9)$$

Donde

DN: diámetro normal
H: altura

Factor: representa al factor de forma el cual puede tener diferentes valores.

Cuadro 7: Factores de forma por género para el cálculo del volumen de madera

Genero	Factor
Abies	0.33
Pinus	0.33
Cupressus	0.3
Juniperus	0.26
Pseudotsuga	0.36
Quercus	0.32
Otros Géneros	0.30

Fuente: Elaboración propia para la investigación.

6.5.4 Estimación de volumen de madera usando una ecuación general

El cuarto método para calcular el volumen de madera también fue obtenido de Méndez y De los Santos (2003). Este método sólo hace uso del diámetro normal y la altura del árbol para calcular su volumen. No se considera el género o especie del árbol, lo que representa una ecuación general para obtener los volúmenes.

$$Volumen = 0.00004 * DN^2 * H \qquad (10)$$

Donde

DN: diámetro normal

H: altura

6.5.5 Estimación de la biomasa

El término biomasa en un sentido más amplio incluye toda la materia viva existente en un instante de tiempo en la Tierra. La biomasa energética también se define como el conjunto de la materia orgánica, de origen vegetal o animal, incluyendo los materiales procedentes de su

transformación natural o artificial. Cualquier tipo de biomasa tiene en común con el resto de provenir de la fotosíntesis vegetal (INE, 2006).

La estimación se biomasa se llevó a cabo considerando que el 50% del volumen corresponde a la biomasa total aérea, dicho factor corresponde con el que reportar el IPCC, Méndez y De los Santos (2011)

$$Biomasa = Volumen\ de\ madera * 0.5 \qquad (11)$$

6.5.6 Estimación de Carbono

La estimación de carbono se realiza aplicando un descuento del 50% a la biomasa calculada.

$$Carbono = Biomasa * 0.5 \qquad (12)$$

6.6 Resumen

En éste capítulo se introdujo la importancia de los Inventarios Forestales describiendo una serie de actividades. En primer lugar se describieron los hechos relevantes acontecidos a nivel mundial que dieron lugar a los primeros Inventarios Forestales Mundiales. Después se proporcionó descripción cronológica de los inventarios forestales llevados a cabo en territorio nacional, las actividades y los principales resultados obtenidos, como fue el método de muestreo que se utiliza y su base de datos.

Posteriormente, se detalló la importancia de calcular el volumen el carbono en bosques y selvas, como indicador para medir el cambio climático. Finalmente, se mostraron los principales métodos y ecuaciones utilizados para calcular estos indicadores, retomando las propuestas de la CONAFOR y de Méndez y de los Santos (2011).

IV MARCO EMPÍRICO

7. METODOLOGÍA DE LA INVESTIGACIÓN

7.1 Tipo de investigación

La presente investigación cae en la categoría cualitativa (Hernández Sampieri *et al.*, 2008), y de diseño de software.

7.2 Población y muestra

Población: La población comprende todos los registros de la base de datos del Inventario Forestal Nacional y de Suelos 2004-2009.

Muestra: La muestra comprende todos los registros que contengan información sobre los árboles de géneros maderables, vivos y muertos en pie, cuyo diámetro normal sea mayor o igual a 7.5 centímetros y menor o igual a 132.5 centímetros y una altura mayor o igual a 5 metros y menor o igual a 47.5 metros, cuya información haya sido levantada en los conglomerados y sitios de medición pertenecientes al Estado de México. Además, se utiliza una muestra adicional del 10% de la información dasométrica de los géneros maderables *Quercus, Pinus, Bursera, Lysiloma y Piscidia*, a nivel nacional, para el entrenamiento de modelos de minería de datos en la clasificación del género arbóreo *Quercus.*

7.3 Enfoque cualitativo

Hernández Sampieri *et al.* (2008) señala que el enfoque cualitativo se enfoca en métodos de recolección de datos a través de descripciones y observaciones sin realizar mediciones numéricas.

Instrumento de recogida de datos cualitativos: Se diseñó una entrevista semiestructurada dirigida a los usuarios de los datos del Inventario Nacional Forestal del Programa Forestal del

Colegio de Postgraduados Campus Montecillo (Anexo 1) y un cuadro de rúbricas para evaluar a las interfaces de análisis, desarrolladas para interactuar con los paquetes de SQL Server 2008, utilizados para manipular, procesar y analizar los datos del Inventario Nacional Forestal y de Suelos 2004-2009 (Anexo 2).

Recolección de datos cualitativos: Se entrevistó a 4 profesores investigadores del Programa Forestal en el Colegio de Postgraduados Campus Montecillo, que hacen uso de la información de la base de datos del Inventario Nacional Forestal. Las entrevistas se enfocaron en el tema de calidad de los datos, metodologías para la extracción de información y herramientas tecnológicas utilizadas para su análisis. Se programó dichas entrevistas para el mes de febrero de 2014 en las oficinas de los propios docentes.

La rúbrica será empleada para evaluar el software al final del diseño aproximadamente en el mes de septiembre de 2014.

Análisis de resultados: Análisis de la información recabada en las entrevistas a través de análisis de discurso y de las rúbricas a través de tendencias en las respuestas.

La información recabada en las entrevistas servirá como punto de partida para el diseño del Data Warehouse y el desarrollo del paquete para extracción, transformación y carga (ETL). En la Figura 15 se muestra de forma gráfica el procedimiento que se va a seguir.

Figura 15. Proceso de análisis cualitativo.
Fuente: Elaboración propia para la investigación.

Diseño de Software: Consiste en la creación de paquetes independientes que operan con SQL Server 2008 y versiones posteriores, y el desarrollo de interfaces para visualizar y analizar la

información procesada por los paquetes; la fuente principal de datos es la base de datos del Inventario Nacional Forestal y de Suelos 2004-2009, en formato de Microsoft Access y ecuaciones para el cálculo de volumen para el género arbóreo *Quercus* en el estado de México, almacenadas en un archivo de Microsoft Excel. El paquete para la extracción, transformación y carga de datos hacia el Data Warehouse, fue desarrollado usando las herramientas del servicio de integración de SQL Server 2008. El paquete para el procesamiento y análisis de la información mediante cubos multidimensionales fue desarrollado usando las herramientas del servicio de análisis de SQL Server 2008. Las interfaces para visualización y análisis fueron desarrolladas usando Visual Studio 2010. En la Figura 16 se muestra el proceso seguido en el desarrollo de los paquetes y las interfaces de análisis.

7.4 Fases de la investigación

Fase 1. Recopilación de información teórica. En ésta fase se llevó a cabo la revisión de diversas publicaciones como libros, artículos científicos y de divulgación, revistas electrónicas, informes, reportes y páginas web, en especial las de la Secretaría de Medio Ambiente y Recursos Naturales (SEMARNAT), Comisión Nacional Forestal (CONAFOR), Instituto Nacional de Ecología (INE), Instituto Nacional de Investigaciones Forestales, Agrícolas y Pecuarias (INIFAP); la información recopilada abarcó diversos temas como son Bases de datos, Data Warehouse, Minería de datos e Inventario Nacional Forestal.

Fase 2. Análisis de la base de datos del Inventario Nacional Forestal y de Suelos 2004-2009: En esta se reconstruyó el esquema de la base de datos donde se muestra sus respectivas entidades, relaciones y cardinalidades.

Fase 3. Desarrollo del paquete de análisis de información a través de modelos de minería de datos. En ésta etapa se utilizó el módulo de análisis de SQL Server 2008 para crear un paquete que permita la generación de modelos de minería de datos a partir de la información de la base de datos del Inventario Nacional Forestal y de Suelos 2004-2009. Los modelos generados tienen como objetivo la clasificación de árboles del género *Quercus*.

Fase 4. Diseño del Data Warehouse. El Data Warehouse fue diseñado a partir de la información recopilada en entrevistas a profesores investigadores del Programa Forestal del Colegio de Postgraduados e investigaciones relacionadas con los inventarios forestales. El diseño del Data Warehouse se enfocó en las tareas de almacenamiento de datos depurados y fuente de información para las consultas sobre volumen de madera, biomasa y carbono.

Fase 5. Desarrollo del paquete para Extracción, transformación y carga de datos (ETL): Durante esta etapa se utilizó las herramientas de SSAS de SQL Server 2008 para automatizar el proceso de extracción, transformación y carga de datos hacia el Data Warehouse. Mediante una serie de objetos se programaron tareas para la selección de información a partir de la base de datos del Inventario Nacional Forestal y de Suelos 2004-2009 y se compiló dentro de un paquete para automatizar este proceso y tenerlo disponible cada vez que se realice una actualización de la base de datos.

Fase 6. Desarrollo del paquete de análisis multidimensional. En ésta etapa se utilizó el módulo de análisis de SQL Server 2008 para crear un paquete de análisis de información a partir del Data Warehouse. Este paquete consta de un cubo de análisis multidimensional que contempla 2 dimensiones para generar información a diferentes niveles de agregación.

Fase 7. Desarrollo de interfaces de consulta: Para la consulta de la información contenida en el cubo multidimensional se desarrollaron dos interfaces de escritorio, una aplicación cliente servidor y una aplicación portable para investigadores. Adicionalmente se ejemplificó el uso de Microsoft Excel para llevar a cabo un análisis de la información mediante tablas y gráficas dinámicas.

Fase 8. Redacción de resultados: Está fase de llevó de manera conjunta con forme se fueron desarrollando las fases del 2 al 7, y consistió en describir las metodologías empleadas y en la representación gráfica de los resultados obtenidos.

Fase 9. Elaboración del documento final: Se integró el trabajo final a partir del marco introductorio, marco teórico, marco contextual, resultados, conclusiones y recomendaciones.

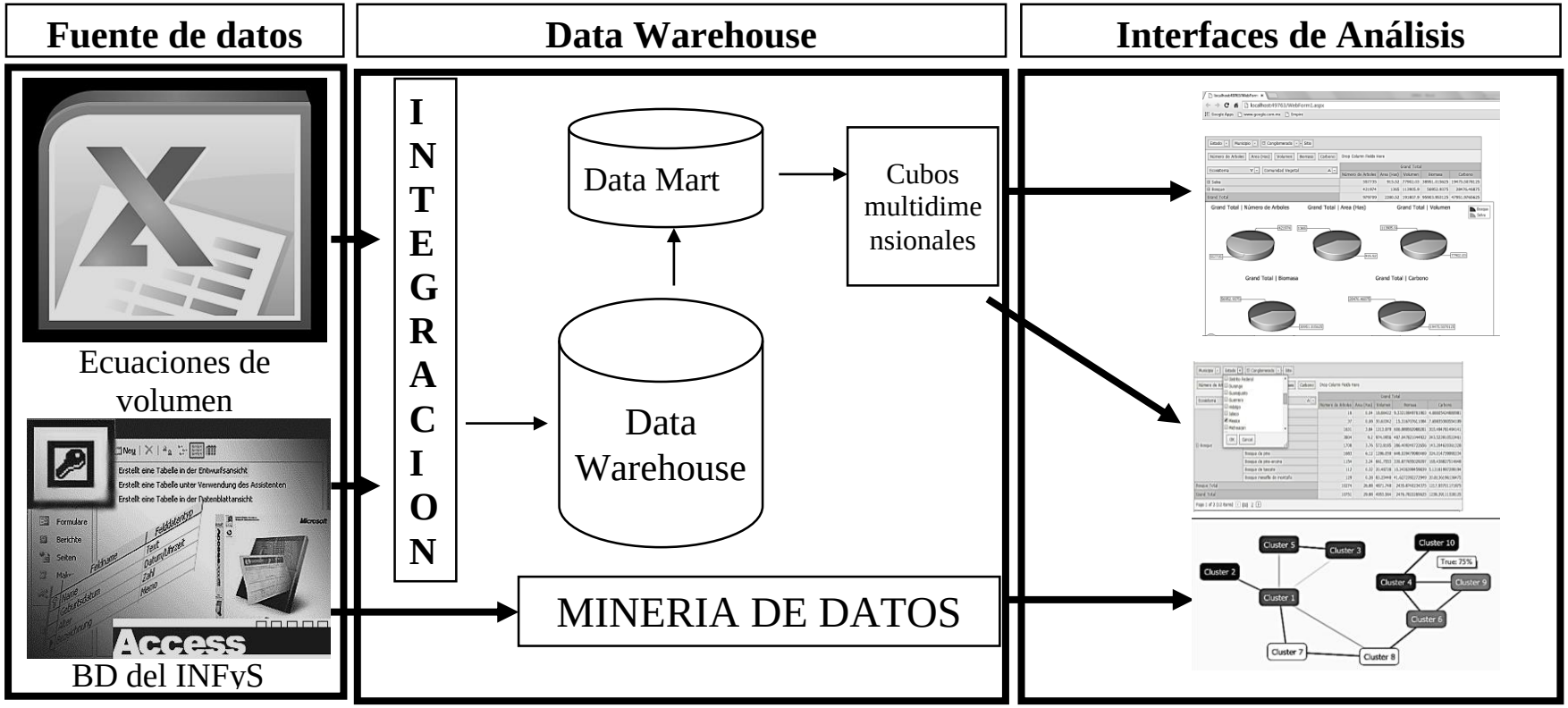


Figura 16. Proceso de desarrollo de aplicaciones de visualización y análisis de información.
Fuente: Elaboración propia para la investigación.

7.5 Proceso de obtención de resultados

Para comenzar con el diseño del software plasmado en la Figura 16, se hizo un Análisis de datos de las entrevistas semiestructuradas aplicadas a profesores (Anexo 1) del cual se obtuvieron los siguientes resultados:

1. *Acceso de la Información de la base de datos del Inventario Nacional Forestal y de Suelos 2004-2009:* Los docentes del CP contestaron que no es fácilmente accesible ya que no es transparente, se debe solicitar por oficio.
2. *La información del INFyS que más utiliza.* Los Investigadores del CP utilizan más las variables dasométricas y geográficas para hacer mapas o Sistemas de información Geográficos.
3. *Problemática más frecuente al procesar la información forestal*. La información es inconsistente, redundante y presenta muchos datos faltantes.
4. *Software que utiliza para analizar los datos*. R, SAS y Excel.
5. *Propuestas de mejora de la calidad de datos*. La mayoría expresa que se debe validar mejor la información que se vacía en las bases de datos.
6. Mejora de la estructura: La mayoría contesto que sí pero no saben mucho de cómputo. Hacen uso del R para investigaciones pero les cuesta mucho adaptar los datos.
7. Mejora de acceso: La mayoría considera que hacer un sistema mejor de bases de datos que les permita acceder a la información en línea.
8. Conocimiento sobre las base de datos multidimensionales: Solo uno dijo que sí, la mayoría pregunto qué era eso.
9. Conocimiento sobre Data Warehouse: La mayoría contesto que no.
10. Conocimiento sobre minería de datos: Sobre ese tema si han escuchado y utilizado en investigaciones con el programa de Estadística del CP.

Después de analizar la información y de entrevistar a los diferentes docentes se puede determinar que:

- El sistema computacional que contiene la Base de Datos del INFyS (2004-2009) es complicada
- La INFyS tiene muchos errores
- Los investigadores requieren de un software que les facilite la obtención de datos de una manera más eficiente y accesible desde cualquier lugar a través de Internet
- Los docentes necesitan de una herramienta computacional que les ayude en cálculos muy particulares en un menor tiempo y con mayor precisión.

Con la información recabada en las entrevistas, se diseñó del Data Warehouse y el desarrollo del paquete para extracción, transformación y carga.

El proceso de obtención de resultados y productos, de la presente investigación, se llevó a cabo siguiendo los pasos que se muestran en la Figura 17. Los resultados se presentan más adelante en capítulos independientes

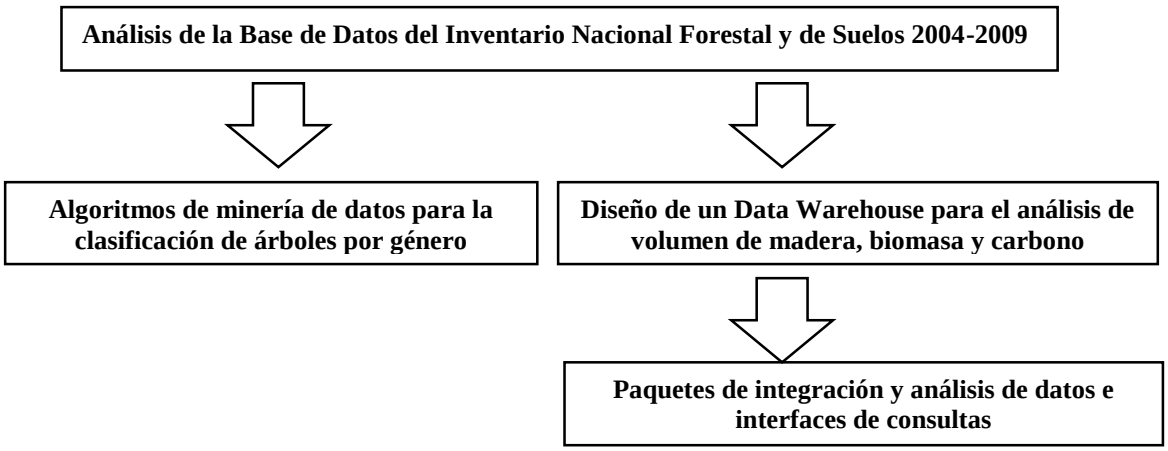


Figura 17. Procedimiento para la obtención de resultados y productos de la investigación.
Fuente: Elaboración propia para la investigación

Análisis de la base de datos del Inventario Nacional Forestal y de Suelos 2004-2009: En ésta sección se analizó el tipo de información que se almacena en la base de datos del Inventario Nacional Forestal y de Suelos 2004-2009; posteriormente, se reconstruyó su diagrama, se identificaron el conjunto de entidades, atributos, relaciones y cardinalidades. Después, se analizaron las 12 reglas, propuestas por Edgar F. Codd en 1984, que deben cumplir los sistemas de bases de datos relacionales y finalmente se seleccionaron aquellas reglas relacionadas con el

diseño de base de datos, las restantes se relacionan con el sistema gestor de bases de datos y no son de competencia para la presente investigación.

Algoritmos de minería de datos para la clasificación de árboles por género: En esta sección se probaron cuatro modelos de minería de datos, para la clasificación del género arbóreo *Quercus,* usando como datos de entrenamiento, y validación la información del INFyS 2004-2009.

Diseño de un Data Warehouse para el análisis de volumen de madera, biomasa y carbono: Para el diseño del Data Warehouse se siguió la metodología propuesta por Ralph Kimball que consiste en identificar el proceso de análisis, definir los niveles de granularidad, las dimensiones y las tablas de hechos y finalmente la modelación del proceso.

Paquetes de integración y análisis de datos e interfaces de consultas: En ésta sección se describen los paquetes desarrollados para la integración, transformación y carga de datos hacia el Data Warehouse así como el de análisis multidimensional. Se describe cada uno de los objetos que componen los paquetes y sus diferentes tareas. Finalmente se detalla en el uso de las interfaces desarrolladas para interactuar entre los usuarios finales y los paquetes mencionados anteriormente.

8. ANÁLISIS DE LA BASE DE DATOS DEL INVENTARIO NACIONAL FORESTAL Y DE SUELOS 2004-2009

En la sección que se describe se lleva a cabo un análisis detallado de la base de datos (BD) del Inventario Nacional Forestal y de Suelos (INFyS) 2004-2009, el cual contempla tres etapas como se muestra en la Figura 18.

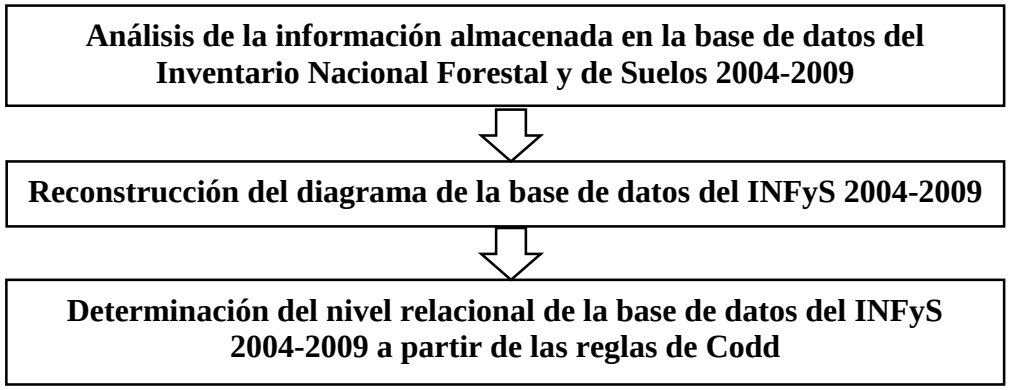


Figura 18. Análisis de la base de datos del Inventario Nacional Forestal y de Suelos.
Fuente: Elaboración propia para la investigación

La primera etapa del análisis está enfocada en dar a conocer la información que se registra y almacena en la BD del INFyS; así como describir las diferentes actividades que la originan. En la segunda etapa se reconstruye el modelo conceptual de la BD para tener una representación visual de las relaciones y cardinalidades de todas las tablas que conforman la BD del INFyS 2004-2009. En la tercera etapa se explican y analizan las reglas de Codd establecidas para los sistemas de bases de datos relacionales con la finalidad de seleccionar aquellas relacionadas con el diseño de bases de datos relacionales y utilizarlas para determinar el nivel relacional de la base de BD.

8.1 Análisis de la información almacenada en la base de datos del INFyS 2004-2009

Para entender la distribución de las tablas dentro de la base de datos del INFyS 2004-2009, primero es importante describir la información que guarda cada una de ellas o por lo menos la

información que contienen las tablas más representativas. Para ello a manera de resumen se describirá el proceso completo desde la toma de datos en campo hasta el almacenamiento de la información validada en la base de datos del servidor central.

Las características del muestreo en campo, cliente de captura y los controles incorporados para garantizar la integridad calidad de los datos fueron tomados de CONAFOR (2012). La secuencia de las actividades realizadas durante todo el proceso se observa en la Figura 19.

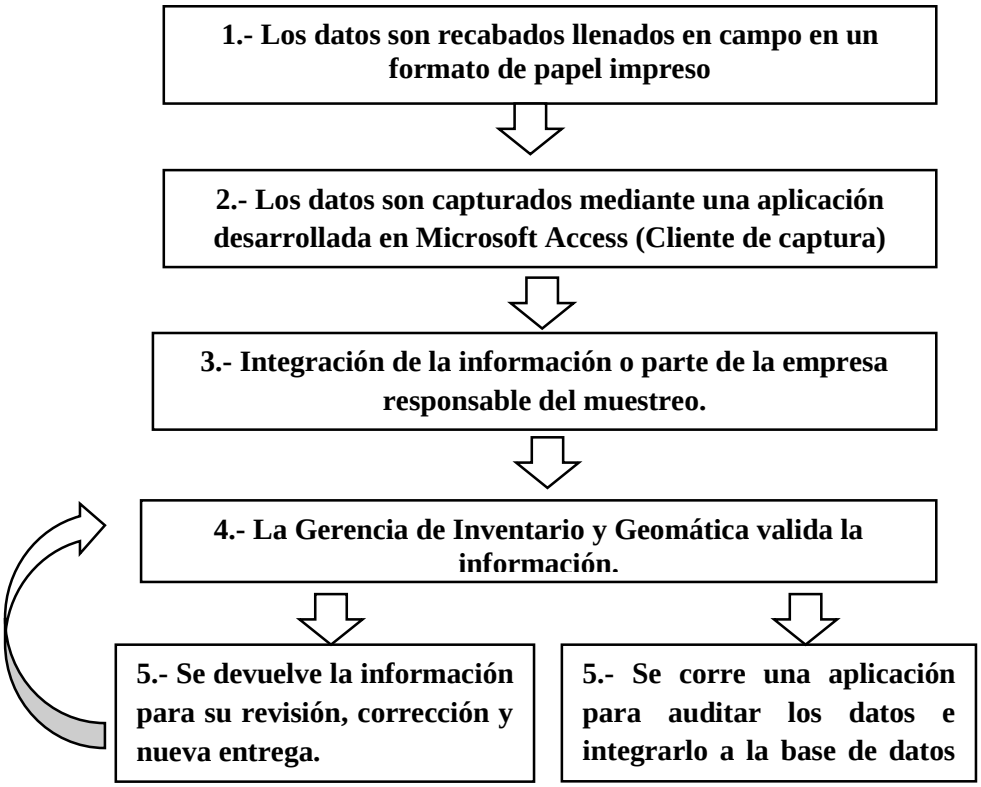


Figura 19: Proceso de recopilación y validación de datos
Fuente: Elaboración propia para la investigación

Los datos en campo son recabados en formatos impresos en papel, que forman parte del anexo H del Manual de Muestreo de Campo, el cual se encuentra disponible en la liga http://www.conafor.gob.mx:8080/documentos/docs/8/4125CNF-24_INFyS.pdf . Este manual fue diseñado con la finalidad de presentar los informes y los instructivos correspondientes para el levantamiento de la información de campo del Inventario Nacional Forestal y de Suelos; y a partir del año 2010, ampliar la información del levantamiento de datos de campo de acuerdo a

los principales ecosistemas donde se ubica cada conglomerado o unidad de muestreo en el país. Como ya se había mencionado anteriormente, para el año 2010 se incorporaron nuevas variables a los formatos de campo del INFyS sobre el tema de sanidad forestal y pastizales naturales contemplado anteriormente dentro de otra categoría. El Anexo H del manual consta de 6 formatos enlistados a continuación:

- Formato e instructivo para bosques.
- Formato e instructivo para selvas, peten, manglar y comunidades subacuáticas.
- Formato e instructivo para zonas áridas, semiáridas, palmar y galería.
- Formato e instructivo para suelos.
- Formato e instructivo de conglomerados sin cubierta vegetal.
- Formato e instructivo de conglomerados justificados.

Posterior al levantamiento de datos en campo, se procede a la captura de la información en un formulario de Microsoft Access conocido como cliente de captura. La flexibilidad que les brinda un sistema gestor de bases de datos como Access es la movilidad de la información, Microsoft Access guarda la información en un solo archivo y a diferencia de muchos otros gestores, no se necesita requerimientos especiales de hardware para poder ser funcional; además de poder realizar las modificaciones del archivo como integración de nuevos registros, modificación de los registros existentes y eliminación de los mismos desde cualquier computadora que tenga instalado Microsoft Access.

El cliente de captura es un formulario que consta de módulos con características adecuadas a la planeación del inventario en cada fase. Cuenta módulos especializados para la captura de información con secciones y objetos dispuestos en el mismo orden que el formato impreso usado en campo, correspondiente a los formatos de bosques, selvas y otro para el de comunidades áridas y semiáridas. Esto módulos permiten agregar expedientes completos de conglomerados, editarlos, eliminarlos, visualizar o imprimir, en forma de reporte para facilitar su revisión; además de incorporar distintas validaciones y controles automatizados de calidad con parámetros preestablecidos, implementados en cada una de las secciones con la finalidad de minimizar el error de captura y proporcionar información adicional de la ubicación y tipo de

vegetación esperada para cada conglomerado. Los datos que incorpora el cliente de captura provienen del conjunto de datos vectoriales de las Cartas de Uso de Suelo y Vegetación del INEGI, serie III y IV, escala 1:250 000 (CONAFOR, 2012).

En la CONAFOR se desarrolló un módulo encargado especialmente para la tarea de la integración de la información; éste es el encargado de validar todo el contenido de las tablas y los registros asociados al conglomerado. En caso de que se detecte información incompleta, sea invalida o que el registro ya existe en la base de datos de destino la importación es rechazada. Para el caso de registros existentes se tiene la opción de reemplazar por el nuevo registro o eliminar de la base de datos los registros repetidos. La empresa encargada del levantamiento de datos en campo entrega a la CONAFOR los formatos impreso y las bases de datos integradas (Aldana, 2012).

En la Gerencia de Inventario Forestal y Geomática la información pasa por un segundo punto de control donde se coteja el contenido del formato en papel contra el digital con la finalidad de detectar inconstancias entre ambos. Aquellos marcados como inconsistentes son devueltos para su revisión, corrección y nueva entrega, sino son almacenados en una base de datos dentro del servidor central conocido como TULE (CONAFOR, 2012). El proceso de integración y validación de la información se muestra en la Figura 20.

La base de datos central de igual forma que la base de datos del cliente de captura, está diseñado bajo un esquema relacional, pero administrado por un sistema gestor de base de datos más potente, como es Microsoft SQL Server.

Figura 20. Integración y control de calidad de los datos del INFyS 2004-2009.
Fuente: Tomado de CONAFOR (2012)

De manera similar como actúa el módulo de integración de información en el cliente de captura, para la importación de los datos, desde los archivos entregados por las empresas a la base de datos central, se desarrolló una aplicación que realiza una segunda auditoría a los datos para reforzar la calidad e integridad de la información. No se encontró información sobre la arquitectura de la aplicación encargada de la auditoria e integración de la información, pero al usar SQL Server como sistema gestor, se supone que ésta fue desarrollada con Visual Studio C, C++ o .Net. En la presente investigación se desarrolló un módulo similar a esta aplicación pero usando la tecnología incorporada en SQL Server 2008 Enterprise, las herramientas de SSIS, las cuales son capaces de realizar el proceso de validación e integración de información a partir de archivos Access de forma muy eficiente.

8.2 Reconstrucción del diagrama de la base de datos del INFyS 2004-2009

En esta sección se realizará un análisis detallado de la base de datos de acuerdo a como se lista en la Figura 21. Primero se localizarán las tablas que conforman la BD del INFyS 2004-2009, correspondientes a las reportadas por la Gerencia de Inventario y Geomática. Posteriormente se hace una revisión de llaves primarias y secundarias de éstas tablas. Después se analizará las cardinalidades de cada una de las tablas y finalmente se reconstruirá el Diagrama Entidad-Relación.

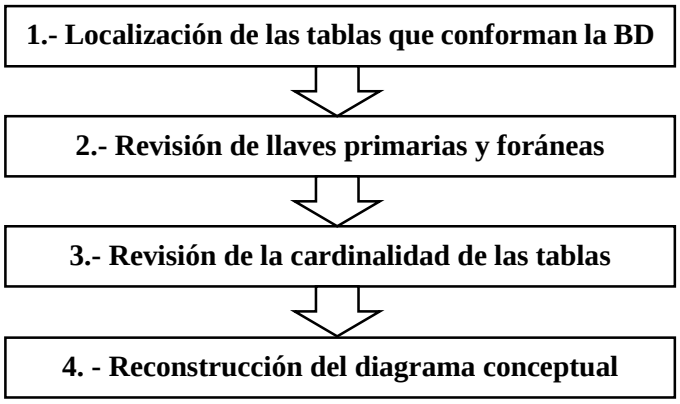


Figura 21. Proceso de análisis para la base de datos del INFyS 2004-2009
Fuente: Elaboración propia para la investigación

La Gerencia de Inventario y Geomática, en CONAFOR (2012), reporta que el diseño y estructura de la BD están basados en un modelo Entidad-Relación, con el objetivo de satisfacer las siguientes necesidades:

- Mantener una independencia lógica y física de los datos
- Evitar la redundancia de información, para garantizar la integridad y calidad de los datos.
- Realizar consultas complejas optimizadas.

Como se ha estado enfatizando, la importancia de la BD del INFyS radica en tener a la mano información actualizada y de calidad, que pueda ser empleada para la elaboración de reportes nacionales e internacionales; así como servir de instrumento para la toma de decisiones en las políticas públicas. Con base en estos argumentos se debe garantizar que la información almacenada en dicha BD, no presente problemas de inconsistencia, con respecto a la integridad y calidad de los datos, reforzándose de esta manera los objetivos que reporta la Gerencia de Inventario y Geomática en cuanto al diseño empleado para la BD.

La base de datos del INFyS está conformado por 67 tablas distribuidas de la siguiente manera: 2 tablas principales, 22 tablas secundarias y 43 tablas del tipo catálogo.

Las dos tablas principales son la tabla **TblConglomerado** y la tabla **TblSitio**. Para tener una descripción más detallada del tipo de información que se almacena en estas tablas es necesario definir algunas fases del proceso de levantamiento de información en campo.

La información recopilada en cada conglomerado hace referencia a las condiciones físicas y geográficas de la zona. La forma que adoptan corresponde a una parcela circular de una hectárea (56.42 metros de radio), en la cual se evalúan 4 unidades de muestreo secundarias o sitios. El tipo de conglomerado en todas las clases de vegetación corresponde a una "Y" invertida, variando la forma de las unidades de muestreo secundarias: rectangular para el caso de las selvas y circular para el resto de la vegetación, el área en ambos diseños de la parcela es la misma. La distancia entre un conglomerado y otro varía según el tipo de vegetación que se trate, tal y como se describe a continuación.

- 5 X 5 kilómetros en bosques de coníferas, coníferas latifoliadas, y bosque mesófilo, así como selvas altas, medianas y manglares.
- 10 X 10 kilómetros en bosque bajo abierto, selvas bajas, matorral subtropical y vegetación semiárida.
- 20 X 20 kilómetros en vegetación de zonas áridas.

La importancia de esta tabla radica en que cada conglomerado es una unidad de muestreo primaria utilizada para llevar a cabo las estimaciones de los principales indicadores y parámetros. La llave principal de esta tabla se encuentra en el campo **IdConglomerado**. A continuación se listan las tablas relacionadas con la tabla **TblConglomerado**.

1. TblCaracEspFlora
2. TblCoordPtoCtrl
3. TblCuerpodeAgua
4. TblDiversidadXEstrato
5. TblEpifita
6. TblImpactoAmbiental
7. TblIncendio
8. TblJustificado
9. TblSuelo

La tabla **TblSitio** almacena los atributos de la información recabada en cada unidad de muestreo secundaria. Su llave primaria es el campo llamado **IdSitio**, a través del cual se vincula con las tablas que registran la información específica de la vegetación encontrada dentro del área de cada uno de los cuatros sitios del conglomerado, en el caso de que los cuatro hayan sido accesibles, de lo contrario no se registra información.

La CONAFOR (2012) reporta que la información almacenada en esta tabla **TblSitio** corresponde a la que cumpla con la siguiente descripción

a) En el sitio de 400 metros cuadrados (radio 11.28 metros) se mide y registra el arbolado cuyo diámetro normal sea mayor o igual a 7.5 cm.
b) En el sitio de 12.56 metros cuadrados se mide y registra, por género, la frecuencia y algunas variables cualitativas del repoblado (regeneración natural), cuyas planta o

árboles pequeños tengan como mínimo 25 cm de altura, hasta la altura que alcancen, siempre que su diámetro normal sea menor de 7.5 cm. Así mismo, se registran los arbustos representativos de las comunidades áridas, e incluso especies invasoras y de pastos nativos e inducidos.

c) En el sitio de 1 metro cuadrado se miden y consignan las plantas herbáceas, helechos, líquenes, musgos y otras características de la superficie del suelo presentes en el subestrato herbáceo.

Las tablas secundarias relacionadas directamente con la tabla de sitios se listan a continuación

1. TblArboladoBosqueSelva
2. TblArboladoSubBosqueSelva
3. TblArbSubMuestraOtrasCom
4. TblCobertura
5. TblCoberturaOtrasC
6. TblCoordenada
7. TblCoordenadaSitio
8. TblRepobladoBosque
9. TblRepobladoOtrasCom
10. TblRepobladoSelva
11. TblVegMayorOtrasCom
12. TblVegMenorBosqueSelva
13. TblVegMenorOtrasCom

Además de las tablas referidas, la base de datos cuenta con 43 tablas auxiliares de tipo catálogo que almacenan los atributos y descripciones de las claves utilizadas en las tablas mencionadas anteriormente. A continuación se listan los 43 catálogos de la BD del INFyS 2004-2009.

1. CatAbundancia
2. CatAccesibilidad
3. CatAniosIncendios
4. CatCarta150
5. CatCategoriaSueloXProfundidad
6. CatCausaImpacto
7. CatCoberturaXVeg
8. CatCondicion
9. CatCuerpoAgua
23. CatImpactoVegSueloH2O
24. CatMantillo
25. CatMercadoEspecie
26. CatMunicipio
27. CatNivelAfectacion
28. CatNivelAfectacionH2O
29. CatTenencia
30. CatTipoAcceso
31. CatTipoConglomerado

10. CatDanio
11. CatDatosAutomaticos
12. CatDegradacion
13. CatEpifitaTipo
14. CatErosion
15. CatEspecie
16. CatEspFlora
17. CatEstado
18. CatExposicion
19. CatFisiografia
20. CatFisonomia
21. CatFormatoTipo
22. CatGenero
32. CatTipodeEstratos
33. CatTipoIncendio
34. CatTipoVegetacionInegiGeneral
35. CatTrozaTipo
36. CatUsoActualCA
37. CatUsoEspecie
38. CatUsoLocalReg
39. CatUsoSuelo
40. CatUsoSueloSinCubiertaVegetal
41. CatVegetacionSecundaria
42. CatVigorArboladoBosqueYSelvaArbolEtapa
43. CatVigorRepobladoBosqueYSelva

En cumplimiento con el esquema relacional bajo el cual está estructurada la base de datos se puede mencionar la existencia de llaves primarias utilizadas como identificadores de cada una de las entidades o tablas pertenecientes a la base de datos. Todas las tablas de la base de datos se relacionan entre sí a través de claves primarias y claves foráneas. En la Figura 22 se muestra el esquema de la base de datos, las entidades que la conforman, así como sus relaciones y cardinalidades y en el anexo 3 se muestra el mismo diagrama dividido en tres secciones para ser analizado con mayor detalle.

8.3 Determinación del nivel relacional de la base de datos del INFyS 2004-2009

En la presente sección se revisa las 12 reglas propuesta por Codd en 1984 que sirven como normas que deben cumplir las bases de datos para ser consideradas como relacionales; además se realiza un análisis de cada una de las reglas y se selecciona aquellas involucradas con el diseño de bases de datos, para aplicarlas a la base de datos del Inventario Nacional Forestal y de Suelos 2004-2009 y obtener una medida de su nivel relacional.

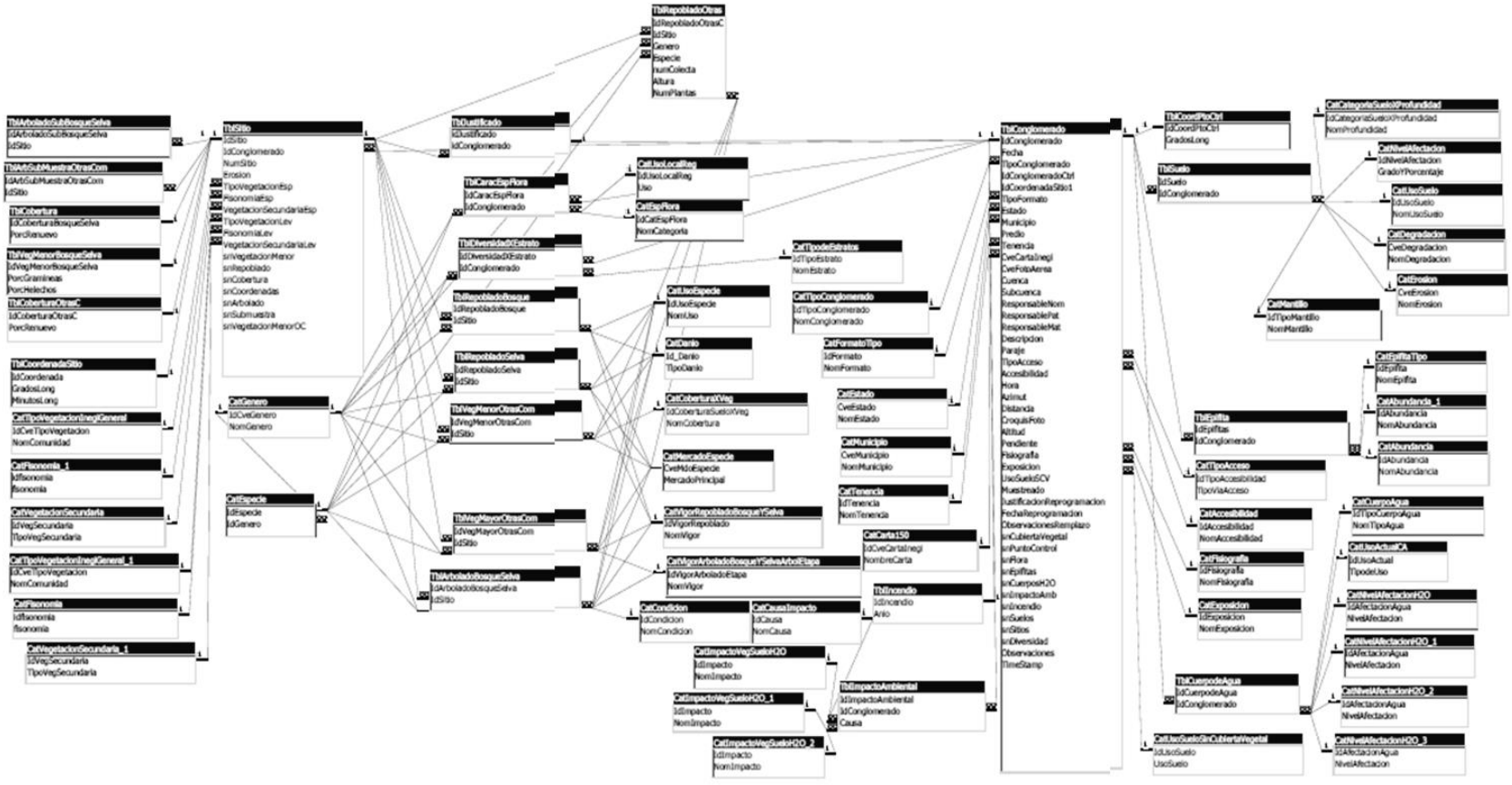


Figura 22. Diagrama reconstruido de la base de datos del INFyS 2004-2009
Fuente: Elaboración propia a partir de la base de datos del INFyS 2004-2009

El “Modelo Relacional” fue publicado por Edgar Frank Codd, en 1970, donde se proponía estructuras de datos simples y lenguajes sencillos para los usuarios, Departamento de Lenguajes y sistemas Informáticos (2004). Este modelo se ha propagado con gran éxito desde su aparición, de tal manera que para los años 80 existían numerosos sistemas que decían estar basados en un modelo relacional pero que carecían de algunas características fundamentales en su diseño. Con la finalidad de diferenciarla de otras bases de datos, Codd en 1985 postuló una serie de reglas, en los artículos E.F. “Is Your DBMS Really Relational?” y “Does Your DBMS Run by the Rules?”, publicadas por la revista *Computerworld*, en octubre 14 y 21 de 1985, respectivamente, (samples.jbpub.com/9781449606008/Chap04Codds TwelveRules.docx). Se considera importante mencionar que el uso de los manejadores de bases de datos actuales, diseñados para la gestión de bases de datos relacionales, implementan una serie de restricciones que garantizan el cumplimiento de la mayoría de las reglas aplicables. El uso de Microsoft Access y SQL Server como sistemas gestores para la base de datos del INFyS garantizan un cierto nivel relacional, pero lo esfuerzos se ven rebasados al momento de diseñar la base datos y garantizar la integridad de la información.

La revisión de la metodología empleada para el manejo y gestión de la base datos del inventario, no proporciona información sobre alguna diferencia, entre el modelo y los diseños de las bases de datos del inventario, gestionados por Microsoft Access y SQL Server, por lo cual, se hace la suposición que éstos son iguales en ambos sistemas gestores.

Las 12 reglas propuestas por Codd para bases de datos relacionales se mencionan en las siguientes subsecciones, donde también se proporciona una breve descripción del significado de cada regla y su correspondiente aplicación en el diseño de bases de datos.

Regla No. 1 – Regla de la información

“Toda la información es presentada explícitamente en el nivel lógico, exactamente de una sola manera, en forma de valores en una tabla”.

Toda la información, incluyendo nombres de tablas, nombres de vistas, nombres de columnas, y los datos de las columnas deben estar almacenados en tablas dentro de las bases de datos. Las tablas que contienen tal información constituyen el Diccionario de Datos. Por tanto los metadatos (diccionario, catálogo) se representan exactamente igual que los datos de usuario y se puede usarse el mismo lenguaje, por ejemplo SQL, para acceder a los datos y a los metadatos (regla 4).

Los resultados de la búsqueda realizada dentro de la base de datos del INFyS se listan a continuación.

- La información correspondiente a las tablas, nombre y descripción, se muestra dentro de la tabla **TABLAS**.
- La llave primaria de cada tabla no se especifica dentro de la tabla correspondiente.
- La información de todas las tablas que forma la base de datos, no se proporciona de manera completa.
- La información correspondiente a los catálogos, nombre del catálogo, llave primaria y descripción, se muestran en la tabla **CATALOGOS**.
- La CONAFOR (2012) reporta 43 catálogos y sólo se presenta información de 42 catálogos.
- La información relacionada con los campos de cada tabla, no se proporciona detalladamente, en la tabla correspondiente, lo que dificulta saber el dominio contenido de los datos almacenados.

En la Figura 23 se muestra la informacion de la tablas de la base de datos del INFyS 2004-2009.

TABLAS

CODIGO	NOMBRE_TABLA	DES	ID
1	TblSitio	Información general de los sitios muestreados en el conglomerado	
2	TblConglomerado	Atributos generales del conglomerado o unidad de muestreo primaria (UMP).	
3	TblArboladoBosqueSelva	Aspectos dasométricos del arbolado de bosques y selvas	
4	TblArboladoSubBosqueSelva	Información de la submuestra del arbolado de bosques y selvas	
5	TblArbSubMuestraOtrasCom	Información de la submuestra de la vegetación mayor de las comunidades áridas y semi	

Figura 23. Información de las tablas de la base de datos del INFyS
Fuente: Elaboración propia para la investigación a partir de la base de datos del Inventario Nacional Forestal 2004-2009.

En la Figura 24 se muestra la información de los catálogos presentes en la base de datos del Inventario Nacional Forestal y de Suelos 2004-2009.

CATALOGOS

Id	NOMBRE	CLAVE	DES	Haga clic para agregar
1	CatAbundancia	IdAbundancia	NomAbundancia	
2	CatAccesibilidad	IdAccesibilidad	NomAccesibilidad	
3	CatAniosIncendios	IdAnioIncendios	Anio	
4	CatCarta150	IdCveCartaInegi	NombreCarta	
5	CatCategoriaSueloXProfundidad	IdCategoriaSueloXProfundidad	NomProfundidad	
6	CatCausaImpacto	IdCausa	NomCausa	
7	CatCoberturaXVeg	IdCoberturaSueloXVeg	NomCobertura	
8	CatCondicion	IdCondicion	NomCondicion	
9	CatCuerpoAgua	IdTipoCuerpoAgua	NomTipoAgua	

Figura 24. Información de los Catálogos de la base de datos del inventario.
Fuente: Elaboración propia para la investigación a partir de la base de datos del Inventario Nacional Forestal 2004-2009.

El resultado del análisis realizado arroja que los metadatos son suficientes para conocer a detalle la información contenida en la base de datos del Inventario Nacional Forestal y de Suelos 2004-2009, por lo cual se considera que no cumple con lo especificado en la presente regla.

Regla No. 2 – Regla del Acceso garantizado

"Cada uno de los datos debe ser lógicamente accesible al ejecutar una búsqueda que combine el nombre de la tabla, su clave primaria, y el nombre de la columna".

La aplicación de esta regla consiste en que dado el nombre de una tabla, el valor de la clave primaria y el nombre de la columna requerida, se debe encontrar uno y solamente un valor. Por esta razón la definición de claves primarias para todas las tablas es prácticamente obligatoria. En la base de datos del INFyS, todas las tablas tienen un nombre único, al igual que sus campos y todas cuentan con una llave primaria, por tal motivo se considera que esta regla se cumple.

Regla No. 3 – Tratamiento sistemático de valores nulos

"El sistema debe ser capaz de representar valores nulos en forma sistemática, independientemente del tipo de datos (dominio). Los valores nulos deberán ser distinto de cero o cualquier otro número o cadenas vacías o en blanco".

Está regla se enfoca principalmente en Sistema Gestor de Bases de Datos, el cual debe ser capaz de soportar el uso de valores nulos para aquello valores que sean desconocidos. Ésta regla no aplica en el diseño de la base de datos.

Regla No. 4 - Descripción de la base de datos

"La descripción de la base de datos es representada en el nivel lógico de la misma manera que los datos ordinarios (en tablas y columnas) que debe ser accesible a los usuarios autorizados".

La información contenida en tablas y vistas, debe ser almacenada en tablas para que sean accesibles de la misma forma que las tablas con datos ordinarios, por ejemplo un comando SQL. Usando la información de las Figuras 23 y 24 se puede observar que las tablas **TABLAS** y **CATALOGOS** cuentan con una llave primaria y nombres de columnas únicos, por lo cual se puede acceder a esta información aplicando la regla de la información. La base de datos del Inventario Nacional Forestal cumple con la presente regla.

Regla No. 5 - La regla del sub-lenguaje Integral

"Debe haber al menos un lenguaje que sea integral para soportar la definición de datos, manipulación de datos, definición de vistas, manipulación de datos (interactivo y por programación) restricciones de integridad, y control de autorizaciones y transacciones (begin, commit y rollback)".

Esto significa que debe haber por lo menos un lenguaje con una sintaxis bien definida que pueda ser usado para administrar completamente la base de datos. Esta regla no es aplicable al diseño de las bases de datos relacionales

Regla No. 6 - La regla de la actualización de vistas

"Todas las vistas que son teóricamente actualizables, deben ser actualizables por el sistema mismo".

La mayoría de las Sistemas Gestores de Bases de Dato Relacionales permiten actualizar vistas simples, pero deshabilitan los intentos de actualizar vistas complejas. Esta regla no es aplicable al diseño de las bases de datos relacionales

Regla No. 7 - La regla de inserción, actualización y borrado

"La capacidad de manejar una base de datos con operadores simples aplica no sólo para la recuperación o consulta de datos, sino también para la inserción, actualización y borrado de datos".

El lenguaje SQL, son soportados por Microsoft Access y SQL Server, lo que permite llevar a cabo operaciones que no regresan valores de una tabla, esta operaciones se realizan con las cláusulas SELECT, UPDATE, DELETE e INSERT (leer, actualizar, eliminar y agregar registros, respectivamente). Esta regla no es aplicable al diseño de las bases de datos relacionales.

Regla No. 8 - Independencia física

"El acceso de usuarios a la base de datos a través de terminales o programas de aplicación, debe permanecer consistente de manera lógica aun cuando se realicen cambios en los datos almacenados, o en el método de acceso a los datos".

El comportamiento de los programas de aplicación y de la actividad de usuarios vía terminales debería ser predecible basados en la definición lógica de la base de datos, y éste comportamiento debería permanecer inalterado, independientemente de los cambios en la definición física de ésta. Esta regla no es aplicable al diseño de las bases de datos relacionales.

Regla No. 9 - Independencia lógica

"Los programas de aplicación y las actividades de acceso por terminal deben permanecer lógicamente inalteradas cuando quiera que se hagan cambios (según los permisos asignados) en las tablas de la base de datos".

La independencia lógica de los datos especifica que los programas de aplicación y las actividades de terminal deben ser independientes de la estructura lógica, por lo tanto los cambios en la estructura lógica no deben alterar o modificar estos programas de aplicación. Esta regla no es aplicable al diseño de las bases de datos relacionales.

Regla No. 10 - Independencia de la integridad

"Todas las restricciones de integridad deben ser definibles en los datos, y almacenables en el catálogo, no en el programa de aplicación"

Las reglas de integridad

1. Ningún componente de una clave primaria puede tener valores en blanco o nulos (ésta es la norma básica de integridad).
2. Para cada valor de clave foránea deberá existir un valor de clave primaria concordante. La combinación de estas reglas aseguran que haya integridad referencial.

La revisión detallada de cada una de las tablas de la base de datos del inventario, permite concluir que ninguna clave primaria tiene valores en blancos o nulos. De igual forma para clave foránea existe se encuentra relacionado a un valor de clave primaria, por lo cual se garantiza la correcta aplicación de la presente regla.

Regla No. 11 - La regla de la distribución

"El sistema debe poseer un lenguaje de datos que pueda soportar que la base de datos esté distribuida físicamente en distintos lugares sin que esto afecte o altere a los programas de aplicación".

El soporte para bases de datos distribuidas significa que una colección arbitraria de relaciones, bases de datos corriendo en una mezcla de distintas máquinas y distintos sistemas operativos y que esté conectada por una variedad de redes, pueda funcionar como si estuviera disponible como en una única base de datos en una sola máquina. Esta regla no es aplicable al diseño de las bases de datos relacionales.

Regla No. 12 - Regla de la no-subversión

"Si el sistema tiene lenguajes de bajo nivel, estos lenguajes de ninguna manera pueden ser usados para violar la integridad de las reglas y restricciones expresadas en un lenguaje de alto nivel (como SQL)".

Algunos productos solamente construyen una interfaz relacional para sus bases de datos, aunque no sean completamente relacionales, lo que hace posible la subversión (violación) de las restricciones de integridad. Esta regla no es aplicable al diseño de las bases de datos relacionales.

Determinación del nivel relacional

El análisis por separado a cada una de las reglas de Codd, permitió determinar que sólo las reglas 1, 2, 4 y 10 están relacionadas con el diseño de las bases de datos. En la Figura 25 se muestra el proceso de validación las 4 reglas de Codd que fue aplicado a la base de datos del Inventario Nacional Forestal y de Suelos 2004-2009.

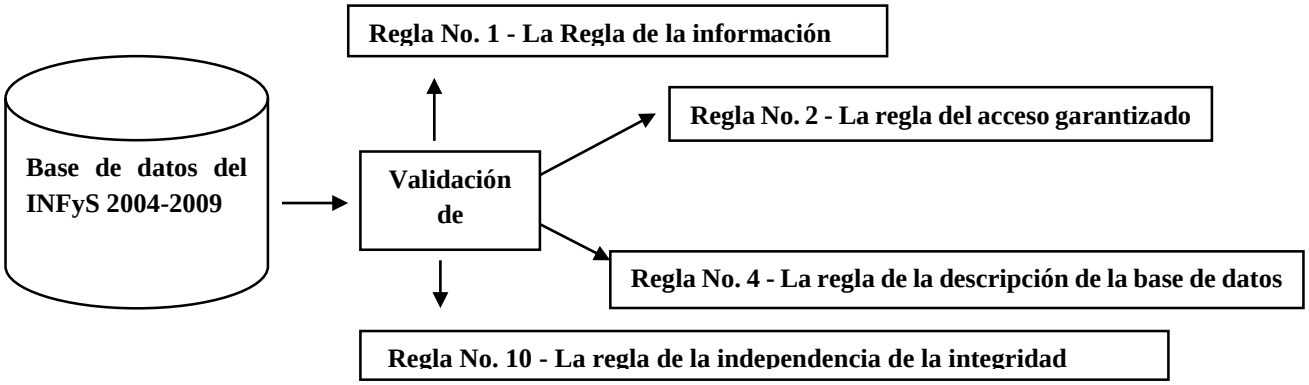


Figura 25: Proceso de validación de las reglas de Codd.
Elaboración: Propia para la investigación.

En el Cuadro 8 se registraron los resultados del análisis de la aplicación de las reglas anteriormente mencionadas, con el objetivo de obtener un cociente para medir el grado relacional de la base de datos. Como sólo 4 de las 12 reglas están relacionadas con el diseño de la base de datos, el procedimiento para medir el nivel relacional de la base de datos resulta del cociente del número de reglas que se cumplen entre el número de reglas que se validaron, como se muestra en la ecuación 13.

Cuadro 8. Registro de información de cumplimiento de las Reglas de Codd.

Regla	Cumple	No cumple
Regla No. 1 - La Regla de la información		X
Regla No. 2 - La regla del acceso garantizado	X	
Regla No. 4 - La regla de la descripción de la base de datos	X	
Regla No. 10 - La regla de la independencia de la integridad	X	

Fuente: Elaboración propia para la investigación

$$Nivel\ relacional = \frac{Reglas\ que\ se\ cumplen}{Total\ de\ reglas\ validadas} = \frac{3}{4} = 0.75 \qquad (13)$$

El resultado obtenido en este sencillo análisis fue que el nivel relacional de la base de datos del inventario es un 75% relacional en lo que respecta al diseño y modelo. Con éste resultado no se puede considerar a ésta base de datos como completamente relacional, sino, que se encuentra en una categoría intermedia, entre relacional y no relacional, por lo que se le puede considerar como semi-relacional.

8.4 Resumen

En éste capítulo se llevaron a cabo una serie de análisis a la base de datos del INFyS 2004-2009 con la finalidad de entender la información que allí se almacena y la forma en que accede a ésta. La primera etapa del análisis se enfocó en dar a conocer la información que se registra y almacena en la base de datos; así como describir las diferentes actividades que la originan. En la segunda etapa se reconstruyó el modelo conceptual de la base de datos con la ayuda de las herramientas de Microsoft Access. Finalmente, en la tercera etapa se explicaron y analizaron las reglas de Codd, seleccionándose aquellas que se refieren al diseño de base de datos relacionales, para posteriormente verificar que la base de datos del INFyS 2004-2009 se apegara a esas normas. Los resultados más importantes fueron la reconstrucción del diorama de la base de datos y la determinación de su nivel relacional el cual se catalogó como semi-relacional en función del número de reglas que satisfizo.

9. ALGORITMOS DE MINERÍA DE DATOS PARA LA CLASIFICACIÓN DE ÁRBOLES POR GÉNERO

En el capítulo 5 se mencionó que la información del inventario se utiliza para la generación de reportes nacionales e internacionales, así como investigaciones forestales. En ese mismo capítulo se dieron a conocer algunas recomendaciones dadas por quienes hacen uso de la información del inventario para generar estos reportes e investigaciones y de los cuales se rescató el tema de garantizar la integridad de los datos.

Las bases de datos del Inventario Nacional Forestal representan una importante fuente de información que puede ser exitosamente explotada, con la minería de datos, para extraer información novedosa o para resolver algún problema específico relacionado con el análisis de información de la misma.

Uno de los principales inconvenientes que se presenta desde la integración de los datos, hasta el análisis de los mismos, se relaciona con la calidad de la información que se puede ver afectada por información incompleta, inconsistente o quizá redundante, debido a errores de captura o errores técnicos en la medición. De acuerdo con CONAFOR (2012), durante la etapa de integración se realiza una inspección minuciosa de la información levantada en campo, antes de ser integrada a la base de datos del servidor central. Esta inspección consiste en cotejar el contenido del formato en papel contra el digital con la finalidad de detectar inconstancias entre ambos. Aquellos marcados como inconsistentes son devueltos para su revisión, corrección y nueva entrega; los que no, son almacenados dentro del servidor central. En esta etapa sólo se verifica que no existan errores de captura de información pero no se detectan datos incongruentes como pueden ser datos atípicos o información incorrecta.

En los reportes e investigaciones se eliminan partes de las inconsistencias de los datos mediante filtros que consideran un rango de altura o diámetro normal, esto deja fuera a los datos atípicos pero no elimina la posible información incorrecta. Sin duda alguna encontrar inconsistencias en el mundo de datos que se genera en el inventario forestal representa un reto que no es fácil de realizar. Este tipo de problemas pueden ser atacados con ayuda de la minería de datos.

Para probar lo anterior se llevarán a cabo comparaciones de diferentes algoritmos de minería de datos con el objetivo de seleccionar uno o más modelos que mejor clasifiquen al género arbóreo *Quercus*, a nivel nacional. El proceso a seguir se desarrollará de acuerdo con la Figura 26.

Figura 26: Proceso para la aplicación de modelos de minería de datos.
Fuente: Elaboración propia para la investigación

En las siguientes secciones se describe a detalle cada una de las etapas del proceso de selección del modelo de minería de datos para la clasificación del género arbóreo *Quercus*, usando los datos del Inventario Nacional Forestal y de Suelos 2004-2009.

9.1 Selección del género arbóreo de estudio

La selección del género arbóreo de estudio considera aspectos como la cantidad de información disponible y la distribución de este a nivel nacional.

En primer lugar se tomó una muestra a nivel nacional, de todos los géneros arbóreos que cumplieron con las restricciones establecidas por la CONAFOR en su procedimiento de

estimación; diámetro normal entre 7.5 centímetros y 132.5 centímetros y altura total entre 5 metros y 47.5 metros. La tabla **TblArboladoBosqueSelva**, donde se registra la información del arbolado, contiene 1, 267, 542 registros de los cuales al aplicar el filtro para eliminar los datos atípicos se reduce a 979, 139 registros, es decir sólo se considera el 77% del total de individuos muestreados. En la Figura 27 se aprecia los porcentajes correspondientes a los registros considerados en la muestra después de aplicar el filtro y la cantidad de datos que se desechan.

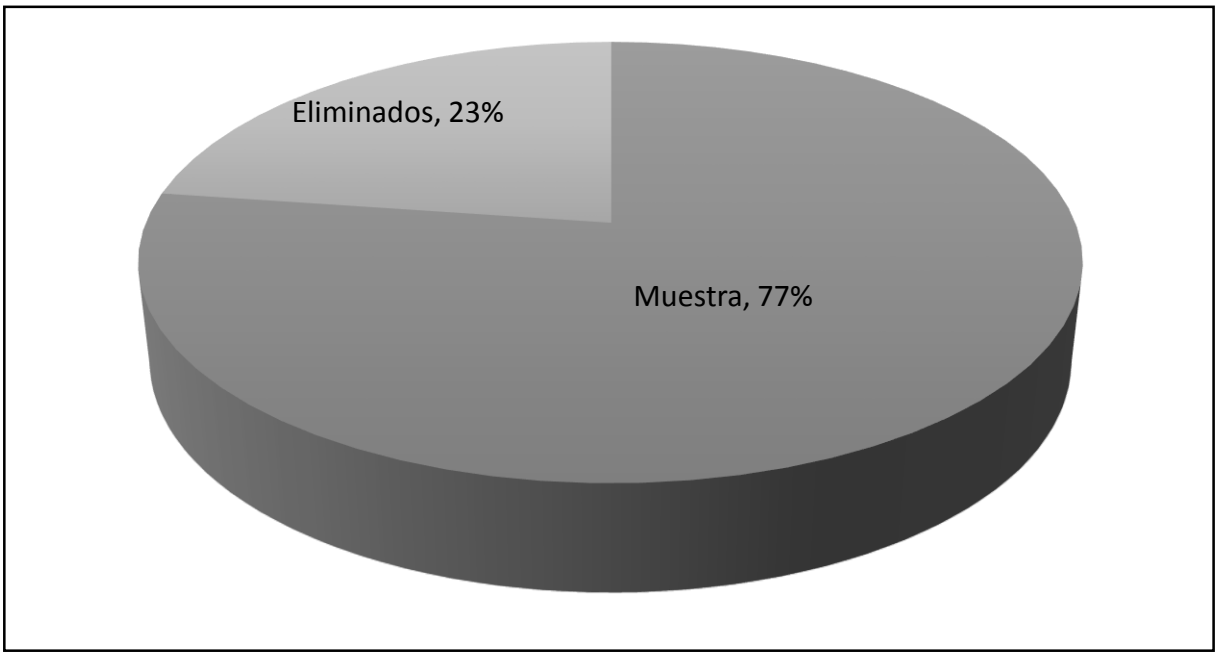


Figura 27. Porcentaje de registros considerados al aplicar el filtro para eliminar datos atípicos.
Fuente: Elaboración propia para la investigación a partir del INFyS 2004-2009.

En segundo lugar se contabilizaron los individuos por género, sin tomar en cuenta su distribución por estados. Los 5 géneros más importantes, de acuerdo al número de individuos considerados en la muestra se representan en el Cuadro 9.

Cuadro 9. Distribución de individuos por género considerados en la muestra.

Genero	Número de Individuos
Quercus	205511
Pinus	133457
Bursera	54344
Lysiloma	45606
Piscidia	27497
Otros	512724

Fuente: Elaboración propia para la investigación a partir del INFyS 2004-2009.

Quercus es el género arbóreo con mayor cantidad de información recopilada a nivel nacional, en el INFyS 2004-2009, con 205, 511 individuos (21%), seguido por *Pinus* con 133, 457 individuos (14%) y *Bursera* con 54, 344 individuos (5%). En total estos tres géneros suman el 40% del total de datos útiles del arbolado. En la Figura 28 se muestran los porcentajes de información de los cinco principales géneros, de acuerdo a la cantidad de información recopilada en el inventario forestal.

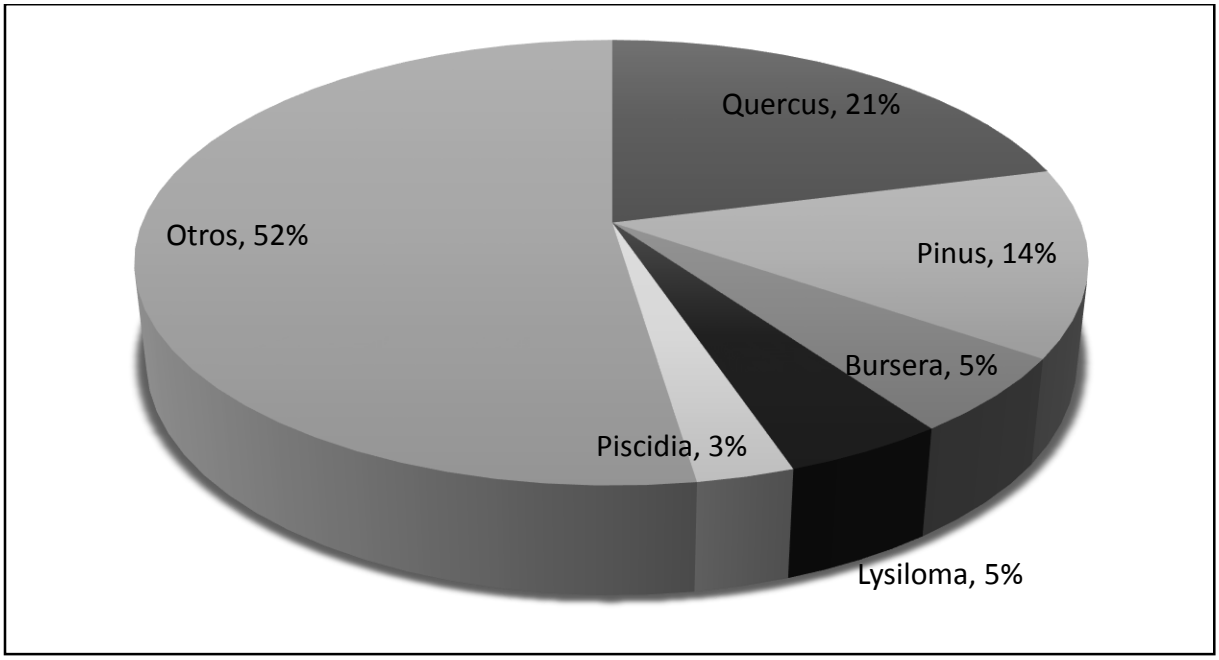


Figura 28: Porcentaje de información útil por género en el INFyS 2004-2009
Fuente: Elaboración propia para la investigación a partir del INFyS 2004-2009

En la distribución espacial dentro del territorio mexicano de los tres principales géneros con mayor información podemos encontrar que *Quercus* y *Pinus* se encuentran en el 87% de las entidades federativas de la república mexicana, comparten los mismos estados, siendo los pertenecientes a la península de Yucatán (Campeche, Quintana Roo y Yucatán) los únicos estados sin información de estos géneros; lo que respecta a *Bursera*, se encontró que se distribuye en el 84% de las entidades. Considerando la información anterior se selecciona al género *Quercus* para obtener modelos que ayuden en su identificación y que puedan ser utilizados durante el proceso de validación de datos o como criterio para llevar a cabo las inspecciones de control para garantizar la calidad de la información. La distribución por estados para el género *Quercus*, se muestra en el Cuadro 10.

Cuadro 10. Distribución por estado del género *Quercus* (después de eliminar datos atípicos).

Estado	Número de individuos
Aguascalientes	690
Baja California	43
Baja California Sur	257
Chiapas	6707
Chihuahua	41083
Coahuila	1680
Colima	417
Distrito Federal	71
Durango	27744
Guanajuato	4041
Guerrero	10473
Hidalgo	2921
Jalisco	21219
México	4683
Michoacán	9407
Morelos	317
Nayarit	8266
Nuevo León	3920
Oaxaca	27238
Puebla	1556
Querétaro	1459
San Luis Potosí	5121
Sinaloa	4555
Sonora	8797
Tabasco	17
Tamaulipas	4940
Tlaxcala	342
Veracruz	1246
Zacatecas	6301

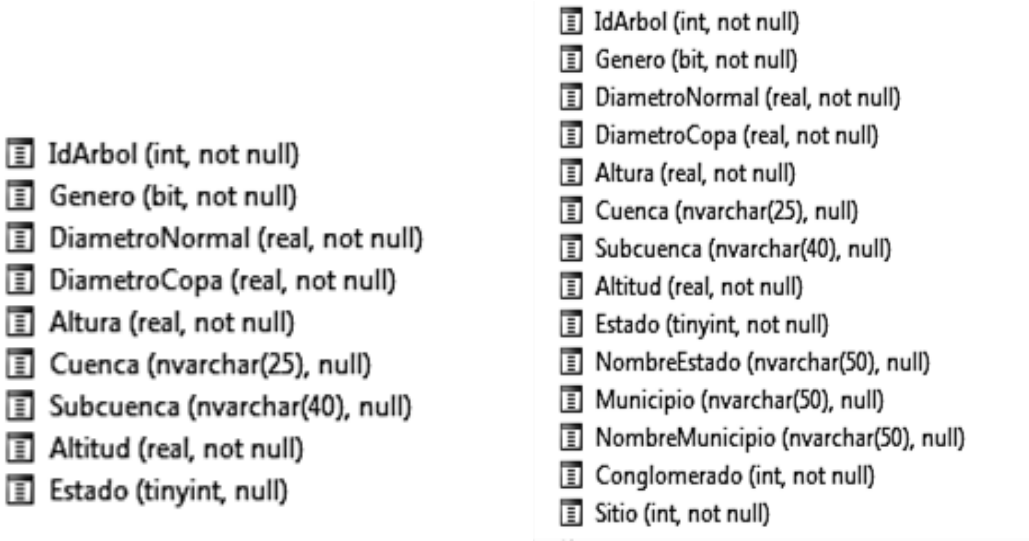
Fuente: Elaboración propia para la investigación a partir del INFyS 2004-2009.

9.2 Selección de las variables de entradas para los modelos de minería

La selección de variables a utilizar en los modelos de minería de datos es de suma importancia ya que de ellas dependerá el éxito que tenga el modelo. Dicha selección de variables tomó como punto de partida el trabajo de Méndez y de los Santos (2011) donde se realizó una depuración de la base de datos del inventario forestal a partir de gráficas por género y el análisis de sus relaciones dasométricas básicas como la relación diámetro normal-altura total, diámetro normal-diámetro de tocón, área basal total-volumen total, entre otras a fin de detectar y corregir errores presentes en la base de datos. Algunos de estos criterios como, área basal total-volumen total,

necesita que la información sea procesada para encontrar esta relación lo que no permite que sea un método práctico al momento de realizar una validación rápida. De este método se pueden rescatar las variables diámetro normal y altura total que también son utilizados en algunas ecuaciones para estimar el volumen y que se encuentran en la misma tabla donde se registra la especie arbórea; adicionalmente se agregará el diámetro de copa y por supuesto la variable respuesta que sería el género.

En la base de datos del inventario se proporciona información adicional en la tabla **TblConglomerado**, que tiene que ver con las condiciones físicas del terreno, y que puede someterse a pruebas para medir la posibilidad de usarse como criterios en la validación. Después de realizar un análisis de la información almacenada la tabla mencionada, se seleccionó como variables de entrada a la altitud donde se levantaron los datos. Considerando la variable anterior se tiene un total de 4 variables de entrada, de las cuales sólo altitud es del tipo discreta y las restantes del tipo continua; es importante mencionar que en minería de datos no importa la cantidad de variables que se utilicen ya que los mismos métodos utilizados se encargaran de no utilizar las que menos favorecen a los modelos.



a) Campos de la tabla de entrenamiento b) Campos de la tabla de predicción

Figura 29: Definición de los campo de las tablas de entrenamiento y de predicción.
Fuente: Elaboración propia para la investigación.

La definición de las tablas que serán usadas para el entrenamiento de los modelos, así como para guardar las predicciones debe realizarse anteriormente al proceso de minería; estas tablas pueden estar en la misma base de datos del inventario, sin embargo es recomendable crear una nueva base de datos sólo para la validación de datos, con la finalidad de tener un mayor control organizacional de la información; no se olvide que la base de datos del inventario cuenta con 63 tablas y agregarle otra más complicaría algunas operaciones. Para este ejemplo se creó en SQL Server una nueva base de datos llamada **validación** que contiene las tablas **Quercus** y **ValQuercus**, tabla de entrenamiento (Figura 29, inciso a) y de predicción (Figura 29, inciso b), respectivamente. Como nota importante, se debe garantizar que ambas tablas contengan, al menos, los campos de las variables utilizadas por los modelos. En la Figura 29 se muestran las definiciones de los campos que contienen cada una de las tablas mencionadas.

9.3 Preparación de los datos para minería

Un proceso de minería no consiste sólo en saber utilizar las herramientas de análisis, sino que se necesita conocer a detalle la información que se va a procesar, analizar las posibles transformaciones de las variables y tipos de datos antes y durante el proceso y finalmente conocer cómo trabaja cada una de las técnicas empleadas.

Las características de cada árbol muestreado para el inventario forestal se encuentran en la tabla **TblArboladoBosqueSelva** y la correspondiente al medio donde se desarrollan, se encuentra en la tabla **TblConglomerado**. Para la generación de los modelos se utilizó la información directamente de la base de datos del INFyS 2004-2009; las variables seleccionadas, que caracterizan físicamente al arbolado, como el diámetro normal, el diámetro de copa y la altura total, son del tipo continuas, mientras que la que caracteriza al medio es del tipo discreta.

El género al que pertenece cada árbol también se encuentra registrado en la tabla **TblArboladoBosqueSelva** en forma de variable cualitativa, la forma en que se representa cada uno es mediante una serie de claves que los identifica de manera única. Como existe una gran diversidad de géneros registrados en la base de datos, sólo se utilizará la información de los

cinco géneros más importantes, que aparecen en el Cuadro 8. El comando SQL empleado para extraer la información para el entrenamiento de los modelos se presenta en el Cuadro 11.

Cuadro 11. Comando SQL para extraer los datos de entrenamiento.

```
SELECT IdArboladoBosqueSelva, Genero, DiametroNormal, DiametroCopa,
AlturaTotal, TblConglomerado.Estado, Cuenca, Subcuenca, Altitud

FROM (((TblArboladoBosqueSelva INNER JOIN TblSitio ON
TblArboladoBosqueSelva.IdSitio = TblSitio.IdSitio) INNER JOIN
TblConglomerado ON TblSitio.IdConglomerado =
lConglomerado.IdConglomerado))

WHERE  DiametroNormal >= 7.5 AND DiametroNormal <= 132.5 AND AlturaTotal
>= 5 AND AlturaTotal <= 47.5 AND DiametroCopa is not null and Genero IN
(808, 741, 149, 579, 746)
```

Fuente: Elaboración propia para la investigación a partir del INFyS 2004-2009.

Los modelos de minería que se generen tendrán el objetivo de realizar una clasificación de un árbol, como perteneciente o no al género *Quercus*. Como la muestra contiene información de cinco géneros diferentes, entre ellos *Quercus*, es necesario realizar una transformación al campo género que permita distinguir los datos de esté genero con respecto a los demás, similar al proceso de definición de variables dummy en regresión lineal. En la transformación es necesario sustituir la clave de género por un indicador, de este modo si los datos del árbol pertenecen al género que se quiere clasificar el valor que tomará el campo género será 1 y 0 para otros géneros. Debido a la gran cantidad de información que se desprende de esta consulta (449, 691 registros), el cambio de valor de la variable se lleva a cabo con las herramientas de integración de SQL Server. En la Figura 30 se muestra los objetos contenidos dentro de este paquete.

MacLennan *et al.* (2009) remarcan que los algoritmos de minería pueden ser sensitivos al número de atributos que se incluyan para el análisis y que al aumentar la cantidad de estos, se requieren mayores exigencias de CPU y memoria para procesarlos. Además menciona que no todos los atributos son igualmente importantes en términos de la precisión de la predicción. De acuerdo con Mood (1974) la cantidad de registros obtenidos, aproximadamente medio millón, permite aplicar el teorema central de límite, que enuncia la tendencia asintótica de la media muestral a la distribución normal, lo que hace posible trabajar con una muestra aleatoria del

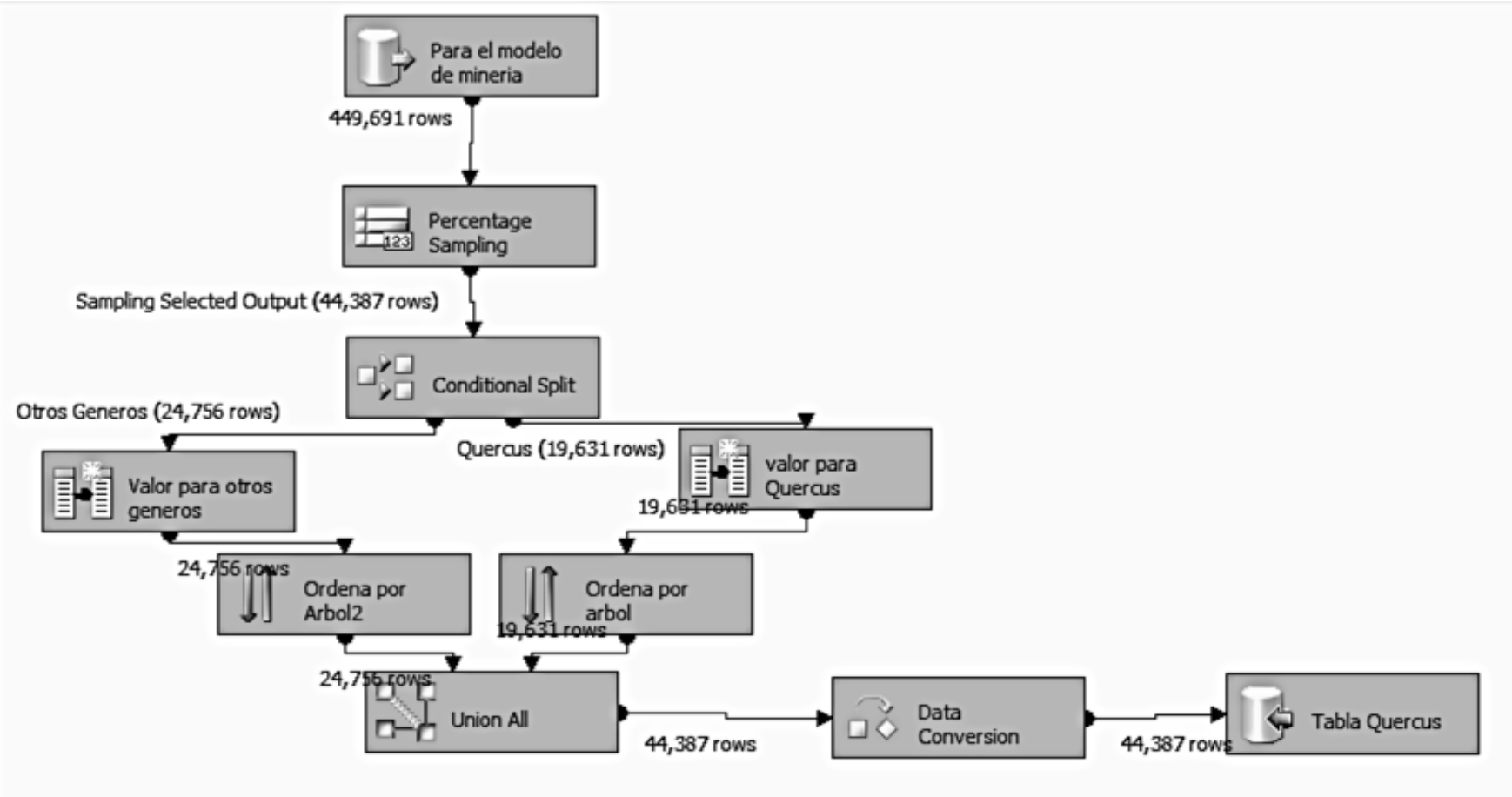


Figura 30: ETL para la tabla de entrenamiento de los modelos de minería de datos.
Elaboración: Propia para la investigación.

10% de los datos seleccionados, con la finalidad de no presentar inconvenientes de hardware al momento de procesar los modelos. En la segunda muestra se obtuvieron 19, 631 registros pertenecían al género Quercus y 24, 756 pertenecían a otros géneros (objeto **percentage sampling**). El objeto **conditional split** realiza la separación de la información en *Quercus* y otros géneros. Los objetos etiquetados como **valor para** realiza los cambio al campo género sustituyendo el valor original correspondiente a la clave de género por el valor de 1 si es *Quercus* y 0 si son otros géneros. Los objetos etiquetados como **ordenar por** ordenan los datos usando la clave original de arbolado. El objeto **Union All** integra la información nuevamente en una tabla. El objeto **Data Conversion** transforma los datos del campo género convirtiéndolos en datos del tipo bite para su correcta integración. Finalmente el objeto **Tabla Quercus** guarda la información en la tabla **Quercus** de la base de datos **validación**.

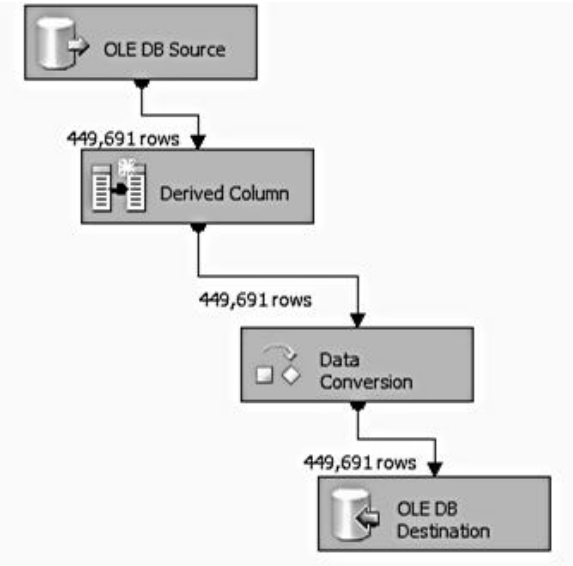


Figura 31: Paquete de integración para la tabla de predicciones.
Fuente: Elaboración propia para la investigación.

Para la integración de los datos en la tabla **ValQuercus** se utilizó otro paquete que contiene los objetos que se muestran en la Figura 31, donde el objeto **OLE DB Source** extrae los datos que van a ser validados, teniendo como fuente la base de datos del inventario. El objeto **Derived Column** asigna un valor de cero a el campo de género ya que este al ser el campo de predicción e necesario que se encuentre definida en la tabla que se usará para realizar las predicciones (El valor de este campo no es considerado al realizar las estimaciones, sólo es un requisito para

llevar a cabo la operación. El objeto **Data Convertion** realiza la transformación de los datos del campo genero al tipo bite. El objeto **DB Destination** guarda la información en la tabla **ValQuercus** de la base de datos Validación.

9.4 Implementación del Servicio de análisis de SQL Server para la generación de modelos de minería de datos

En el servicio de análisis, integrado al entorno de Business Intelligence Develop Studio de SQL Server 2008, se contemplan 7 algoritmos de minería de datos, los cuales son: Árboles de decisión, reglas de asociación, segmentación, regresión lineal, regresión logística, Bayes naïve y redes neuronales. Cada uno de estos tienen sus propias restricciones para modelar un conjunto de datos, la más importante es el tipo de variable, continua o discreta. Los algoritmos de minería como regresión lineal, Bayes naïves y reglas de asociación, trabajan principalmente con variables continuas; el resto de algoritmos permiten una combinación de variables, discretas y continuas.

Las variables seleccionadas para la generación de los modelos de minería son una combinación de variables cuantitativas y cuantitativas, pero la variable a predecir fue transformada de manera dicotómica, por lo cual no se puede utilizar todos los algoritmos proporcionados por SQL Server 2008.

El proceso de generación de modelos de minería de datos con SQL Server 2008 comienza con la definición de una estructura de minería, como se observa del lado izquierdo de la Figura 32, se muestran los modelo de minería definidos, así como la condición de las variables utilizadas (llave, predicción y entrada). Posteriormente, se agregan uno por uno los algoritmos a usar en la fase de entrenamiento del modelo de minería; para el ejemplo que se desarrolla los algoritmos utilizados son: arboles de decisión, análisis clúster, red neuronal y regresión logística. La importancia de la estructura de minería radica en que las variables que usarán los modelos son las mismas definidas en la estructura de minería; la forma en que cada algoritmo las utiliza también es como se define en la estructura.

Structure	ArboldeDesicion	Cluster	RedNeuronal	RegresionLogistica
	Microsoft_Decision_Trees	Microsoft_Clustering	Microsoft_Neural_Network	Microsoft_Logistic_Regr
Altitud	Input	Input	Input	Input
Altura	Input	Input	Input	Input
Diametro Copa	Input	Input	Input	Input
Diametro Normal	Input	Input	Input	Input
Genero	PredictOnly	PredictOnly	PredictOnly	PredictOnly
Id Arbol	Key	Key	Key	Key

Figura 32: Estructura y algoritmos de minería.
Fuente: Elaboración propia para la investigación.

SQL Server, por ser un sistema relacional, hace uso de tablas para representar un conjunto de entidades que a su vez, necesitan de una llave primaria para identificar a cada entidad del conjunto y así evitar la duplicidad; este mismo comportamiento es heredado a la estructura de minería, para el ejemplo el identificador del arbolado funge como llave primaria y sólo es utilizado para tarea específica, mientras que el atributo género es utilizado como la variable de predicción, los demás atributos son utilizados como variables de entrada.

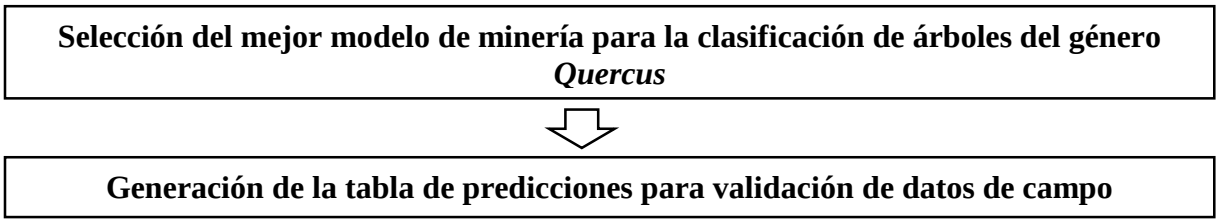


Figura 33. Procedimiento para validar información de campo con respecto al género *Quercus*.
Fuente: Elaboración propia para la investigación.

La Figura 33 muestra las tareas a desarrollarse con la ayuda de SSAS en la obtención de modelos que puedan ser usados para la validación de datos provenientes del muestreo de campo durante la etapa de recolección de datos para los inventarios forestales.

9.4.1 Selección del mejor modelo de minería para la clasificación de árboles del género Quercus

La selección del mejor modelo se realiza con la gráfica de elevación, que se muestra en la Figura 34; en ésta se gráfica en un mismo plano un modelo óptimo que ajusta a los datos y un modelo aleatorio, además de las gráficas generadas por cada uno de los modelos entrenados. De acuerdo

con la gráfica de la Figura 34, el modelo que mejor se ajusta a los datos es el árbol de decisión, con un score de 0.88, seguido por el modelo clúster, la red neuronal y la regresión logística, con el 0.85, 0.82 y 0.82, respectivamente.

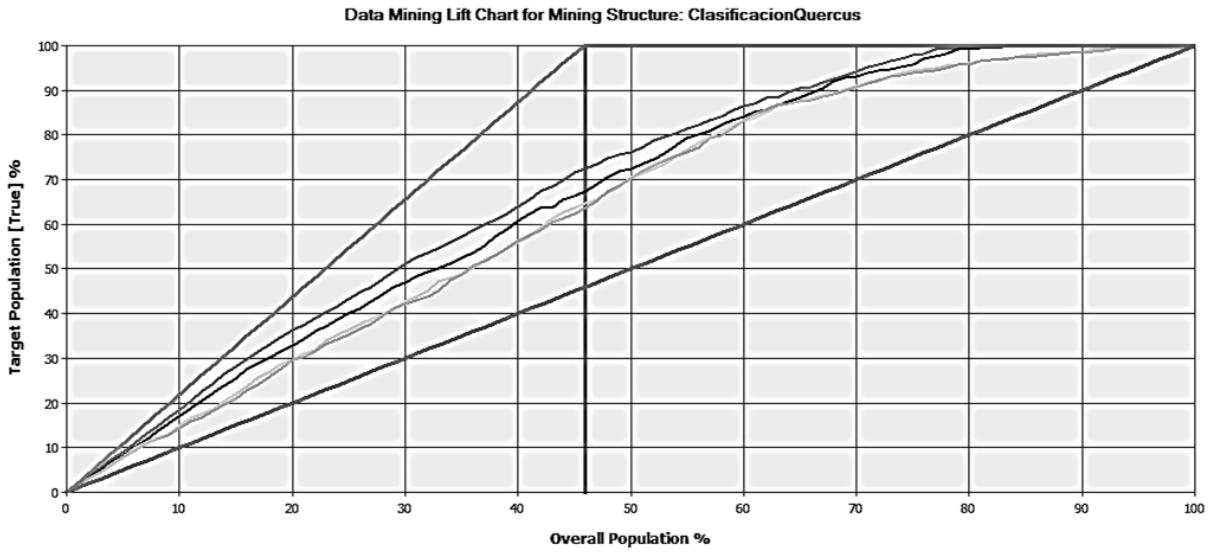


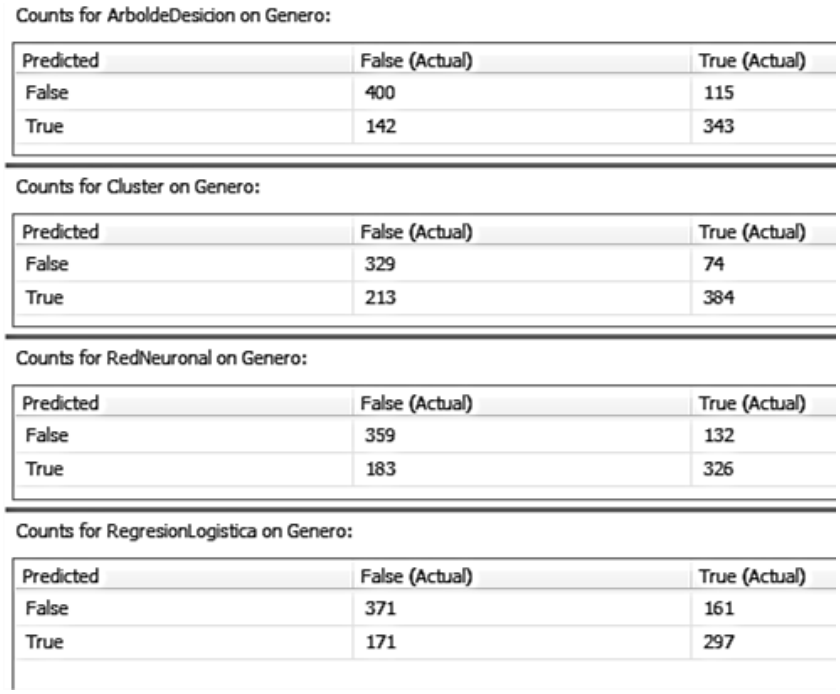
Figura 34: Gráfico de elevación para la comparación de modelos de minería de datos.
Fuente: Elaboración propia para la investigación.

La evaluación del modelo para una tarea de clasificación, como ya se había mencionado anteriormente, se lleva a cabo utilizando la razón de precisión que se obtiene dividiendo el número de clasificaciones correctas por el número total de instancias. Hernández Orallo *et al.*, (2004) aseguran que la precisión es un buen indicador de cómo se comportará el modelo para datos futuros similares a los de prueba, pero que no garantiza que el modelo sea correcto, más bien, indica que si usamos la misma técnica en una base de datos con datos similares a los de prueba, la precisión media será similar a la obtenida con éstos.

La precisión de los modelos se obtiene de la suma el número de individuos clasificados correctamente entre el total de instancias presentes en la prueba, como se ejemplifica en la ecuación (14).

$$Presición = \frac{Número\ de\ Individuos\ clasificados\ correctamente}{Total\ de\ individios\ en\ la\ muestra} \qquad (14)$$

La matriz de clasificación proporcionada por el software (Figura 35) es usada para confirmar la precisión de cada modelo en la clasificación de los árboles del género *Quercus*.



Counts for ArboldeDesicion on Genero:

Predicted	False (Actual)	True (Actual)
False	400	115
True	142	343

Counts for Cluster on Genero:

Predicted	False (Actual)	True (Actual)
False	329	74
True	213	384

Counts for RedNeuronal on Genero:

Predicted	False (Actual)	True (Actual)
False	359	132
True	183	326

Counts for RegresionLogistica on Genero:

Predicted	False (Actual)	True (Actual)
False	371	161
True	171	297

Figura 35: Matriz de clasificación obtenida mediante SSAS.
Fuente: Elaboración propia para la investigación.

El modelo de árbol de decisión clasificó 400 árboles como falsos cuando realmente no pertenecían al género *Quercus* y 343 como verdadero cuando realmente pertenecían a éste género, por lo cual su precisión es

$$Presición = \frac{743}{1000} = 0.74 = 74\% \qquad (15)$$

La ecuación anterior se aplica para cada uno de los modelos con los datos de la matriz de clasificación, obteniéndose los porcentajes que se muestran en el Cuadro 12.

Cuadro 12. Precisión obtenida por los modelos de clasificación.

Clúster	71%
Red Neuronal	69%
Regresión Logística	67%

Fuente: Elaboración propia para la investigación.

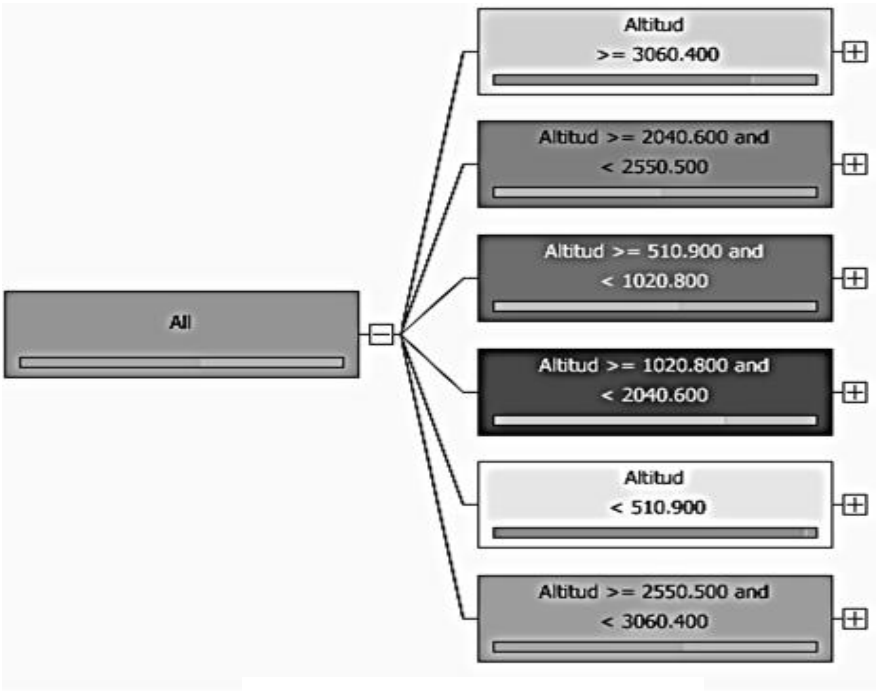
En las siguientes secciones se lleva a cabo un breve análisis de los modelos de minería que mejor se ajustaron a los datos, y una pequeña introducción del algoritmo usado por SQL Server 2008 para los dos modelos restantes; recuérdese que uno de los objetivos de la tesis consiste sólo en encontrar el mejor modelo para la clasificación de árboles del género *Quercus*, para obtener información sobre los métodos de análisis de los demás modelos, se puede recurrir a la bibliografía de la presente tesis.

Análisis mediante el algoritmo de árboles de decisión

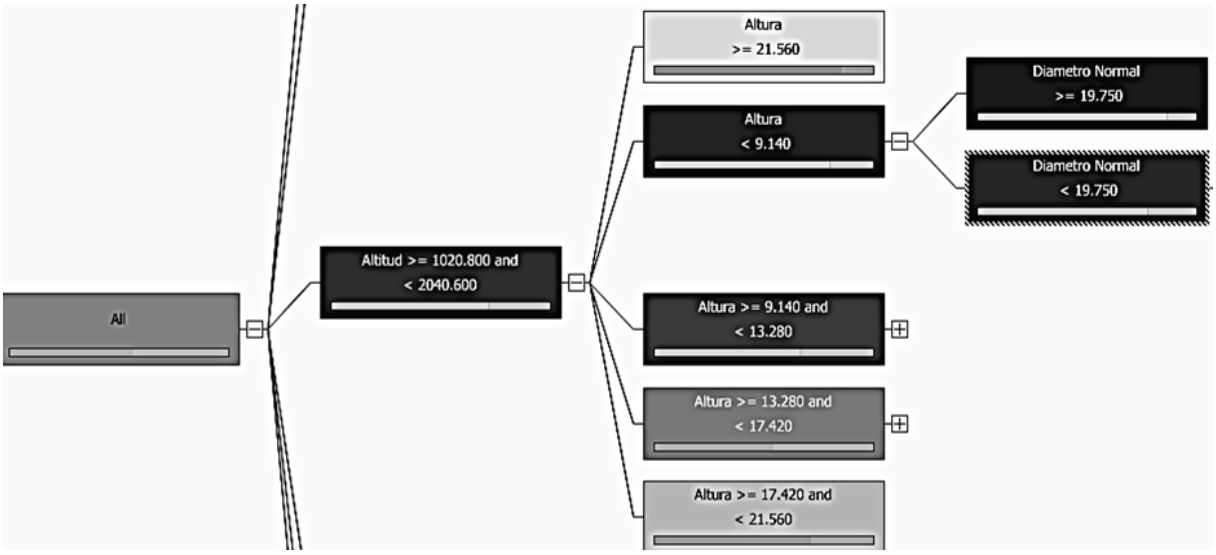
El primer modelo por analizar corresponde al árbol de decisión. El algoritmo que utiliza SQL SERVER 2008, para ésta técnica, es un algoritmo híbrido desarrollado por un grupo de investigación de Microsoft que soporta tareas de clasificación, regresión y asociación (MacLennan *et al.*, 2009).

En la Figura 36 a) se observa el gráfico del árbol de decisión con el nodo raíz y un nivel de análisis. Los tonos más obscuros representan los nodos que describen de mejor manera a los datos, lo que indica que el género Quercus se encuentra distribuido con una mayor probabilidad en altitudes de 1020-2040 metros sobre el nivel del mar. Continuando con el análisis se tiene, en la Figura 36 b), el segundo y tercer nivel del árbol, donde se aprecian las variables altura y diámetro normal, respectivamente. En el segundo nivel, las alturas menores o iguales a 9.14 metros son las que más peso tienen y en el tercer nivel, los diámetros mayores o iguales a 19.75 centímetros.

La información presentada en el párrafo anterior se puede traducir en los árboles que crecen a una altitud entre 1020 y 2040 metros sobre el nivel del mar, que presentan una altura menor o igual a 9.14 metros, y además, tienen un diámetro mayor o igual a 19.75 centímetros, tienen mayor probabilidad de pertenecer al género *Quercus*.



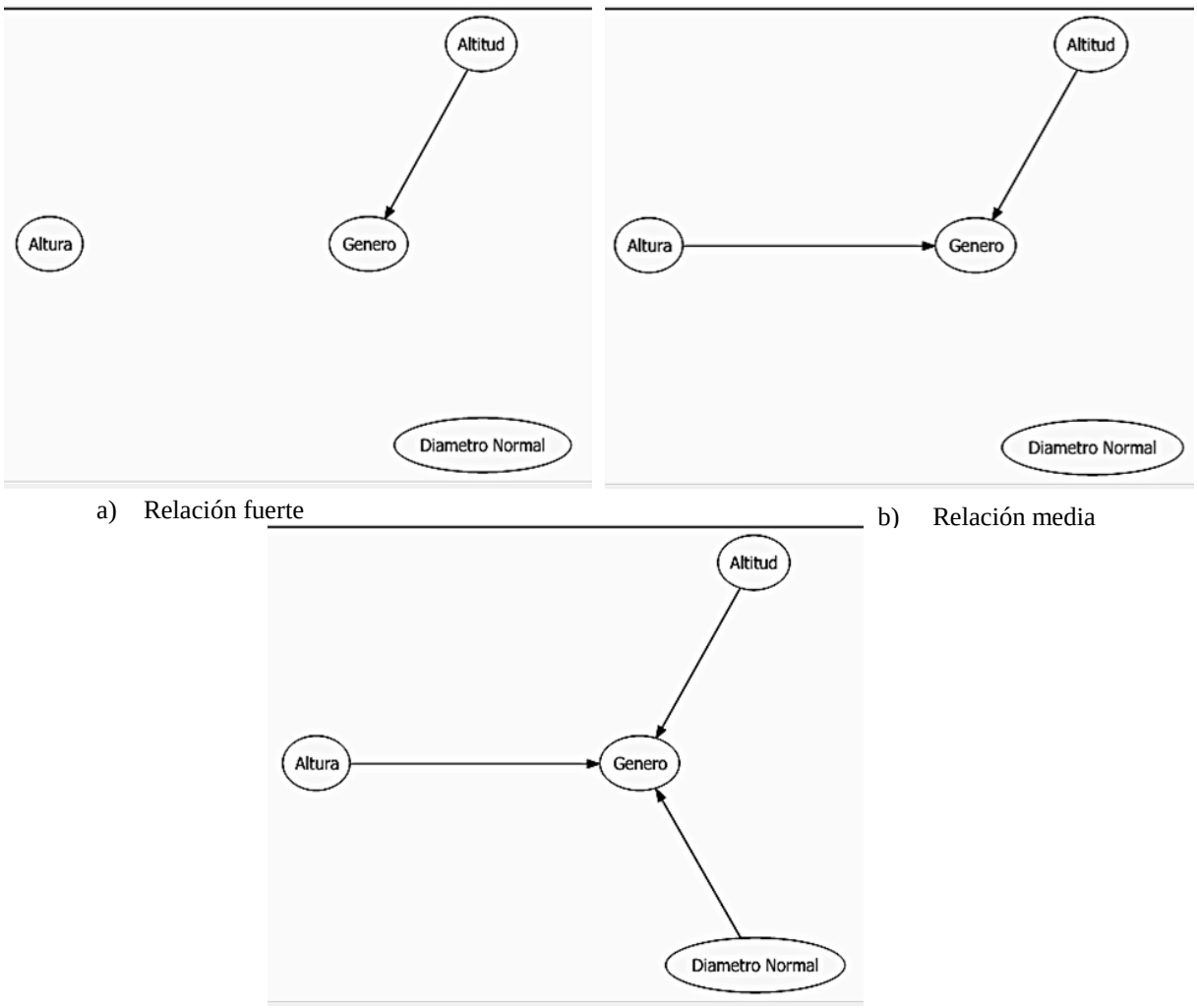
a) Primer nivel del árbol



b) Segundo y tercer nivel del árbol

Figura 36: Gráfico de árbol de decisiones para la clasificación del género Quercus
Fuente: Elaboración propia para la investigación

La detección de las variables que más influyen, también se puede llevar a cabo mediante la gráfica de red de dependencia que se genera al momento de procesar el modelo; la red generada para el modelo que se analiza se muestra en la Figura 37 incisos a), b) y c).



a) Relación fuerte

b) Relación media

c) Relación débil

Figura 37: Diagrama de relaciones del árbol de decisiones para la clasificación del género Quercus
Fuente: Elaboración propia para la investigación

La red de dependencia muestra que la relación más fuerte, entre el modelo y las variables, corresponde a la variable altitud, Figura 37 a), la segunda relación en aparecer es la que se da con la variable altura, Figura 37 b), y la tercera relación, Figura 37 c), se presenta con la variable diámetro normal. Para éste ejemplo, el orden que se sigue al analizar la información con el visor de árbol de decisión, es el mismo orden con que aparecen las variables en la red de dependencia; en muchos casos y cuando se tienen muchas variables, el orden en que aparecen las variables en el árbol y en la red de dependencia, puede variar.

Análisis mediante el algoritmo de clústeres (Clustering)

El servicio de análisis tiene herramientas que permiten visualizar un mismo modelo de diferentes maneras. Una vista, por si sola, no permite un análisis completo de los clústeres pero en conjunto su potencial de análisis es mucho mayor.

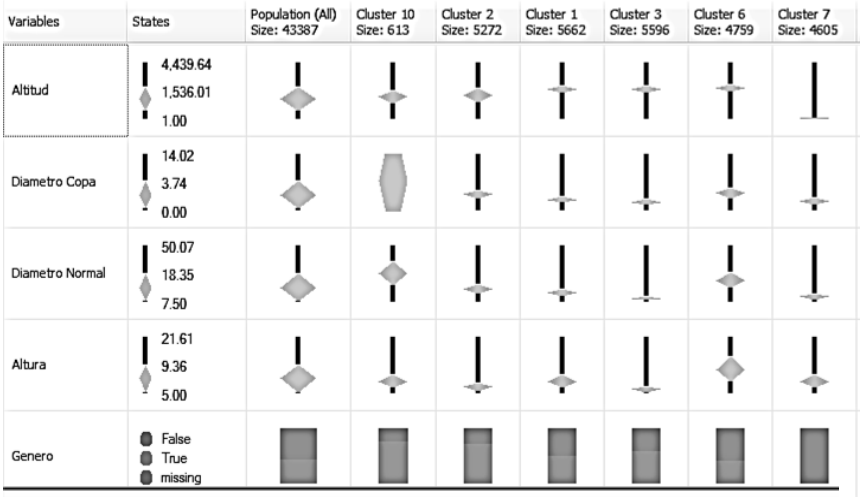
MacLennan *et al.* (2009) proponen los seis pasos básicos siguientes para realizar un análisis detallado de los clústeres:

1. Observar de manera general cada uno de los clústeres.
2. Seleccionar un clúster y determinar sus diferencias contra el resto de la población.
3. Determinar la diferencia del clúster seleccionado con el clúster más cercano.
4. Verificar que los supuestos sean ciertos.
5. Etiquetar el clúster.
6. Repetir los pasos 2 al 5 para todos los clústeres.

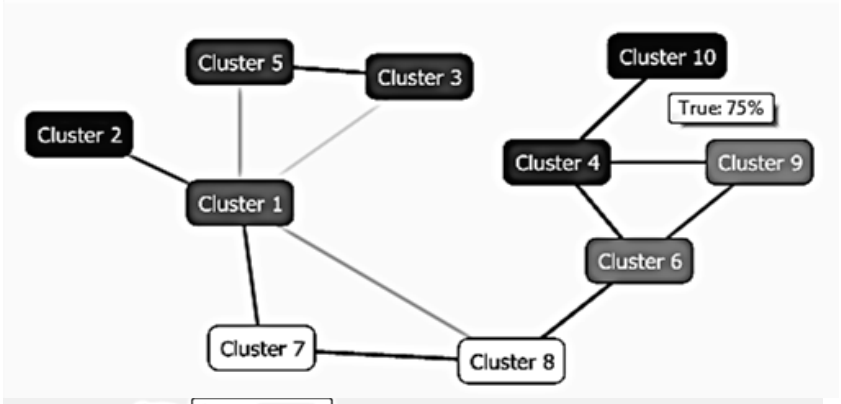
En el análisis que se lleva a cabo, las primeras dos etapas se realizan con ayuda del visor de perfile y el visor de diagrama de clústeres, inciso a) y b) de la Figura 37, respectivamente. En la Figura 37 a) se comparan las características de la población en general y de cada uno de los clústeres; destaca la información del clúster 10 donde se puede observar que la variación en las medidas del diámetro de copa es más alta que la población en general y el diámetro normal también es un poco mayor que el de la población. En la Figura 37 b) las tonalidades más obscuras de los clústeres indican que las características de las variables que los forman, describen de mejor manera a la variable que se quiere predecir, para este caso es el clúster 10.

En la Figura 38 c) se observa que las probabilidades de pertenecer al clúster 10 se acuerdo a sus características son: diámetro normal entre 25.5 y 50.1 centímetros, altura entre 6.6 y 9.4 metros y que crezcan a una altitud entre 1536 y 2.188.8 metros. Finalmente, la Figura 38 d) muestra la principal características que hace diferente al clúster 10 del resto, y que como se pudo observar con el visor de perfiles de la Figura 38 a), es la amplia consideración del diámetro de copa.

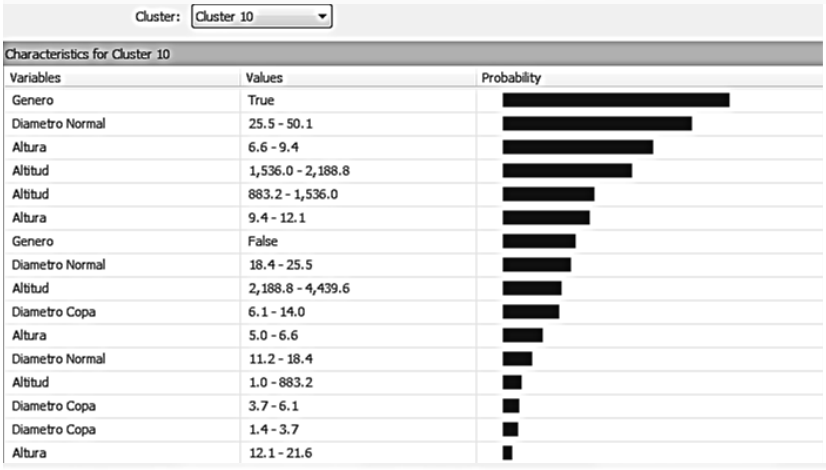
a) Visor de perfiles



b) Visor de Diagrama



c) Características del Clúster más significativo



d) Discriminación de clústeres

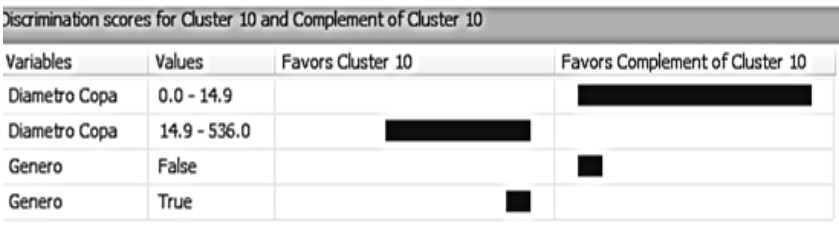


Figura 38: Análisis de Clúster
Fuente: Elaboración propia para la investigación

9.4.2 Generación de la tabla de predicciones para validación de datos de campo

Hernández Orallo *et al.* (2004) mencionan que después de generar el modelo con un conjunto de entrenamiento, éste se puede usar para predecir la clase de los datos de prueba; SSAS permite utilizar el modelo de minería creado, mediante la estimación de una probabilidad en función de las variables consideradas en la estructura de minería. Como ya se había mencionado, SQL Server es un sistema gestor de bases de bases de datos relacionales, por lo cual se expresa sólo mediante tablas (Date, 2001).

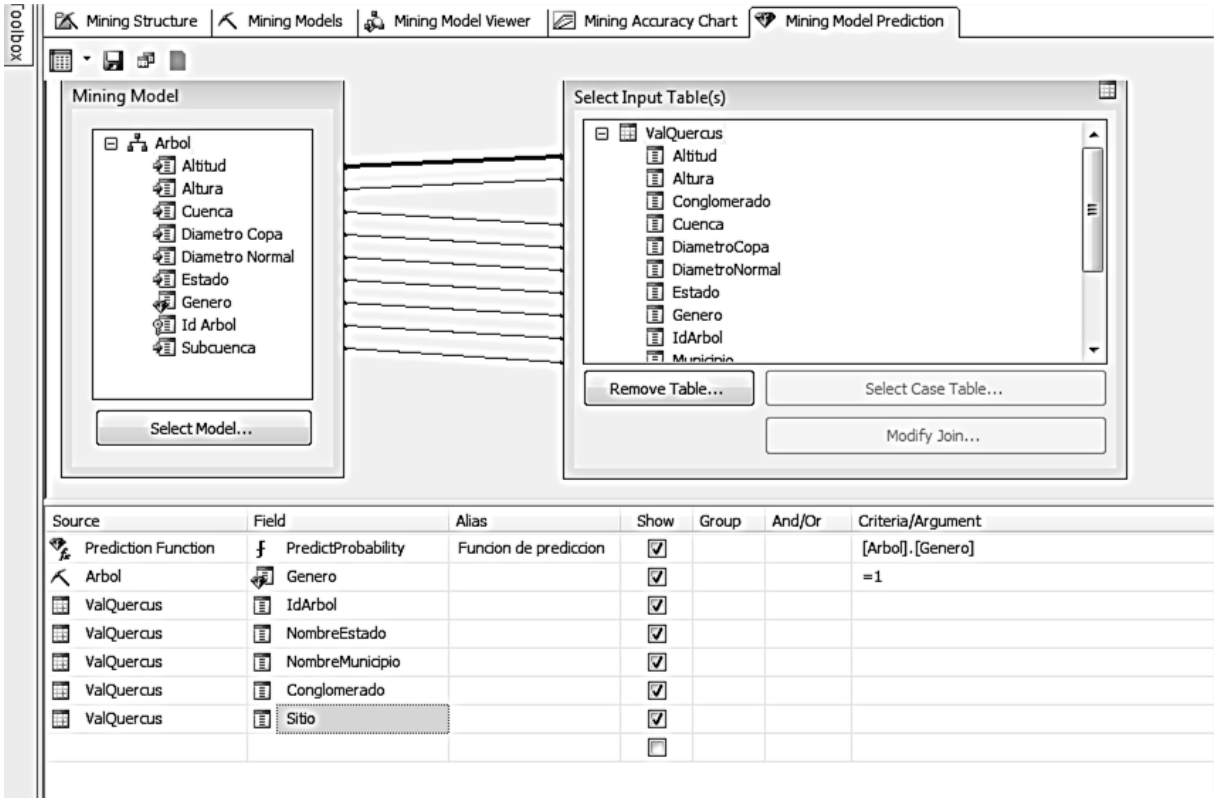
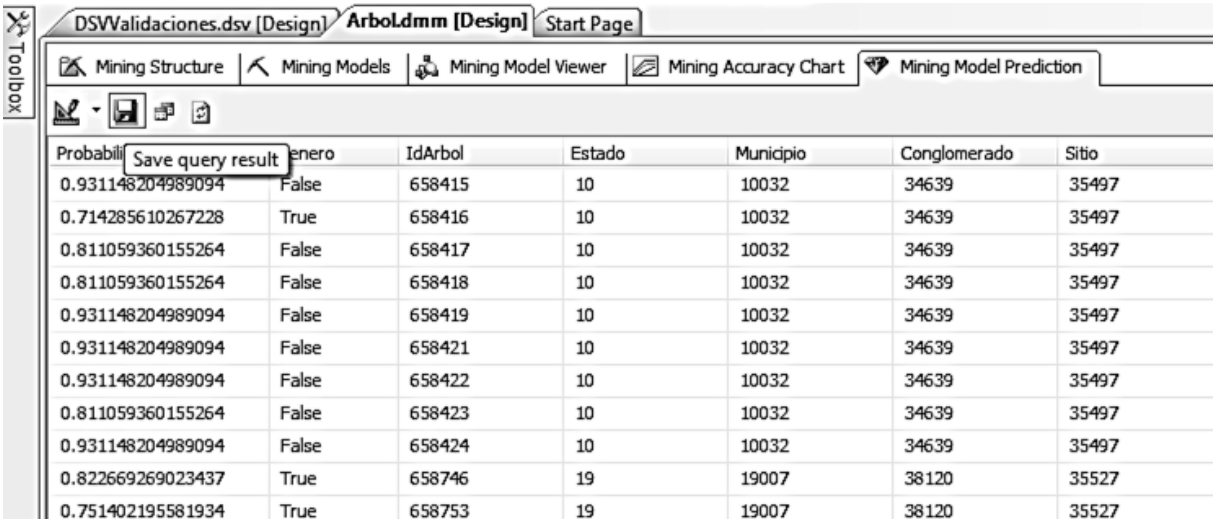


Figura 39: Definición de la consulta de predicción.
Fuente: Elaboración propia para la investigación.

Para crear una tabla de predicciones usando el modelo seleccionado, en la pestaña Predicción del Modelo de Minería se crea una correspondencia entre el modelo de minería que será usado y la tabla **ValQuercus**, como se muestra en la Figura 39, generada con anterioridad, y que

contiene la información de todos los árboles muestreados en el Inventario Nacional Forestal y de Suelos 2004-2009 y los atributos correspondientes a su ubicación.

La nueva tabla que se genera al ejecutar la consulta contiene tanto la información de ubicación de todos los árboles, tanto los del género *Quercus* como de otros géneros, su respectiva ubicación y la probabilidad asignada por el modelo, tal y como se muestra en la Figura 40.



Probabili	enero	IdArbol	Estado	Municipio	Conglomerado	Sitio
0.931148204989094	False	658415	10	10032	34639	35497
0.714285610267228	True	658416	10	10032	34639	35497
0.811059360155264	False	658417	10	10032	34639	35497
0.811059360155264	False	658418	10	10032	34639	35497
0.931148204989094	False	658419	10	10032	34639	35497
0.931148204989094	False	658421	10	10032	34639	35497
0.931148204989094	False	658422	10	10032	34639	35497
0.811059360155264	False	658423	10	10032	34639	35497
0.931148204989094	False	658424	10	10032	34639	35497
0.822669269023437	True	658746	19	19007	38120	35527
0.751402195581934	True	658753	19	19007	38120	35527

Figura 40: Consulta de predicción
Fuente: Elaboración propia para la investigación.

La información generada a partir de ésta tabla puede guardarse en la base de datos y ser usada para seleccionar los sitios donde se realizan las inspecciones mencionadas en CONAFOR (2012), como medio para garantizar la confiabilidad de los datos. La ventaja de usar la información de esta tabla radica en que los sitios donde se realiza la inspección no son seleccionados al azar, en vez de esto la visita se realiza a los sitios donde existe una alta probabilidad de encontrar inconsistencias. Una posible consulta para seleccionar los sitios de inspección, se muestra en el Cuadro 13, donde se selecciona la información de los árboles que son del género *Quercus*, pero que tienen una baja probabilidad de que realmente pertenezcan a éste género (probabilidad menor del 60%).

Cuadro 13. Consulta para seleccionar los posibles sitios de inspección.

```
select * from [Tabla de predicciones] where probabilidad < 0.6 and
Genero= 'true'
```

Fuente: Elaboración propia para la investigación.

La consulta selecciona 30, 862 registros de 212, 274, lo que implica que el 28% de la información del arbolado perteneciente al género *Quercus* es susceptible de verificación.

9.5 Resumen

En este capítulo se definieron, construyeron y probaron cuatro modelos de minería de datos para la clasificación del género arbóreo *Quercus*. La fuente de información fue la base de datos del INFyS 2004-2009, para el cual se tomó, como datos de entrenamiento y validación, un muestreo simple aleatorio de aproximadamente el 10% de la información dasométrica de los géneros maderables *Quercus, Pinus, Bursera, Lysiloma y Piscidia*, a nivel nacional. El resultado más destacable fue la selección del mejor modelo, realizado mediante comparaciones, para as cuales se recurrió a dos criterios, la gráfica de elevación y la matriz de clasificación, ambos obtenidos al momento de procesar los modelos. Después de la comparación, el mejor modelo de clasificación fue el de árbol de decisión con una precisión del 74%, seguido por el de análisis de conglomerados con una precisión 71% y los modelos de redes neuronales y regresión logística con una precisión del 69% y 67%, respectivamente. Para la generación de los modelos se utilizó el servicio de análisis de SQL Server 2008.

10. DISEÑO DE UN DATA WAREHOUSE PARA EL ANÁLISIS DE VOLUMEN DE MADERA, BIOMASA Y CARBONO

En el Capítulo 6 se mencionó que para el Inventario Nacional y de Suelos 2004-2009 se implementó un diseño de muestreo, una periodicidad y una metodología en el levantamiento, integración, sistematización y procesamiento de la información, que permitiera homogeneizar y hacer compatible los datos de diferentes periodos de muestreo. Siguiendo estas metas se desarrolló la base de datos utilizando Microsoft Access, como sistema gestor de base de datos, para la información recopilada en campo y SQL Server como sistema gestor para el almacenamiento final de la información y análisis de la misma.

De acuerdo con la CONAFOR (2012) la información almacenada en el servidor central es accedida mediante consultas SQL que les garantiza un rápido procesamiento de los datos para su reporte final. Al respecto, se debe mencionar que las consultas para obtener información específica sobre algún parámetro no es muy fácil de generar debido a la gran cantidad de operaciones que se tienen que realizar antes de obtener un producto.

De la misma manera como se desarrolló un método de muestreo que fuera aplicable a todos los niveles de inventario forestal, es necesario desarrollar nuevas técnicas de almacenamiento de información que permitan agilizar los procesos de análisis y al mismo tiempo contemple las etapas de integración futura de información, con mayor detalle, proveniente de los inventarios estatales.

En el capítulo 4 correspondiente a la minería de datos se habló del proceso de descubrimiento de conocimientos en bases de datos el cual se pretende aplicar a la base de datos del inventario para mejorar los tiempos de análisis y presentación de resultados.

El desarrollo de las primeras dos etapas del proceso KDD, de acuerdo Krzysztof *et al.* (2007), puede llevarse a cabo mediante las bases de datos y los Data Warehouse ya que los software empleados para administrar la información proveen una manera eficiente para la recuperación de datos y las herramientas necesarias para preparar y seleccionar los datos para los subsecuentes

pasos en el proceso de descubrimiento de conocimientos. En la Figura 41 se muestra la arquitectura propuesta por Krzysztof *et al.* (2007), para acceder a la información desde una base de datos o un Data Warehouse.

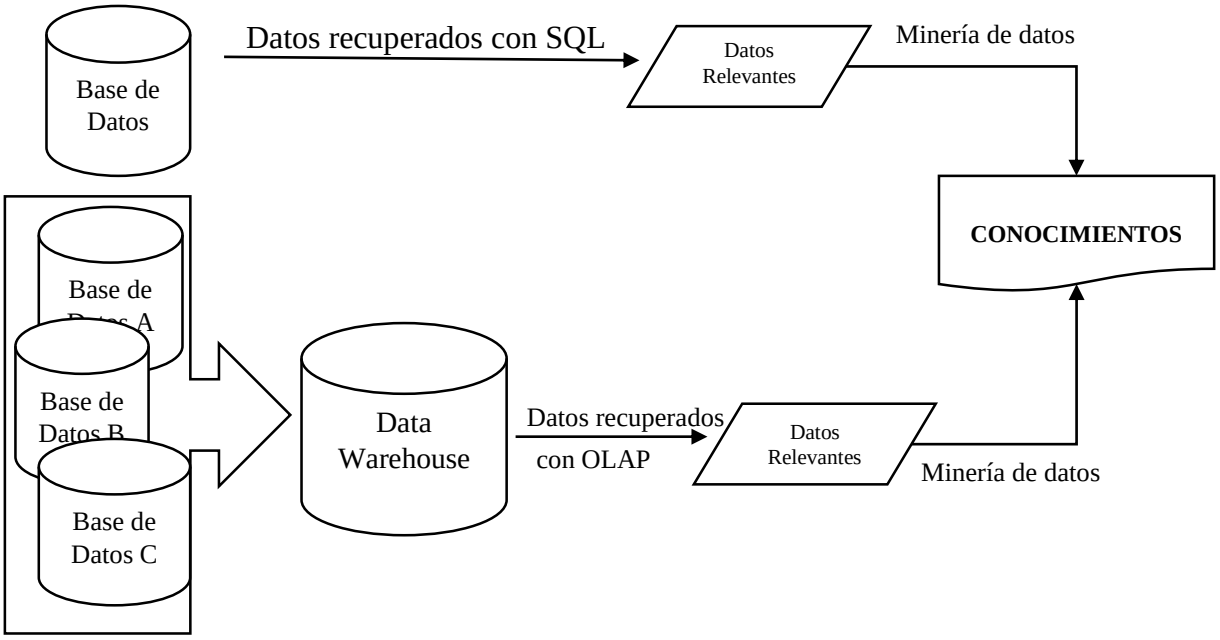


Figura 41: Bases de datos y Data Warehouse y su relación en la generación de conocimientos
Fuente: Modificado de Krzysztof *et al.* (2007)

El diseño de almacenes de datos con fines de investigación pueden agilizar los procesos de análisis y hasta utilizarse para la generación de nuevos conocimientos combinando la información histórica existente con los nuevos descubrimientos (Ruíz *et al.*, 2008). La importancia de los Data Warehouse radica en que la mejor información para ser minada es aquella que ha permanecido almacenada y por lo tanto guardan información histórica de eventos y resultados que con el tiempo puede llegar a combinarse con información novedosa y conjuntamente ampliar el panorama de conocimiento de un fenómeno.

Un caso de éxito en la depuración de datos y rápida accesibilidad a la información la encontramos en Ruíz *et al.* (2008) donde se propuso la implementación de dos aplicaciones informáticas con distintas plataformas de desarrollo que permitió reducir el tiempo de ingreso de los datos colectados en campo, así como el almacenamiento y análisis para la obtención de información estadísticamente validada y en periodos cortos de tiempo. Se utilizaron diferentes

tecnologías de programación para crear los formularios, aunque la columna vertebral del proyecto fue la construcción de un Data Warehouse administrado por SQL Server. Este es uno de los primeros proyectos agronómicos, que hace uso del potencial de los actuales gestores de bases de datos, con la finalidad de facilitar el levantamiento de datos de campo.

Microsoft SQL Server cuenta con una serie de herramientas capaces de facilitar el acceso y análisis de la información provenientes de diferentes fuentes. Dentro de estas herramientas se incluye los proyectos de servicio de integración, para la administración de información desde y hacia diferentes formatos que además cuenta con el potencial de usar programación en diferentes lenguajes para la manipulación de la información. Otro proyecto es el servicio de análisis, cuya principal característica es el procesamiento de grandes volúmenes de información mediante cubos y minería de datos. Dado que en el alcance de la tesis sólo se considera la estimación de volumen de madera, biomasa y carbono, dejando fuera los demás procesos, el método considerado para la integración del Data Warehouse coincide con la visión de Kimball.

Kimball propone 4 pasos fundamentales, que se debe seguir en estricto orden, para la correcta modelación de un proceso, la presente investigación se basó en estos 4 pasos que se muestran de forma jerárquica en la Figura 42.

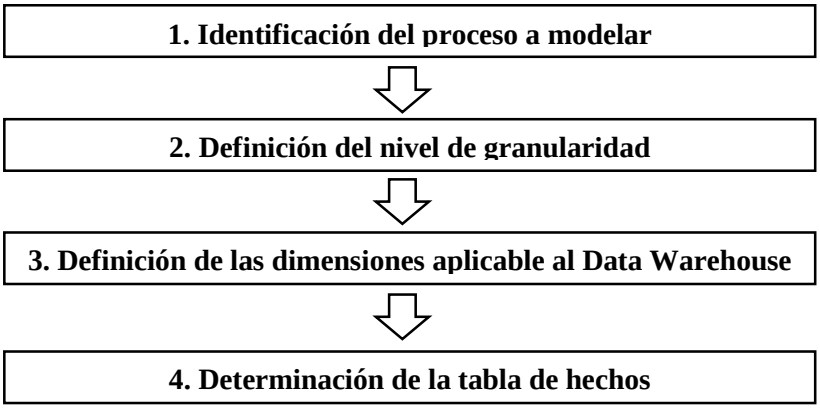


Figura 42: Proceso para el diseño de un Data Warehouse desde la perspectiva de Kimball
Fuente: Elaboración propia para la investigación.

El Data Warehouse será integrado a partir de Data Marts que modelen los diferentes procesos de estimación. Esta decisión también se soporta con el hecho de que el diseño de Data Warehouse propuesto por Inmon está enfocado a modelar todos los procesos, esto obviamente es un procedimiento muy tardado y no funcional para los objetivos de la presente tesis.

El modelo propuesto por Kimball difícilmente se apega a las restricciones de normalidad que gobiernan a las bases de datos relacionales. La mayoría de los Data Marts diseñados bajo éste enfoque se basan en un esquema de estrella que facilita el acceso y análisis de información. En las siguientes secciones se analizará cada una de las etapas propuestas por Kimball hasta obtener el diseño conceptual del Data Mart que modela el proceso de análisis de volumen de madera biomasa y carbono.

10.1 Identificación del proceso a modelar

El primer paso para la construcción de un buen Data Warehouse, desde la perspectiva de Kimball, es el estudio de los diferentes procesos que se llevan a cabo en una empresa con la finalidad de captar los aspectos más relevantes y de esta manera plasmarlo en el diseño conceptual del Data Warehouse. La terminología empleada por Kimball para el diseño de Data Warehouse se apega estrictamente al sector de negocios de una empresa por lo cual es necesario dejar claro que los términos originales se adecuarán a la terminología empleada en el ámbito forestal.

En primer lugar se debe tener presente que la teoría Data Warehouse está enfocado hacia el análisis de negocio y que el INFyS es un proceso general compuesto de varios procesos específicos, por lo que es necesario hacer una analogía entre empresa e Inventario Nacional Forestal y de Suelos, con el fin de adecuar los conceptos teóricos a los datos.

Una manera de ilustrar el potencial que ofrecen los Data Warehouse para el de análisis de datos, en la presente tesis se diseñará un DW con el objetivo de estimar el volumen de madera en bosques y selvas usado las restricciones descrita en el Informe de Resultados 2004-2009 y el método para calcular el volumen mediante ecuaciones de volumen propuesto en Méndez y De los Santos (2011); además se incorporarán algunos requerimientos internacionales como son los cálculos de la biomasa y el carbono. Definido el objetivo y alcance del DW, se entenderá como proceso a la serie de actividades y metodologías empleadas desde el levantamiento de datos en campo hasta la estimación del volumen de madera, biomasa y carbono.

En la sección 3.3 sobre modelado dimensional se menciona que, antes de comenzar a diseñar el esquema del DW, primero es necesario definir el nivel de granularidad de los datos, las dimensiones que ayudarán a analizar la información a diferentes niveles de detalle y finalmente la tabla de hechos que contendrá las métricas o información que será utilizada para el análisis. En las siguientes secciones se profundiza en la metodología empleada para la definición de los objetos mencionados en éste párrafo.

10.2 Definición del nivel de granularidad de los datos

La determinación del nivel de granularidad condicionará los requerimientos de hardware especializado para llevar a cabo las tareas de análisis. En un sentido más coloquial se puede decir que a mayor nivel de detalle en el almacenamiento de los datos, mayor será en tamaño la base de datos y mayor los requerimientos para el equipo encargado de resguardar y analizar la información.

En la sección anterior se seleccionó el proceso de cálculo del volumen de madera en bosques y selvas; como punto de partida se tiene la metodología aplicada en CONAFOR (2012) donde se describe una serie de hechos considerados durante los diferentes procesos de cálculo, incluyendo el de volumen de madera. En primer lugar hay que analizar la metodología desarrollada e implementada en la etapa de muestreo en campo; de acuerdo con éste informe, el diseño de muestreo, tamaño y forma de los sitios, fueron tomados a partir de un estudio previo elaborado por el Instituto Nacional de Investigaciones Forestales, Agrícolas y Pecuarias (INIFAP) para la Secretaría de Medio Ambiente y Recursos Naturales (SEMARNAT) en 2002. Los estimadores utilizados para los cálculos de los parámetros anteriormente mencionados fueron generados por Velasco *et al.* (2003), aquí se menciona que las estimaciones se deben obtener tanto a nivel de formación vegetal como de ecosistema, y los dos estratos en que se presenta actualmente la información nacional.

El diseño de muestreo fue Estratificado Sistemático por Conglomerados en Dos Etapas, la unidad de muestreo primaria (UMP) está representado por los conglomerados y se encuentran

conformados por un máximo de 4 unidades de muestreo secundarias (UMS), representadas por los sitios de muestreo. El número de UMS varía porque algunos sitios no fueron accesibles y por lo consiguiente no se levantó información en ellos.

Los parámetros de interés son estimados a partir de un estimador de razón el cual permite usar una variable auxiliar para realizar las estimaciones por hectárea apoyándose en el área muestreada a nivel de UMP, es decir, se obtiene una variable auxiliar X_i (hectáreas), correlacionada con Y_i (parámetro de interés), para cada unidad de la muestra. Para emplear el estimador de razón, se suman todos los datos de las UMS, tanto para la variable de interés como para la variable auxiliar, y se considera el dato a nivel de UMP, CONAFOR (2012).

La ecuación (16) se aplica a la estimación del volumen, número de árboles y otros parámetros con el detalle de que debe haber un método definido para calcular estos valores a nivel de UMS, para el caso de la presente tesis será el método de cálculo de volumen mediante ecuaciones de volumen propuesta por Méndez y De los Santos (2011).

$$\hat{R} = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i} = \frac{\bar{Y}}{\bar{X}} \qquad (16)$$

La metodología utilizada para las estimaciones define el nivel de granularidad de los datos, de tal manera que la información debe ser almacenada de forma agregada a nivel de UMS, es decir el mayor nivel de detalle en el DW debe ser información resumida a nivel de sitios.

10.3 Definición de las dimensiones aplicables al Data Warehouse

Uno de los aspecto más importantes a considerar en la selección de las dimensiones que se necesitan para el análisis de la información, se relaciona con el alcance que tienen los estimadores. Desde un principio el diseño de muestreo tuvo como base la cartográfica generada por el INEGI en la que la unidad básica de clasificación es la comunidad vegetal cuyos niveles de estimación naturales fueron los subestratos, estratos y ecosistemas, Velasco *et al.* (2003), para el INFyS 2004-2009 esta metodología se respaldó en el uso de datos vectoriales

provenientes de las Cartas de Uso de Suelo y Vegetación, serie III y IV, generadas también por el INEGI, CONAFOR (2012).

La correcta aplicación de los estimadores está condicionada a considerar como subestrato a la condición de vegetación primaria o secundaria y como estrato a la formación vegetal. En la Figura 43 se muestra la disposición de información en tablas para la presentación de informes.

Ecosistema	Estrato/ Formación	Subestrato/Condición de conservación	Comunidad vegetal/tipo de vegetación
Bosques	Coníferas	Vegetación primaria	Bosque de ayarín
			Bosque de cedro
			Bosque de oyamel
			Bosque de pino
			Bosque de táscate
			Matorral de coníferas
		Vegetación secundaria	Bosque de ayarín
			Bosque de cedro
			Bosque de oyamel
			Bosque de pino
			Bosque de táscate
			Matorral de coníferas

Figura 43: Nivel jerárquico para la presentación de informes y resultados del INFyS 2004-2009.
Fuente: CONAFOR (2012).

La jerarquía que se muestra en la Figura 43 en forma de tabla representa las bases para definir la primera dimensión denominada "Vegetación". La base de datos del INFyS 2004-2009 sólo contiene la información correspondiente a Ecosistemas y Comunidad Vegetal, almacenado en la tabla **CatTipoVegetacionInegiGeneral**, por lo que la jerarquía de ésta dimensión queda como en la Figura 44.

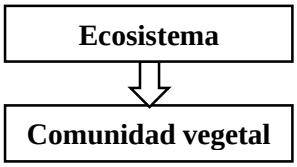


Figura 44: Jerarquía de niveles para la dimensión ESTRATOS.
Fuente: Elaboración propia para la investigación.

Otro aspecto importante que debe ser considerado para la definición de las dimensiones es el alcance del inventario. Como está definido para ser a nivel nacional, las estimaciones de las variables a nivel estatal, municipal o predial no están permitidas. La superficie de varios estados o municipios del país no son lo suficientemente grandes, considerando el tipo de muestreo y la distancia entre conglomerados, para garantizar un tamaño adecuado de muestras. Considerando estas limitantes que entra en conflicto con la perspectiva del INFyS que fomenta la realización de inventarios estatales, se logró incluir en el diseño de muestreo la flexibilidad que permite aumentar unidades de muestreo según las prioridades de información y los recursos disponibles, por lo que se puede adaptar para obtener información a nivel regional, incrementando la intensidad de muestreo en las áreas con mayor interés, CONAFOR (2012). Considerando está información, es conveniente incluir una segunda dimensión en el Data Mart, la cual represente los diferentes niveles de agregación a nivel espacial. Esta dimensión que se denominará "Región" puede no ser incluido en el diseño del Data Mart actual ya que la intensidad con la que se llevó a cabo el muestreo no proporciona la suficiente información para realizar estimaciones a este nivel. Por cuestiones prácticas se diseñará el Data Mart con esta dimensión y se ejemplificará un posible uso. Los niveles jerárquicos quedarán como se muestra en la Figura 45.

Figura 45. Jerarquía de niveles para la dimensión "Región"
Fuente: Elaboración propia para la investigación.

Hasta el momento se ha determinado el nivel de granularidad y dos dimensiones que se usarán para el análisis de la información a diferentes niveles de detalle. El siguiente paso es la determinación de los datos requeridos en la tabla de hechos, cuyos criterios para su determinación se detallan en la siguiente sección.

10.4 Determinación de la tabla de hechos.

La definición de la tabla de hechos está en función de los datos necesarios y suficientes para calcular el volumen de madera. Recuérdese que en la sección 3.3.2 se dijo que la tabla de hechos se encuentra integrada por información relevante de análisis y que por lo general se encuentra agregada.

En la sección 10.2 se propusieron las bases para determinar el nivel de granularidad, llegándose a la conclusión de que el mayor nivel de detalle sería a nivel de Unidad de Muestreo Secundaria, por lo cual, la información debe ser agregada a nivel de UMS. Debido a que la estimación del volumen se realiza a nivel de árbol, entonces durante el proceso de cagar de datos se deben llevar a cabo estas estimaciones, adicionalmente se deben obtener otros indicadores durante el proceso.

Los siguientes pasos indican la dirección que sigue el procesamiento de los datos antes de formar parte de la tabla de hechos.

1. Seleccionar aquellos árboles que cumplen con las restricciones establecidas por la CONAFOR en su método de estimación de volumen.
2. Realizar los cálculos correspondientes a nivel de individuo o árbol, el nivel de detalle más bajo, para calcular el volumen por individuo, usando las ecuaciones de volumen propuestas por Méndez y De los Santos (2011).
3. Obtener los datos agregados por Unidad de Muestreo Secundaria o Sitio, suma del volumen de madera y cantidad de árboles medidos en el sitio.
4. Incluir la clave de sitio que servirá como llave foránea para relacionarse con la dimensión Región.
5. Incluir la clave de vegetación que servirá como llave foránea para relacionarse con la dimensión Estrato.

10.5 Modelación del proceso

En la sección anterior se realizó una revisión de la información que se debe incluir en el Data Warehouse, así como las restricciones a la que está sujeta esta para no incurrir en estimaciones de baja confiabilidad debido a la naturaleza de los datos y el alcance de los estimadores. En esta sección se llevará a cabo la modelación de los procesos de estimación a partir de la base de datos.

El primer paso para modelar el proceso de estimación de volumen es identificar las tablas relacionadas durante esta etapa. En el capítulo 8 se llevó a cabo el análisis de la base de datos del INFyS detectándose que está integrado de la siguiente manera: 67 tablas distribuidas de las cuales 2 tablas son señaladas como principales, 22 tablas secundarias y 43 tablas del tipo catálogo. Para la generación del Data Mart que modelo el proceso de estimación de volumen de madera se tiene que redefinir la importancia de cada tabla en la base de datos, siendo sólo importante aquellas que están relacionada directamente con el proceso.

La sección 10.3 permitió conocer que los estimadores están diseñados para ser aplicados a nivel de estratos y sub estratos, que de acuerdo al análisis realizado de la base de datos del INFyS 2004-2009, los datos correspondientes a éstos niveles se encuentran almacenados en la tabla **CatTipoVegetaciónInegiGeneral** que a su vez se encuentra relacionada directamente con la tabla **Tblsitio** mediante los campos **TipoVegetacionEsp** y **TipoVegetacionLev.** Las tablas **Tblconglomerado** y **TblSitio** representan las unidades de muestreo primarias y secundarias, respectivamente, por lo cual es necesario tener identificada cada una de estas unidades, de esta manera también se incluyen estas tablas al Data Mart. Finalmente, la tabla **TblArboladoBosqueSelva** es de donde se extrae la información agregada por sitio de la información dasométricas del arbolado, de aquí se genera la tabla de hechos.

La mejor manera de detectar las tablas relacionadas directamente es visualizando el diagrama de la base de datos y generar una tabla de procesos, que comúnmente es usada para determinar las tablas relacionadas en uno o más procesos; en el Cuadro 14 se muestra ésta tabla, donde se puede observar que sólo 6 tablas, de 67 totales, están relacionadas con el proceso de cálculo de

volumen de madera. De esas seis tablas, sólo la tabla **TblArboladoBosqueSelva** contiene información numérica que es utilizada para calcular el volumen de madera. De manera técnica las tablas que no tienen información numérica de interés, pero que están involucradas en el proceso, forman parte de las dimensiones.

Cuadro 14. Tablas relacionadas en el proceso de estimación de volumen de madera.

Entidades	Relación	Entidades	Relación
CatAbundancia		CatTrozaTipo	
CatAccesibilidad		CatUsoActualCA	
CatAniosIncendios		CatUsoEspecie	
CatCarta150		CatUsoLocalReg	
CatCategoriaSueloXProfundidad		CatUsoSuelo	
CatCausaImpacto		CatUsoSueloSinCubiertaVegetal	
CatCoberturaXVeg		CatVegetacionSecundaria	
CatCondicion		CatVigorArboladoBosqueYSelvaArbolEtapa	
CatCuerpoAgua		CatVigorRepobladoBosqueYSelva	
CatDanio		TblArboladoBosqueSelva	X
CatDatosAutomaticos		TblArboladoSubBosqueSelva	
CatDegradacion		TblArbSubMuestraOtrasCom	
CatEpifitaTipo		TblCobertura	
CatErosion		TblCoberturaOtrasC	
CatEspecie		TblCoordenada	
CatEspFlora		TblCoordenadaSitio	
CatEstado	X	TblRepobladoBosque	
CatExposicion		TblRepobladoOtrasCom	
CatFisiografia		TblRepobladoSelva	
CatFisonomia		TblVegMayorOtrasCom	
CatFormatoTipo		TblVegMenorBosqueSelva	
CatGenero		TblVegMenorOtrasCom	
CatImpactoVegSueloH2O		TblCaracEspFlora	
CatMantillo		TblCoordPtoCtrl	
CatMercadoEspecie		TblCuerpodeAgua	
CatMunicipio	X	TblDiversidadXEstrato	
CatNivelAfectacion		TblEpifita	
CatNivelAfectacionH2O		TblImpactoAmbiental	
CatTenencia		TblIncendio	
CatTipoAcceso		TblJustificado	
CatTipoConglomerado		TblSuelo	
CatTipodeEstratos		TblConglomerado	X
CatTipoIncendio		TblSitio	X
CatTipoVegetacionInegiGeneral	X		

Fuente: Elaboración propia para la investigación a partir del INFyS 2004-2009.

La tabla **CatVegetacionInegiGeneral** contiene la información de los estratos y subestratos por lo cual ella sola da lugar a la dimensión **Vegetación**. La tabla **TblSitio** y la tabla **TblConglomerado** junto con los catálogos **CatEstado** y **CatMunicipio** son utilizadas como apoyo para agregar la información del arbolado de bosques y selvas; la unión de estas da lugar a la dimensión **Región**. El esquema relacional del proceso de cálculo de volumen de madera queda como se muestra en la Figura 46.

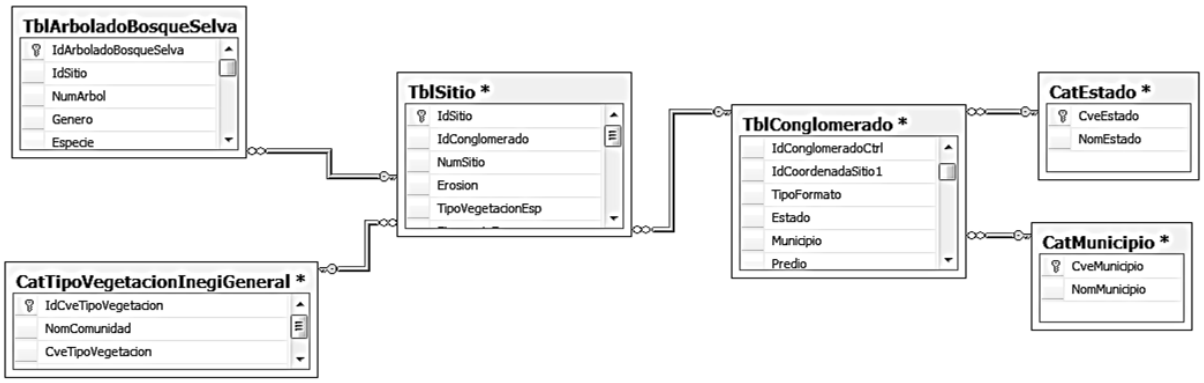


Figura 46. Modelo relacional del Data Mart para el análisis de volumen, biomasa y carbono
Elaboración: Propia para la investigación.

Para que el acceso a los datos se lleve a cabo de una manera más eficiente, se deshizo la normalidad de las tablas que se presentan en la Figura 46 y se creó un esquema de estrella como el que se muestra en la Figura 47.

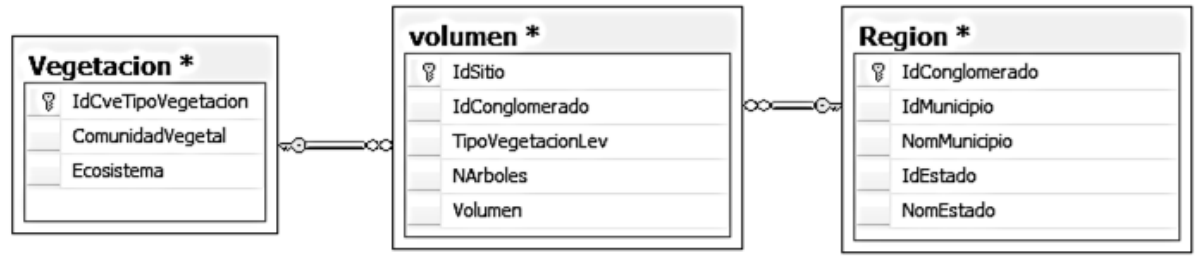


Figura 47. Esquema de estrella del Data Mart para el análisis de volumen, biomasa y carbono
Elaboración: Propia para la investigación.

10.6 Resumen

En este capítulo se diseñó el Data Warehouse para el análisis de volumen de madera, biomas y carbono, basado en el enfoque de Ralph Kimball (Data Warehouse Bus). Éste enfoque consiste de seguir los cuatro pasos básicos siguientes:

- La identificación del proceso a modelar.
- La identificación del nivel de granularidad de los datos.
- La definición de las dimensiones.
- La determinación de la tabla de hechos.

El proceso a modelar fue el cálculo de volumen de madera, biomasa y carbono.

El nivel de granularidad consistió de información agregada a nivel de sitios de muestro.

Las dimensiones definidas fueron un total de dos. La primera dimensión etiquetada como ESTRATOS consistió de dos niveles cuya jerarquía fue Ecosistemas y luego Comunidad Vegetal. La segunda dimensión consistió de tres niveles cuya jerarquía es Estados, Municipio y Conglomerados.

La tabla de hechos se conformó por información agregada a nivel de sitios de los cálculos de volumen, biomasa y carbono, así como la información del número de sitios y el número de hectáreas muestreadas.

11. IMPLEMENTACIÓN DEL DATA WAREHOUSE Y DESCRIPCIÓN DE LOS PAQUETES DE CARGA DE DATOS Y ANÁLISIS

En este capítulo se muestra a detalle el desarrollo e implementación del Data Warehouse para análisis volumen de madera, biomasa y carbono, a nivel nacional, a partir del Inventario Nacional Forestal y de Suelos 2004-2009. En primer lugar se verán el tema relacionados con la creación del Data Warehouse desde el administrador de SQL Server, así como el código SQL usado. En segundo lugar se describen los objetos del paquete de integración desarrollado con el servicio de integración para llevar a cabo la extracción, transformación y carga de los datos. Posteriormente se describen los objetos del paquete creado con el servicio de análisis para procesar los cubos multidimensionales de análisis. Para cada uno de los paquetes se muestra a detalle cada paso a seguir para su correcta ejecución.

Los paquete fueron desarrollados, como ya se había mencionado, para analizar la información de volumen de madera biomasa y carbono a nivel nacional, pero empleado diferentes métodos; para el estado de México se utilizaron ecuaciones de volumen específicas, como las que se presentaron en la sección 6.5, ecuaciones 1 al 7, y una ecuación general para el resto de las entidades federativas, ecuación 10. La intención de realizar un análisis a nivel nacional es mostrar la capacidad de los cubos multidimensionales para procesar grandes cantidades de información de manera eficiente; no se incluyeron las ecuaciones de volumen correspondientes para todos los estados, debido a que existe una gran cantidad de ellas y el tiempo para concluir la tesis no sería suficiente para concluirla en los tiempos establecidos para su realización.

11.1 Creación del Data Warehouse

11.1.1 Creación del Data Warehouse mediante consulta

Este método es universal, usado por muchos Sistemas Manejadores de Bases de Datos, que consiste en realizar una consulta usando un lenguaje de definición de datos, como el que se

muestra en el Cuadro 15, donde se crea la base de datos y se definen la tabla de hechos etiquetada como **Volumen,** y las tablas de dimensiones etiquetadas como **Vegetación** y **Región**. Para crear el Data Warehouse con este método es necesario que el usuario conozca la interfaz de SQL Server Management Studio.

Cuadro 15. Comando SQL para la creación de las tablas del Data Warehouse de análisis.

```
Create database dmm1
USE [dmm1]
CREATE TABLE [dbo].[volumen](
[IdSitio] [int] NOT NULL,
[TipoVegetacionLev] [tinyint] NULL,
[NArboles] [smallint] NULL,
[Volumen] [real] NULL,)

CREATE TABLE [dbo].[Vegetacion](
[IdCveTipoVegetacion] [tinyint] NOT NULL,
[ComunidadVegetal] [nvarchar](50) NULL,
[Ecosistema] [nvarchar](50) NULL,)

CREATE TABLE [dbo].[Region](
[IdSitio] [int] NOT NULL,
[IdConglomerado] [int] NOT NULL,
[IdMunicipio] [int] NOT NULL,
[NomMunicipio] [nchar](50) NULL,
[IdEstado] [int] NOT NULL,
[NomEstado] [nchar](50) NULL,)
```

Elaboración: Propia para la investigación.

11.1.2 Creación del Data Warehouse mediante aplicación

La base de datos o Data Warehouse de análisis para el volumen de madera, biomasa y carbono se puede crear mediante tres métodos. El primer método hace uso de la aplicación desarrollada para ejecutar las tareas sin necesidad de entrar a SQL Server Management Studio. En la Figura 48 se muestra una interfaz con tres botones, el primer botón del lado izquierdo construye todas las tablas del DW con tan sólo seleccionarlo.

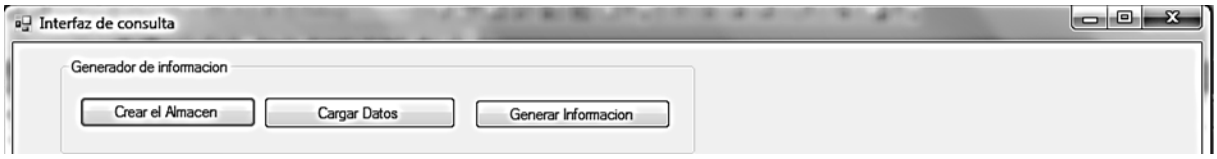


Figura 48: Creación del Data Warehouse de análisis mediante aplicación.
Elaboración: Propia para la investigación.

Los métodos restantes los puede llevar a cabo cualquier usuario de SQL, ya que no se necesita conocimientos avanzados de bases de datos. En las siguiente secciones se detallan los pasos a seguir para creación de la base de datos con el administrador de SQL Server.

11.1.3 Creación del Data Warehouse mediante el asistente de SQL Server.

La creación del Data Warehouse vía el administrador de base de datos de SQL Server es una tara que puede realizar cualquier usuario siguiendo las siguientes instrucciones.

1. Abrir una instancia de SQL Server 2008 o un versión posterior.
2. Posicionar el mouse en la pestaña "Base de datos" del explorador de objetos.
3. Hacer clic con el botón secundario del mouse y seleccionar Nueva base de datos, como en la Figura 49.

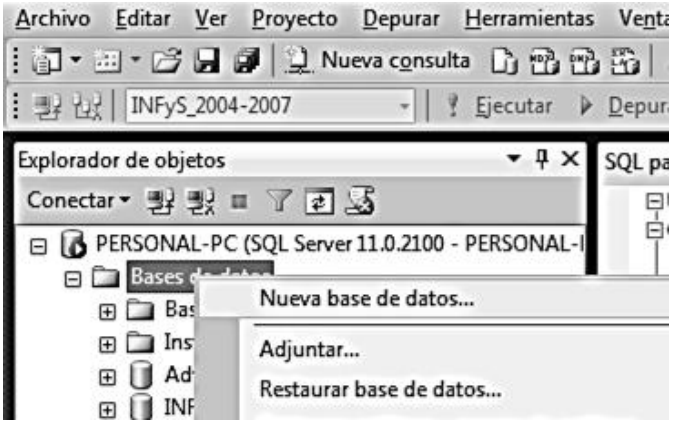


Figura 49. Creación de la base de datos desde el administrador de base de datos
Elaboración: Propia para la investigación.

Al aparecer una ventana como el de la Figura 50, en el recuadro Nombre de la base de datos se coloca "dmm1" y se presiona el botón de aceptar, éste es el nombre de la base de datos y es necesario que se etiquete de esta forma para que los paquetes de carga y análisis, que se describen más adelante, funcionen.

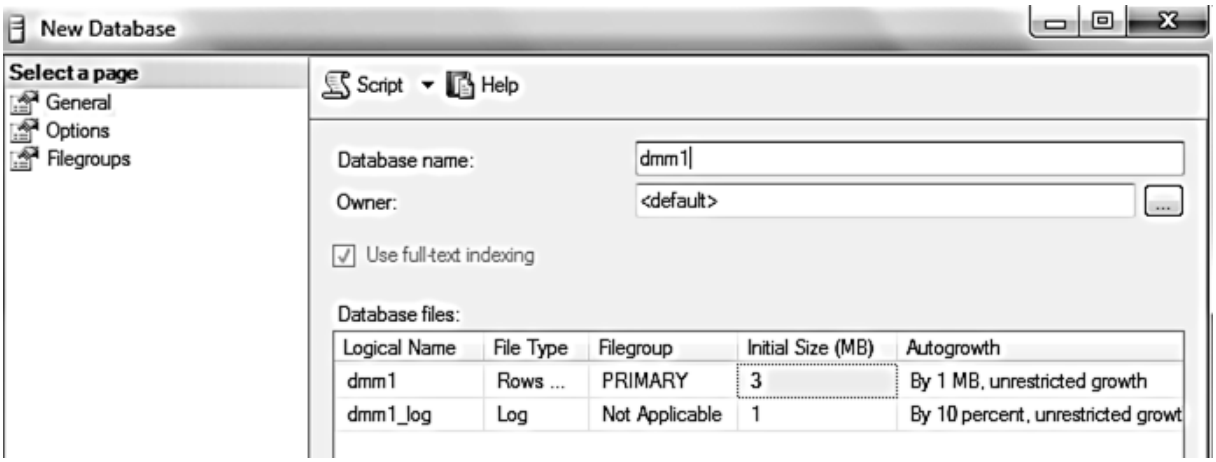


Figura 50: Ventana de propiedades del Data Warehouse
Elaboración: Propia para la investigación.

La tabla de hechos, volumen, se crean de manera similar a la base de datos. Una vez construida la base de datos, ésta se expande haciendo clic en el signo + que aparece a un lado izquierdo de la etiqueta. Se selecciona la etiqueta tabla, y se hace clic derecho para seleccionar la pestaña Nueva tabla en el menú emergente; este procedimiento se muestra en la Figura 51.

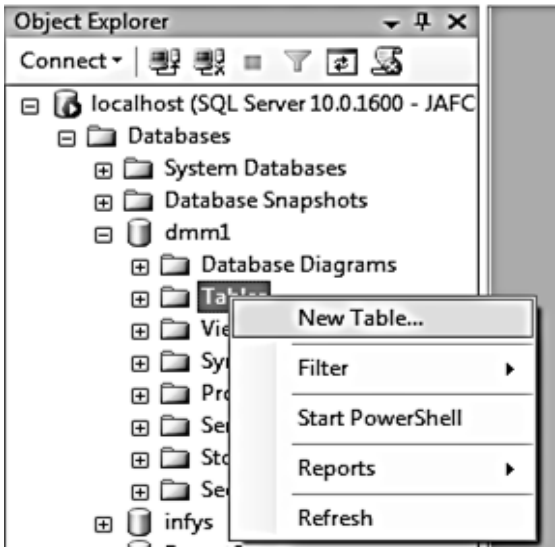


Figura 51. Creación de tablas usando el asistente.
Elaboración: Propia para la investigación.

Después de seleccionar Nueva tabla se mostrará una nueva ventana donde se definen los campos con sus dominios y las restricciones de integridad (las llaves primarias, externas y otras restricciones). En la Figura 52 se muestra la configuración que debe tener cada una de las tablas para que sean compatibles con los paquetes de integración y análisis desarrollados.

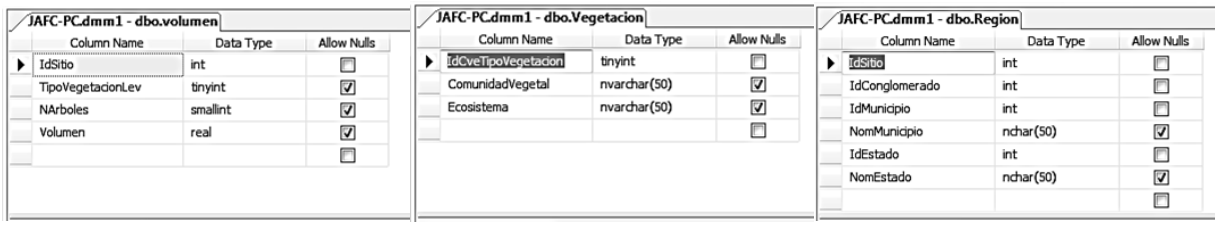


Figura 52. Nombres y configuración de las tablas del Data Warehouse de análisis
Elaboración: Propia para la investigación.

El Tipo de datos o dominio se selecciona haciendo clic en la columna Tipo de Datos. Al finalizar la configuración de la tabla se hace clic derecho en la pestaña dbo.Tabla_ y aparece una ventana emergente donde solicita el nombre que se le dará a cada tabla. Los nombres que se deben proporcionar aparecen en la parte superior de la Figura 52 posterior a la etiqueta dbo.

11.2 Paquete de Integración, Transformación y Carga

La tarea conocida como Extracción, transformación y carga consiste en un proceso automático para recuperar datos desde una o más fuentes de información hacia las tablas que conforman el Data Warehouse. El Paquete ETL, como será conocido a éste paquete, considera todas las restricciones y agregaciones de información requeridas para aplicar las operaciones de análisis mencionadas en el capítulo 3. Para el desarrollo de éste paquete se usó del archivo (MS Access) donde se encuentra la base de datos del Inventario Nacional forestal y de Suelos 2004-2009, una hoja de cálculo (MS Excel), donde se encuentran las ecuaciones de volumen para el estado de México y las herramientas del servicio de integración incluida en la versión Enterprise de SQL Server.

Para ejecutar el paquete de integración es necesario seguirlas instrucciones siguientes:

1. Abrir la aplicación SQL Server Data Tools que se encuentra en la lista de todos los programas.
2. Selecciona abrir proyecto en la ventana de Visual Studio 2010 que aparece.
3. Buscar archivo en la ruta C:\proyecto\Carga ArboladoBosqueSelva y seleccionar el proyecto.

En la solución que se abre se encuentran todos los comandos y tareas usados para la carga de datos desde la base de datos del INFyS, hacia las tablas del Data Warehouse. En la primera pestaña, etiquetada como **control de flujo**, se encuentran dos objetos (Figura 53) que llevan a cabo las tareas ETL y procesamiento del cubo de análisis.

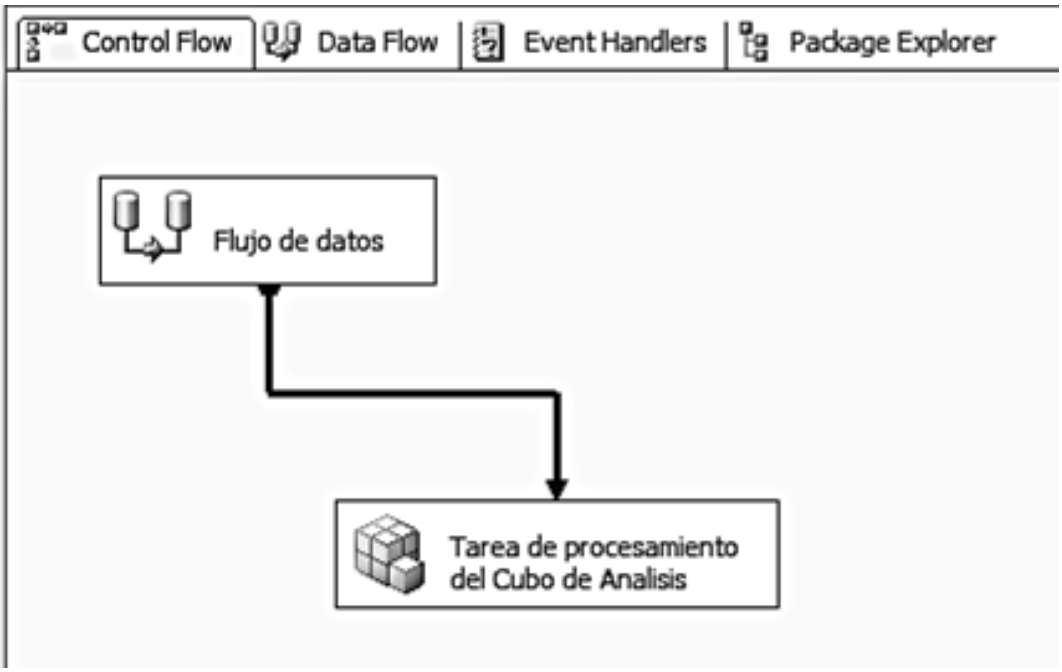


Figura 53: Control de flujo del Paquete ETL.
Elaboración: Propia para la investigación.

El primer objeto etiquetado como Flujo de datos representa el contenedor para una serie de objetos usados en la manipulación y procesamiento de la información desde diferentes fuentes de datos. El segundo objeto, llamado Tarea de Procesamiento de Cubos de análisis, contiene la conexión hacia el paquete de análisis creado con SSIS y su única tarea consiste en procesar las dimensiones y los cubos definidos de este paquete. En las siguientes secciones se describen con mayor detalle cada una de las tareas.

11.2.1 Integración de la tabla de hechos

En la Figura 54 se visualiza la pestaña flujo de datos donde se puede observar una serie de objetos relacionados que contienen conexiones e instrucciones SQL de selección, proyección,

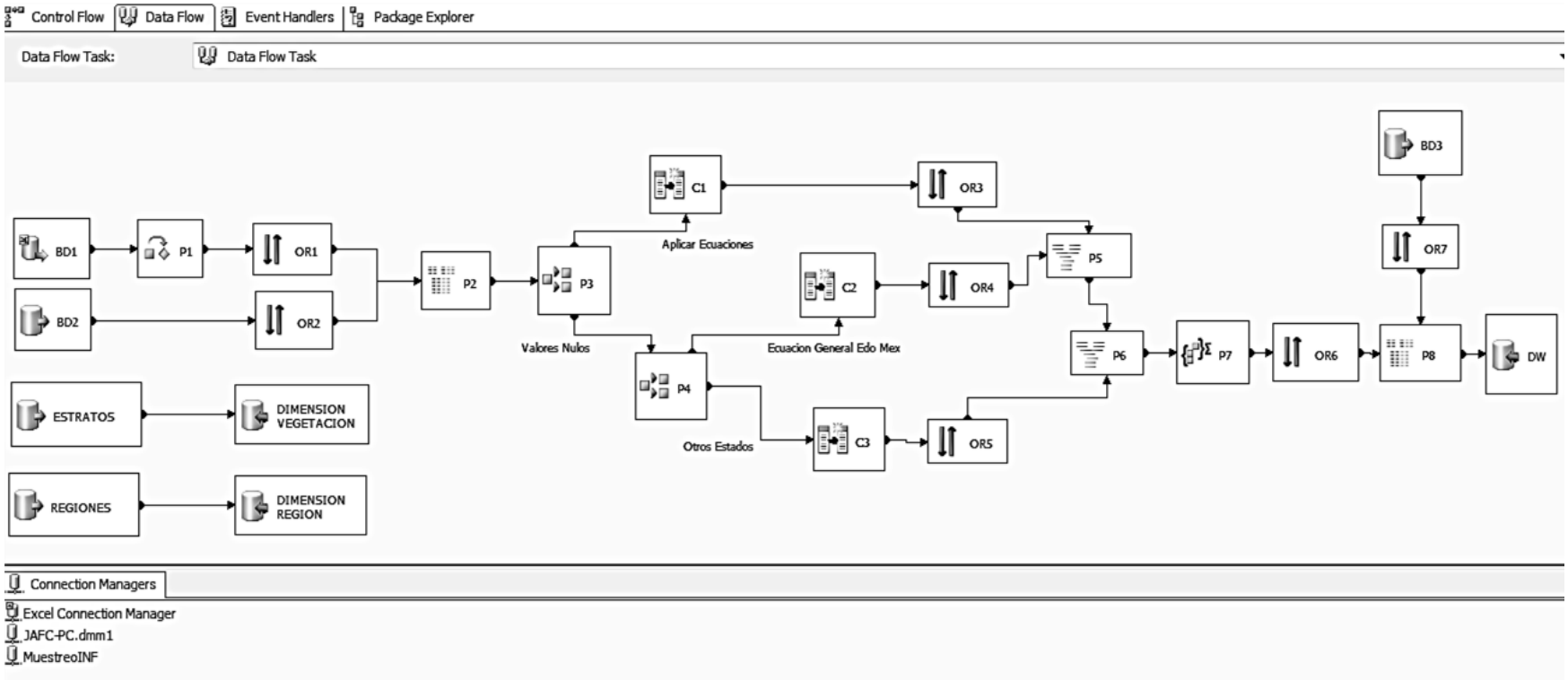


Figura 54. Flujo de datos del Paquete de Integración.
Elaboración: Propia para la investigación.

ordenación, reunión externa y funciones de agregación. Las tareas y funciones que se encuentran implícitas en los objetos se definen a continuación.

- El objeto **BD1**, crea una conexión hacia un archivo en Microsoft Excel que contiene la información de las ecuaciones de volumen aplicables a algunos géneros arbóreos del Estado de México, obtenidas de Méndez y De los Santos (2011).
- El Objeto **P1**, realiza una transformación a los datos correspondientes al campo género para garantizar la compatibilidad del dominio entre los datos provenientes del archivo de Excel y los provenientes de la base de datos del inventario.
- El objeto **OR1**, sólo realiza una ordenación, por género, de los datos provenientes del objeto unificación de dominio.
- El objeto **BD2**, crea una conexión hacia la base de datos del inventario, además de tener implícita una instrucción SQL que extrae la información relevante desde diferentes tablas y con diferentes restricciones. En el Cuadro 16 se encuentra el comando SQL empleado para extraer los datos así como las restricciones usadas por la CONAFOR en su reporte de resultados. La última restricción en la cláusula condicional hace referencia sólo a la vegetación de bosque y selva debido a que son los dos ecosistemas sobre los cuales recae las estimaciones de volumen de madera, biomasa y carbono.

Cuadro 16. Comando SQL usado para extraer los datos para la tabla de hechos.

```
SELECT      IdArboladoBosqueSelva, TblSitio.IdSitio, Genero, DiametroNormal, AlturaTotal, Estado

FROM        ((TblArboladoBosqueSelva INNER JOIN TblSitio ON TblArboladoBosqueSelva.IdSitio =
TblSitio.IdSitio) INNER JOIN TblConglomerado ON TblSitio.IdConglomerado =
TblConglomerado.IdConglomerado)

WHERE       Condicion in (1,2) AND (DiametroNormal >= 7.5) AND (DiametroNormal <= 132.5) AND
(AlturaTotal >= 5) AND (AlturaTotal <= 47.5) AND          (TipoVegetacionLev in (select
IdCveTipoVegetacion from CatTipoVegetacionInegiGeneral  where CveTipoVegetacion IN
('Bosque','Selva')))
```

Elaboración: Propia para la investigación.

- El objeto **OR2**, realiza una ordenación por género, de los datos extraídos de la base de datos del inventario.
- El objeto **P2**, realiza la operación reunión externa por la izquierda para juntar en una sola tabla la información del arbolado con sus respectivas ecuaciones de volumen. Ésta

operación conserva toda la información de la tabla que se encuentra del lado izquierdo (los datos proveniente de la base de datos del inventario) y llena con valore nulos el lado derecho de la tabla que después de aplicar una reunión natural, no se encuentren relacionados.

- El objeto **P3**, realiza una operación condicional que separa en otra tabla, los datos que contienen coeficientes con valores nulos.
- El objeto **C1**, agrega una columna calculada, de volumen de madera, a partir de los datos dasométricos del arbolado y su respectiva ecuación de volumen.
- El objeto **C2**, agrega una nueva columna calculada, de volumen de madera, aplicando la ecuación de volumen general para aquellos géneros arbóreos que no cuentan con una ecuación de volumen específica, para el Estado de México.
- El objeto **P4**, separa la información del Estado de México de los otros Estados.
- El objeto **C3**, agrega una nueva columna calculada, de volumen de madera, que utiliza la ecuación de volumen general para todos los estados.
- El Objeto **OR3** ordena por género los datos provenientes del objeto C1.
- El objeto **OR4** ordena por género los datos provenientes del objeto C2.
- El objeto **OR5** ordena por género los datos provenientes del objeto C3.
- El objeto **P5**, realiza una operación de unión de los datos del Estado de México.
- El objeto **P6**, realiza una operación de unión de los datos del Estado de México con los datos del resto de los estados.
- El objeto **P7**, realiza una operación de agregación de datos agrupándolos por sitio.
- El objeto **BD3**, extra la información de la vegetación encontrada en cada sitio para incorporarla como información adicional a la información agregada con el objeto anterior. Cabe mencionar que en cada sitio sólo se registra un tipo de vegetación de acuerdo a un catálogo incorporado que contiene la información de todos los tipos de vegetación del país. La instrucción SQL correspondiente se muestra en el Cuadro 17.

Cuadro 17. Comando SQL para extraer la información de vegetación en sitios de muestreo.

```
select IdSitio,TipoVegetacionLev from TblSitio
```

Elaboración: Propia para la investigación.

- El objeto **OR6**, ordena por sitio los datos agrupados por el objeto P7.
- El objeto **OR7** ordena por sitio los datos extraídos mediante el objeto BD3.
- El objeto **P8**, lleva a cabo una reunión natural, de la información agregada por sitios con su respectiva vegetación.
- El objeto **DW**, corresponde al almacenamiento de la información en la fuente de destino (Data Warehouse); la tarea se realiza mediante una conexión implícita en este objeto con la tabla de hechos presente en el Data Warehouse de análisis. En la Figura 55 se puede observar la relación existente entre la información procesada mediante las diferentes tareas del objeto del paquete y su destino en el Data Warehouse.

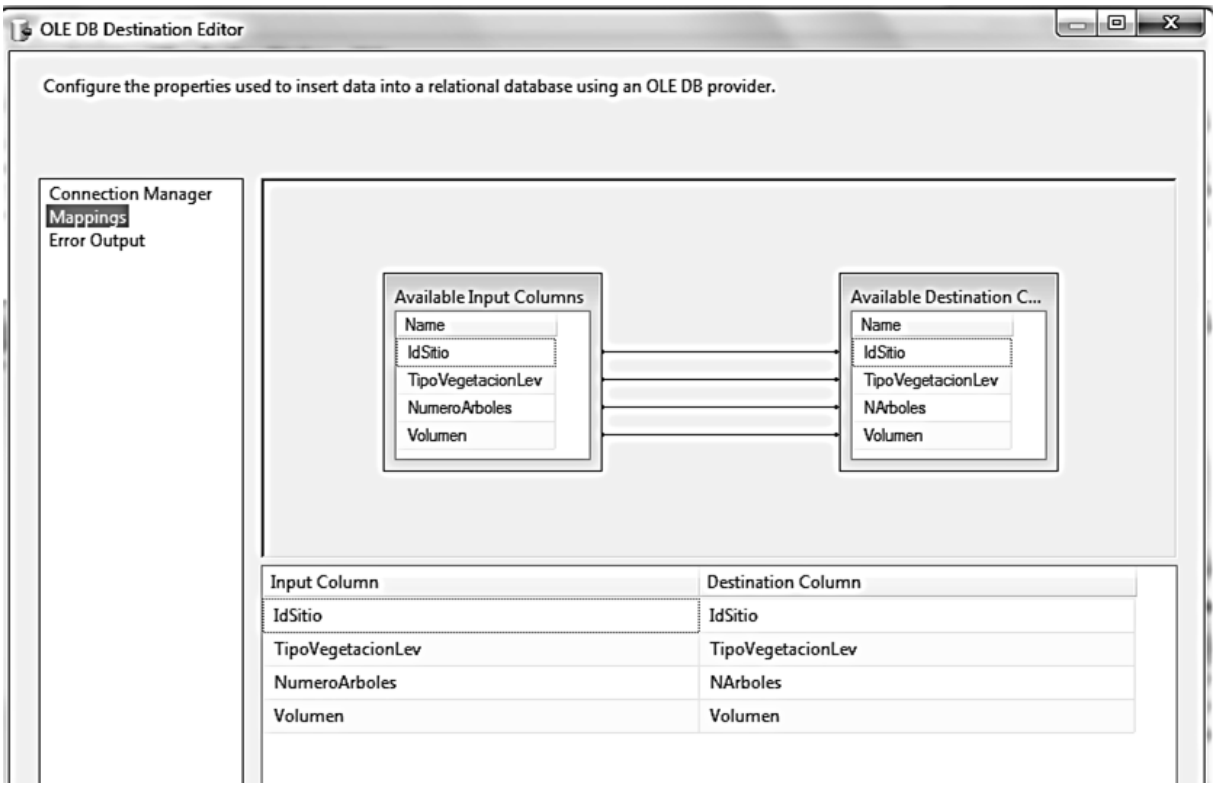


Figura 55: Relación entre los campos de los datos procesados y la tabla de hechos.
Elaboración: Propia para la investigación.

11.2.2 Integración de la dimensión Vegetación

El proceso para extraer la información desde la base de datos del inventario y cargarla en la tabla de dimensión **Vegetación** en el DW, se realiza mediante los objetos etiquetados como **ESTRATOS** y **DIMENSION VEGETACION**, localizados en la parte inferior izquierda del proceso ETL en la Figura 56.

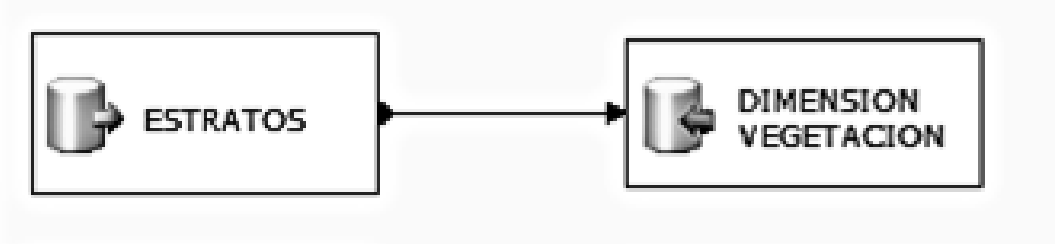


Figura 56: Objetos relacionados en el proceso ETL para la dimensión Vegetación.
Elaboración: Propia para la investigación.

En la Figura 56, el objeto etiquetado como **ESTRATOS** representa una conexión hacia la base de datos del inventario el cual emplea un comando SQL para extraer la información requerida por la dimensión **Vegetación**. El comando SQL utilizado se puede analizar en el Cuadro 18.

Cuadro 18. Comando SQL usado para extraer los datos para la dimensión "Vegetación".

```
SELECT DISTINCT  CatTipoVegetacionInegiGeneral.IdCveTipoVegetacion,
CatTipoVegetacionInegiGeneral.NomComunidad,
CatTipoVegetacionInegiGeneral.CveTipoVegetacion

FROM            ((TblArboladoBosqueSelva INNER JOIN TblSitio ON
TblArboladoBosqueSelva.IdSitio = TblSitio.IdSitio) INNER JOIN
CatTipoVegetacionInegiGeneral ON TblSitio.TipoVegetacionLev =
CatTipoVegetacionInegiGeneral.IdCveTipoVegetacion)

WHERE        (TblArboladoBosqueSelva.Condicion IN (1, 2)) AND
TblArboladoBosqueSelva.DiametroNormal >= 7.5) AND
(TblArboladoBosqueSelva.DiametroNormal <= 132.5) AND
(TblArboladoBosqueSelva.AlturaTotal >= 5) AND
(TblArboladoBosqueSelva.AlturaTotal <= 47.5) AND
(TblSitio.TipoVegetacionLev IN (SELECT   IdCveTipoVegetacion  FROM
CatTipoVegetacionInegiGeneral  WHERE        (CveTipoVegetacion IN
('Bosque', 'Selva'))))
```

Elaboración: Propia para la investigación.

11.2.3 Integración de la dimensión Región

El proceso para extraer la información desde la base de datos del inventario y cargarla en la tabla de dimensión **Región** en el DW, se realiza mediante los objetos etiquetados como **Regionalización** y **Dimensión Región** (Figura 57), ubicado debajo del proceso ETL de la dimensión **Vegetación**.

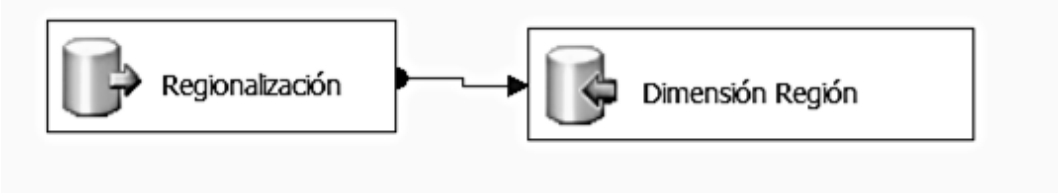


Figura 57: Objetos relacionados en el proceso ETL para la dimensión Región.
Elaboración: Propia para la investigación.

En el llenado de la tabla de dimensión **Región** se garantiza que la cardinalidad uno a muchos, de ésta tabla hacia a la tabla de hechos se mantenga; para cumplir con dicha restricción sólo se selecciona la información no repetida sobre los sitios para cada registro en la tabla de hechos. El comando SQL implícito en el objeto **Regionalización** se puede observar en el Cuadro 19.

Cuadro 19. Comando SQL usado para extraer los datos para la dimensión "Región".

```
SELECT DISTINCT TblSitio.IdSitio, TblConglomerado.IdConglomerado,
CatMunicipio.CveMunicipio,  CatMunicipio.NomMunicipio,
CatEstado.CveEstado, CatEstado.NomEstado

FROM TblArboladoBosqueSelva, CatTipoVegetacionInegiGeneral,
TblSitio, TblConglomerado, CatEstado, CatMunicipio

WHERE Condicion in (1,2) AND (DiametroNormal >= 7.5) AND
(DiametroNormal <= 132.5) AND
(AlturaTotal >= 5) AND (AlturaTotal <= 47.5) AND (TipoVegetacionLev
in (select IdCveTipoVegetacion from CatTipoVegetacionInegiGeneral
where CveTipoVegetacion IN ('Bosque','Selva'))) ANDTblSitio.IdSitio
= TblArboladoBosqueSelva.IdSitio AND TipoVegetacionLev =
IdCveTipoVegetacion AND TblSitio.IdConglomerado =
TblConglomerado.IdConglomerado AND
Estado = CatEstado.CveEstado AND Municipio =
CatMunicipio.CveMunicipio
```

Elaboración: Propia para la investigación.

En la Figura 58 se observa la relación entre los campos generados con la instrucción SQL del Cuadro 20.

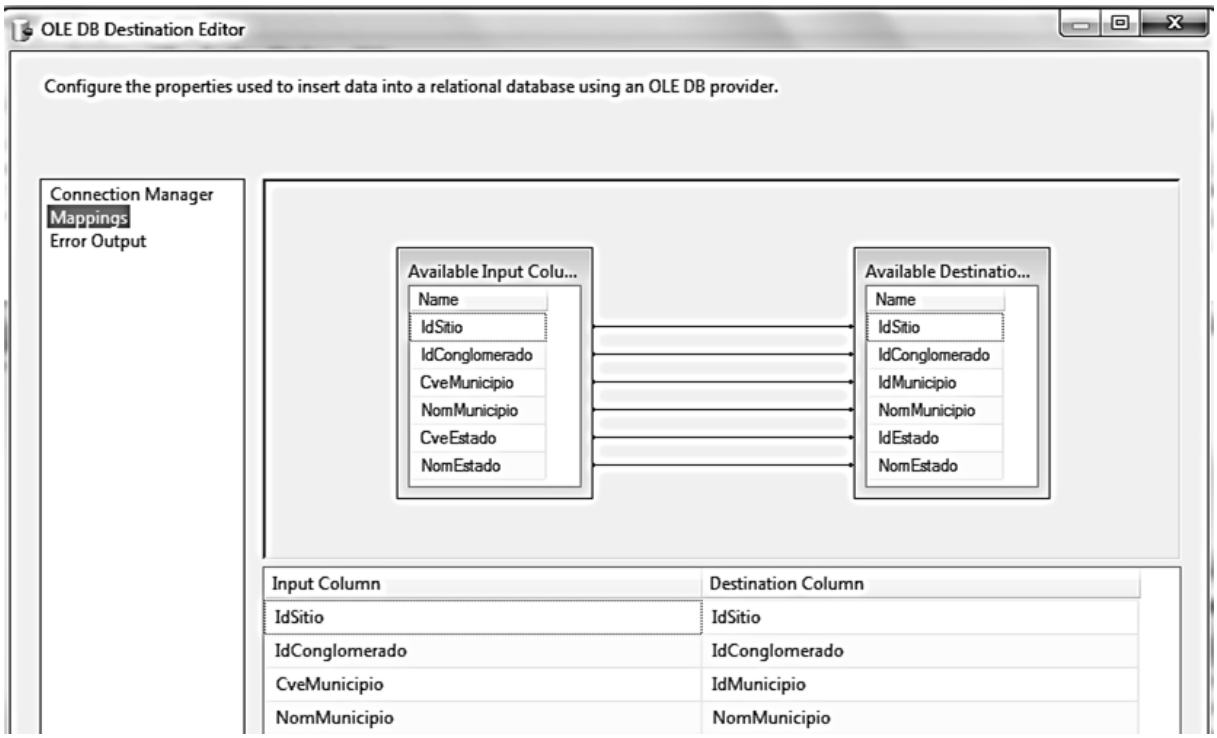


Figura 58: Relación entre los campos de la fuente de origen y la tabla de dimensión “Región”.
Elaboración: Propia para la investigación.

Una vez descrito el funcionamiento de los objetos involucrados en la tarea de carga de datos, en la siguiente sección se detalla, paso a paso, la ejecución de las tareas.

11.2.4 Ejecución de la tarea de carga de datos

La tara de integración se puede llevar a cabo mediante la aplicación, desarrollada en la presente tesis, haciendo clic en el botón Cargar Datos. (Figura 57).

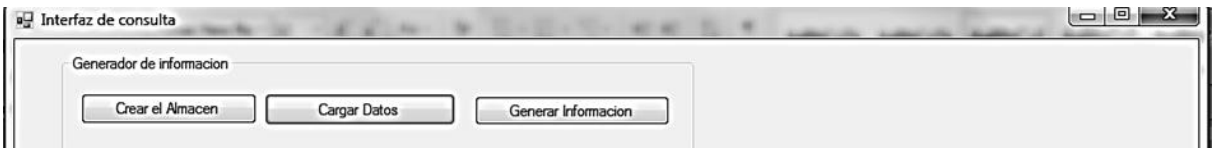


Figura 59: Carga de datos mediante aplicación.
Elaboración: Propia para la investigación.

En SSIS, es necesario abrir el proyecto Integración desde Access que es del tipo Microsoft Visual Studio Solution, que se encuentra en la dirección C:\proyecto\Integracion desde Access.

Para poner en marcha la tarea sólo hay que hacer clic en la pestaña generar (Build en la Figura 60, si la instancia de SQL instalada está en inglés) y seleccionar Generar Integración desde Access.

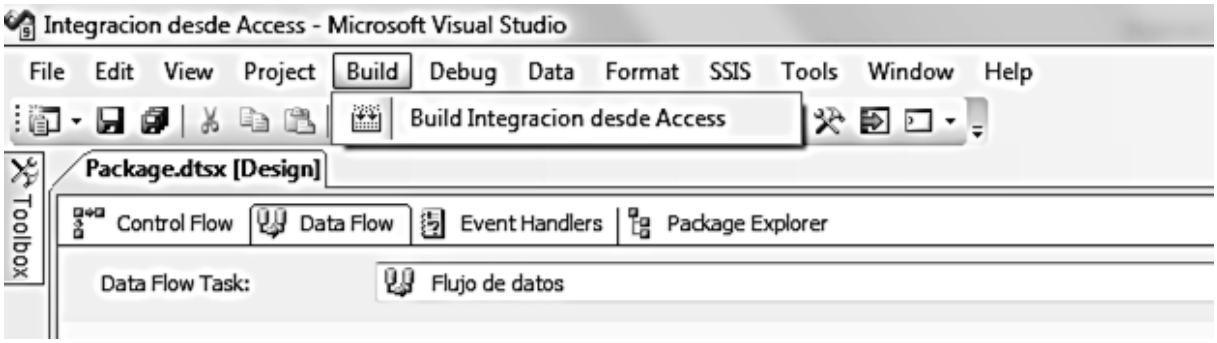


Figura 60: Ejecución del paquete de integración usando SSIS.
Elaboración: Propia para la investigación.

En la pestaña Flujo de datos se puede observar, durante la ejecución, que alguno de éstos procesos se ejecutan en paralelo. En la pestaña progreso se puede ver las estadísticas de la ejecución que incluye el tiempo trascurrido y la cantidad de datos procesados.

11.3 El paquete de análisis de información

El paquete de análisis es el encargado de llevar a cabo el procesamiento de la información que se encuentra en el Data Warehouse y generar los cubos para el análisis de la misma. Este paquete fue desarrollado usando las herramientas del servicio de análisis (SSAS), que contiene una serie de objetos para procesar grandes volúmenes de datos de manera eficiente. En esta sección se detallará cada uno de los objetos involucrados en el análisis multidimensional de los datos del inventario.

11.3.1 Objetos del paquete de análisis

Los objetos que integran el paquete de análisis se encuentran definidos en el explorador de soluciones, tal y como se muestra en la Figura 61.

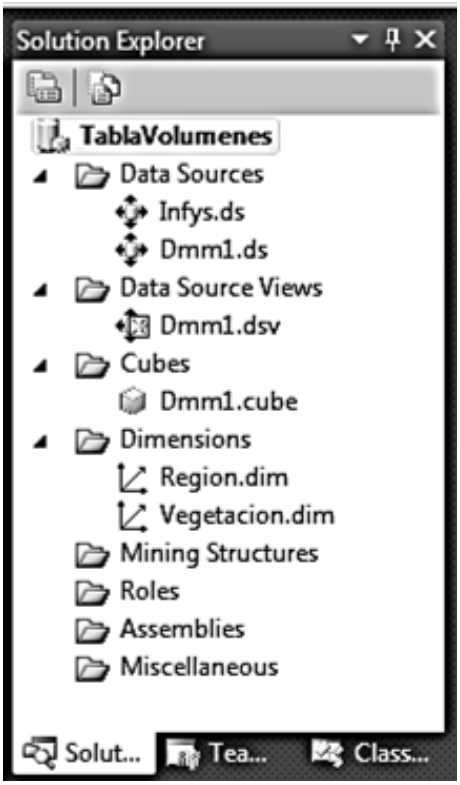


Figura 61: Explorador de soluciones del paquete de análisis de datos.
Elaboración: Propia para la investigación.

Cada uno de estos objetos realiza una tarea particular las cuales de se describen a continuación.

Orígenes de datos (Data Sources): En este se especifica la conexión al Data Warehouse.

Vistas de origen de datos (Data Sources Views): Se definen las vistas, generadas a partir de las tablas del Data Warehouse, que se usan en el proyecto.

Cubos (Cubes): Se define el objeto volumen (cubo de dos dimensiones) que muestra la información agregada de las estimaciones de volumen de madera, biomasa y carbono.

Dimensiones (Dimensions): Corresponde a las dimensiones definidas para la agregación de datos (Vegetación y Región).

11.3.2 Configuración del Servicio de Análisis

SQL Server, al igual que muchos sistemas gestores de bases de datos, contempla dentro de sus funciones la capacidad para administrar a los usuarios de las bases de datos; utiliza diferentes métodos de autentificación para los usuarios y define roles específicos con diferentes tareas y permisos. Estas tareas y permisos pueden ser, por ejemplo, realizar consultas y modificaciones a las tablas de la base de datos. Antes de poder procesar los cubos de análisis definidos en éste paquete, primero es necesario conceder los permisos a la herramientas de análisis para que pueda acceder a la información que se encuentra en las base de datos en SQL Server.

La configuración que se usa, para establecer la conexión, es la de suplantación de origen de datos; en esta se especifica el contexto de seguridad bajo el que se importan o se procesan los datos. De forma predeterminada, los valores de suplantación especifican la cuenta de servicio de SSAS para obtener acceso a los datos. Para usar esta configuración predeterminada, es necesario que la cuenta de servicio bajo la que se ejecuta SSAS tiene permisos de lectura para la base de datos dmm1.

Los pasos para otorgar el permiso de lectura a partir de SSAS se detallan a continuación o se pueden consultar en la siguiente liga (http://msdn.microsoft.com/es-es/library/hh403424.aspx).

En primer lugar hay que determinar la cuenta de servicio, para ello es necesario acceder al Administrador de configuración de SQL Server o la aplicación de consola, Servicios, para ver la información de las cuentas. Si durante la instalación se seleccionó SSAS como instancia predeterminada, entonces el servicio de análisis se ejecuta bajo el nombre de NT Service\MSSQLServerOLAPService. Las instrucciones necesarias para configurar el servicio en dado caso de que no se halla establecido desde un principio, de enumeran en la siguiente lista.

1. Conectar a la instancia del motor de base de datos.

2. Expandir la carpeta Seguridad, haciendo clic con el botón secundario en Inicios de sesión y seleccione Nuevo inicio de sesión. (Figura 62)

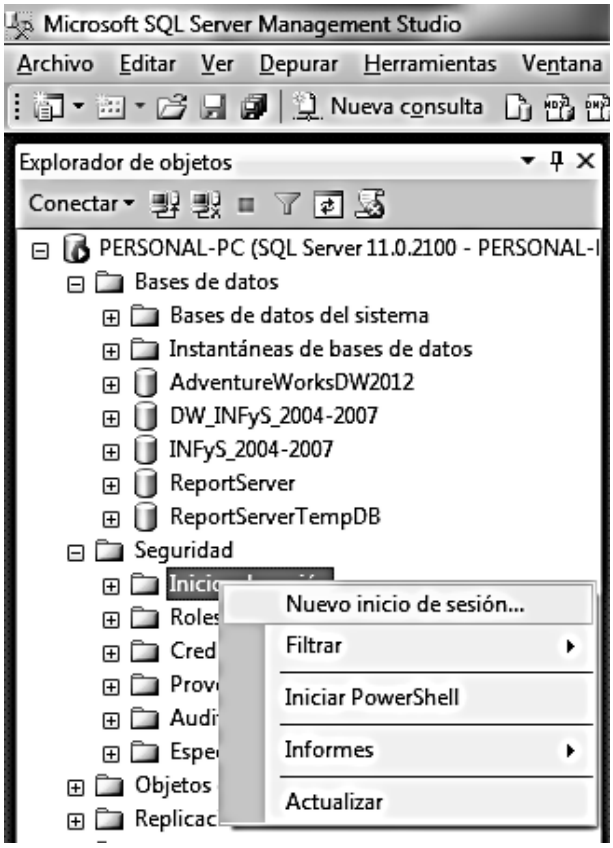


Figura 62: Creación de un nuevo inicio de sesión.
Elaboración: Propia para la investigación.

3 Escribir NT Service\MSSQLServerOLAPService (o la cuenta en la que se ejecuta el servicio, ver Figura 63) en Nombre de inicio de sesión.

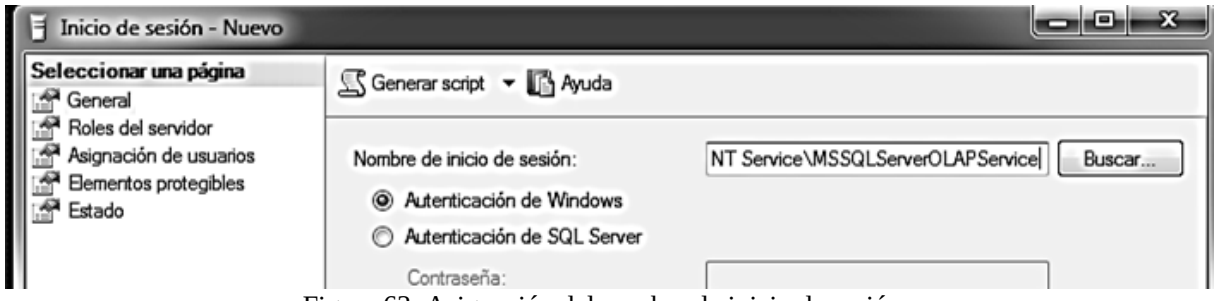


Figura 63: Asignación del nombre de inicio de sesión.
Elaboración: Propia para la investigación.

4 Hacer clic en Asignación de usuarios, en el menú del lado izquierdo.

5 Activar la casilla situada al lado de la base de datos dmm1. En este rol, se debe tener los privilegios de db_datareader y público (Ver Figura 64), o concederlos activando la casilla a su izquierda. Haga clic en Aceptar para aceptar para guardar los cambios.

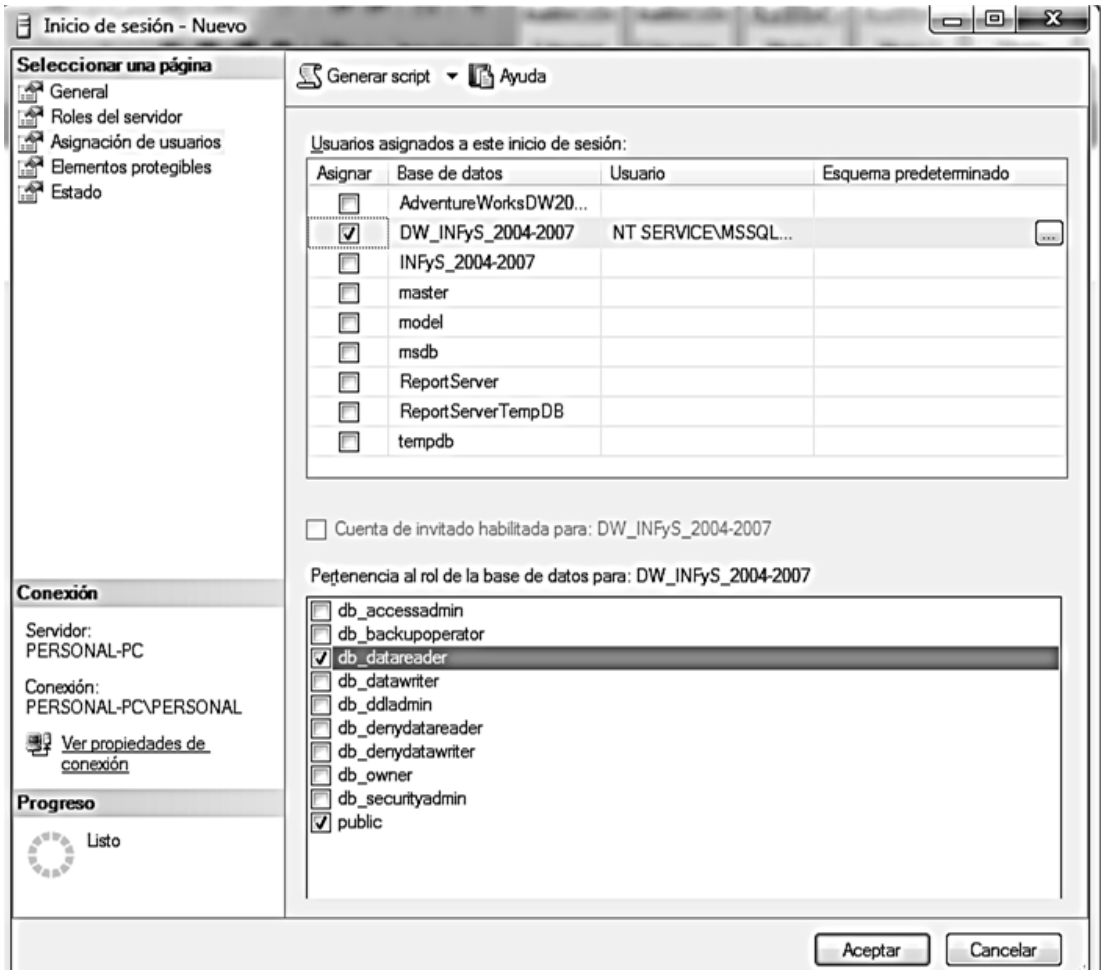


Figura 64: Asignación de usuarios y permisos.
Elaboración: Propia para la investigación.

11.3.3 Procesamiento de los cubos.

La información de interés no se encuentra disponible para consulta; primero se debe llevar a cabo el procesamiento de la información; lo cual consiste en rellenar los atributos de cada objeto accediendo a la información contenida en el Data Warehouse. De manera automática este procesado de información se realizar al ejecutar la tarea de Integración de datos, pero bajo

algunas circunstancias de configuración durante la instalación de SQL Server, no se definen los permisos para acceder al Data Warehouse y el cubo no será procesa.

Para llevar a cabo el procesamiento, de forma manual de la información, siga los siguientes pasos.

1. Hacer clic con el botón secundario sobre Volumen.cube y selecciones abrir.
2. Hacer clic en la pestaña examinador y se verá una pantalla como el de la Figura 65.

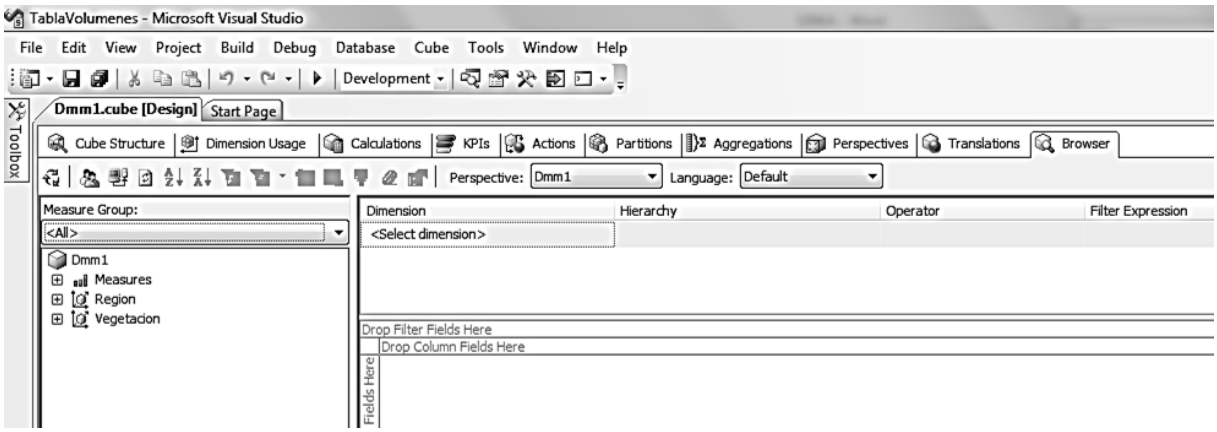


Figura 65: Ventana de diseño del cubo (pestaña examinador).
Elaboración: Propia para la investigación.

3. Hacer clic en el botón proceso que tiene forma de flechas verdes en espiral, que se encuentra debajo de la pestaña Estructura del cubo (parte superior izquierda de la forma), o en la pestaña generar y luego proceso (Figura 66).

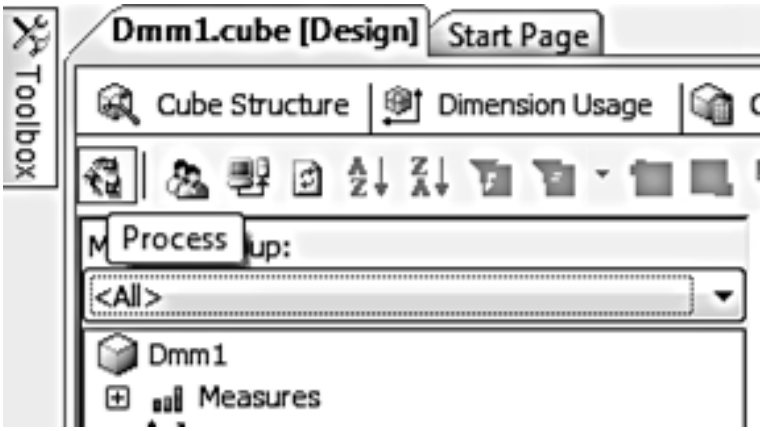


Figura 66: Procesamiento de los cubos de información
Elaboración: Propia para la investigación.

4. Oprimir el botón ejecutar en la ventana que se despliega y se esperar a que termine el procesamiento del cubo. Finalmente, oprima el botón cerrar en cerrar en cada una de las dos ventanas emergentes y la información ya se encontrará disponible para su análisis.

11.3.4 Exportación de los datos a Excel.

La opción de migra la información procesada mediante cubo OLAP no está disponible para la versión 2008 de MS SQL Server, ésta opción se encuentra disponible para las versiones de 2012 en delante; pero cualquier paquete desarrollado en SQL Server 2008 es compatible con todas las versiones que le suceden.

La información procesada en los cubos puede migrarse a un archivo xlsx (Excel) haciendo clic en el icono con forma de hoja de cálculo como se indica en la Figura 67.

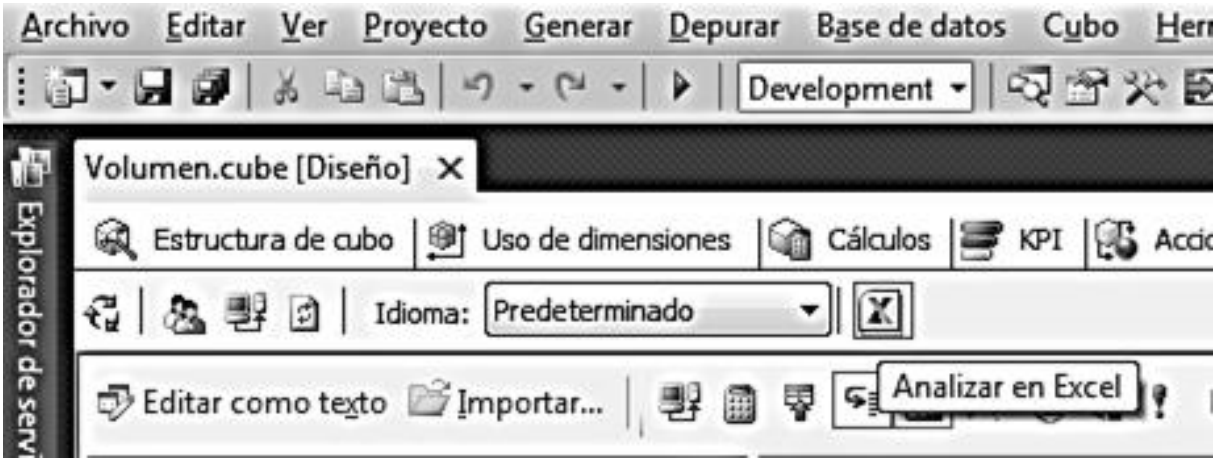


Figura 67: Migración de la información procesada a un archivo de Microsoft Excel.
Fuente: Elaboración propia para la investigación.

En el recuadro que aparece cuando se abre el libro de Excel debe hacer clic en habilitar para establecer la conexión. Con esto esta acción se crea un vínculo entre la hoja de cálculo y la base datos con las dimensiones ya procesadas.

11.4 Resumen

En este capítulo se describió los pasos para implementar el Data Warehouse para el análisis de volumen de madera, biomasa y carbono.

En primer lugar, se explicó paso a paso la construcción del Data Warehouse mediante el asistente de SQL Server 2008 y mediante consulta usando un comando SQL.

En segundo lugar se describió el paquete de integración de datos, con el cual se realiza la tarea conocida como extracción, transformación y carga, que consiste en un proceso automático para recuperar datos desde una o más fuentes de información hacia las tablas que conforman el Data Warehouse. También se describieron los objetos que lo conforman y la tarea específica que realizan. El paquete fue credo con el proyecto del servicio de integración de SQL Server 2008.

En tercer lugar se describió el paquete de análisis de datos, el cual se encarga de llevar a cabo el procesamiento de la información que se encuentra en el Data Warehouse y generar los cubos para el análisis de la misma. Este paquete fue desarrollado usando las herramientas del servicio de análisis de SQL Server.

12. INTERFACES PARA ANÁLISIS DE VOLUMEN DE MADERA, BIOMASA Y CARBONO

En la presente sección se ejemplifica el uso de diferentes interfaces, para el análisis y visualización de la información procesada mediante el cubo OLAP. La interfaz principal corresponde al examinador de datos que es una herramienta gráfica de consultas, integrado al servicio de análisis de SQL Server (SSAS). La segunda interfaz es un archivo de MS Excel, generado a partir del cubo, con el objetivo de explotar el potencial de las tablas y gráficos dinámicos. Las dos interfaces restantes fueron desarrolladas para servir como intermediario entre los usuarios finales y la información del cubo; se desarrollaron una interfaz de escritorio y una interfaz web que emulan al examinador de datos de SSAS. Todas las interfaces anteriores comparte la característica de permitir un análisis dinámico de los datos en tiempo relativamente menores, comparado contra una consulta tradicional en SQL. En la Figura 68 se muestran las interfaces a utilizar para la visualización y análisis de la información.

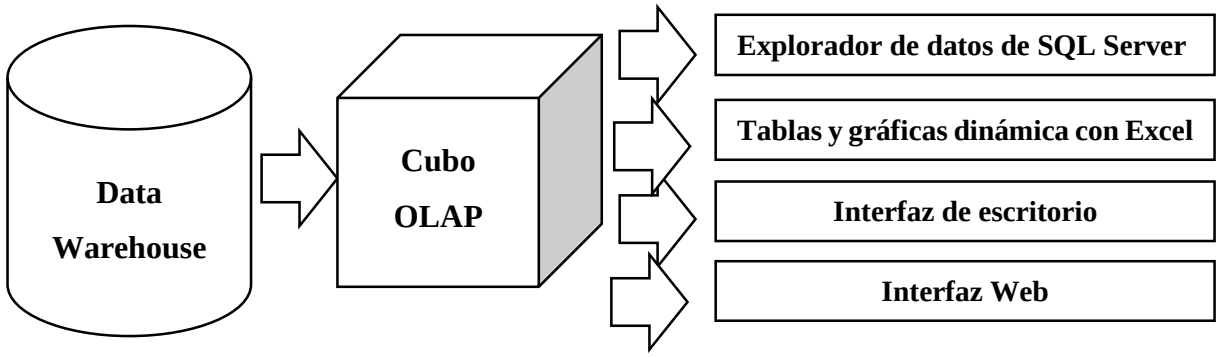


Figura 68. Interfaces para la visualización y análisis de la información
Fuente: Elaboración propia para la investigación.

La información sujeta de análisis, como se establece en uno de los objetivos, corresponde a los cálculos de volumen de madera, biomasa y carbono; adicionalmente se muestra la información del número de árboles usados en el cálculo y el área medida en hectáreas que es usada como variable auxiliar al momento de realizar las estimaciones del volumen de madera, biomasa y carbono, sobre un área determinada. En las siguientes subsecciones se realiza un análisis de los datos, usando las interfaces mencionadas en éste párrafo; los análisis realizados con cada

interfaz son iguales con el objetivo de obtener los mismos resultados; los análisis de datos agregados a nivel de ecosistema y comunidad vegetal son a nivel nacional a menos que se especifique que es para el Estado de México.

12.1 Análisis de datos mediante el examinador del servicio de integración

La información procesada mediante un cubo OLAP puede ser accedida de diferentes formas, la más cómoda para realizar un análisis rápido de la información es mediante el explorador de datos integrados como herramienta del servicio de integración de SQL Server 2008.

El explorador de datos puede ser accedido desde la ventana explorador y consiste en un área dinámica para la exploración de datos; su principal característica es que permite realizar las operaciones básicas aplicadas a los cubos OLAP como por ejemplo: drill Down, roll up y pivotaje, adicionalmente incluye un área para filtrado de información para obtener las operaciones de Slice y Dice. En la Figura 69 se muestra el explorador de datos con la información agregada a nivel de ecosistemas.

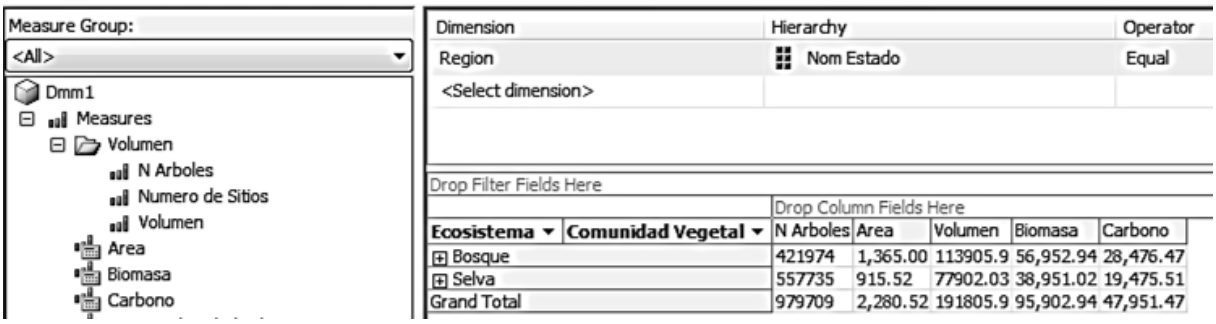
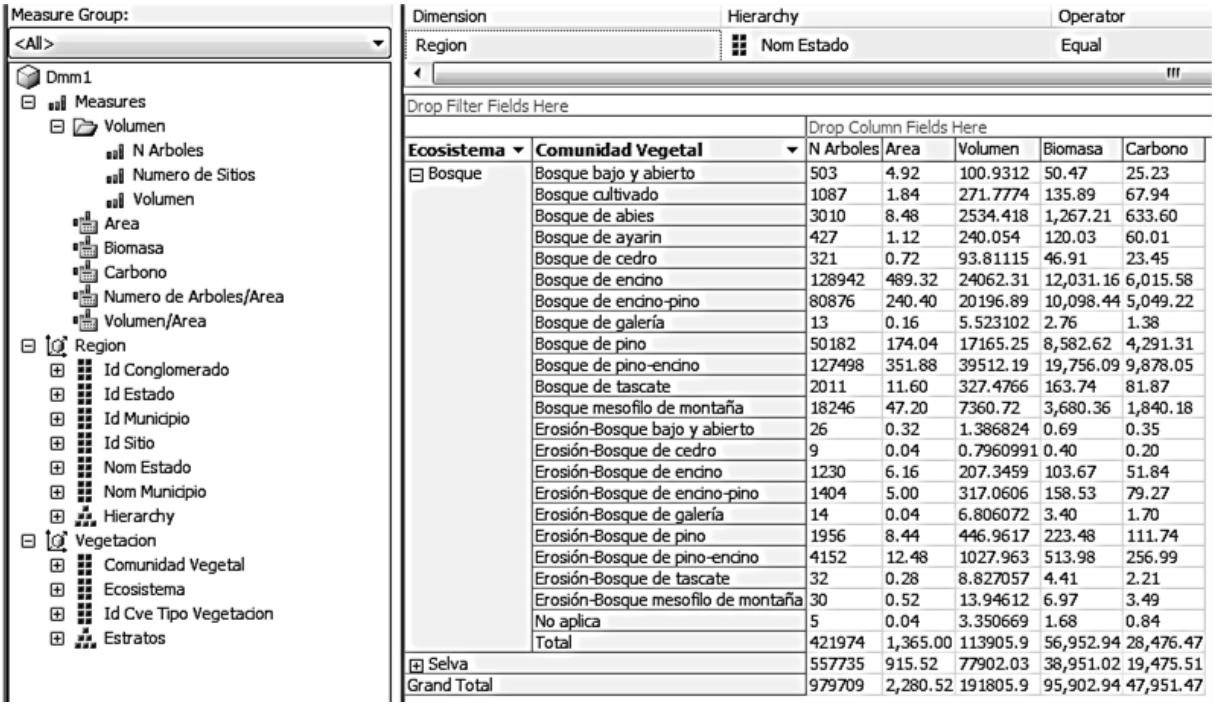


Figura 69: Análisis de información agregada ecosistema, mediante el examinador del servicio de análisis de SQL Server.
Fuente: Elaboración propia para la investigación

El análisis de los datos se realiza al llevar las medidas y campos de las dimensiones (o jerarquías completas) desde el grupo de medidas hacia el área de columnas, filas o medidas del examinador de datos; literalmente se arrastran los objetos hasta el área de interés. En la Figura 70 se muestra

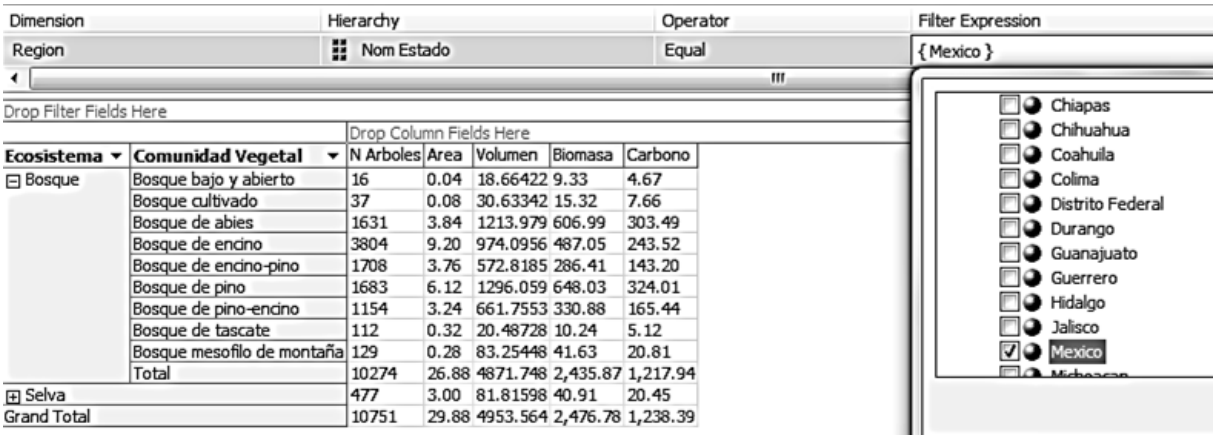
la misma información con un mayor nivel de detalle, agregado primero a nivel de comunidad vegetal y después a nivel de ecosistema.



Ecosistema	Comunidad Vegetal	N Arboles	Area	Volumen	Biomasa	Carbono
Bosque	Bosque bajo y abierto	503	4.92	100.9312	50.47	25.23
	Bosque cultivado	1087	1.84	271.7774	135.89	67.94
	Bosque de abies	3010	8.48	2534.418	1,267.21	633.60
	Bosque de ayarin	427	1.12	240.054	120.03	60.01
	Bosque de cedro	321	0.72	93.81115	46.91	23.45
	Bosque de encino	128942	489.32	24062.31	12,031.16	6,015.58
	Bosque de encino-pino	80876	240.40	20196.89	10,098.44	5,049.22
	Bosque de galería	13	0.16	5.523102	2.76	1.38
	Bosque de pino	50182	174.04	17165.25	8,582.62	4,291.31
	Bosque de pino-encino	127498	351.88	39512.19	19,756.09	9,878.05
	Bosque de tascate	2011	11.60	327.4766	163.74	81.87
	Bosque mesofilo de montaña	18246	47.20	7360.72	3,680.36	1,840.18
	Erosión-Bosque bajo y abierto	26	0.32	1.386824	0.69	0.35
	Erosión-Bosque de cedro	9	0.04	0.7960991	0.40	0.20
	Erosión-Bosque de encino	1230	6.16	207.3459	103.67	51.84
	Erosión-Bosque de encino-pino	1404	5.00	317.0606	158.53	79.27
	Erosión-Bosque de galería	14	0.04	6.806072	3.40	1.70
	Erosión-Bosque de pino	1956	8.44	446.9617	223.48	111.74
	Erosión-Bosque de pino-encino	4152	12.48	1027.963	513.98	256.99
	Erosión-Bosque de tascate	32	0.28	8.827057	4.41	2.21
	Erosión-Bosque mesofilo de montaña	30	0.52	13.94612	6.97	3.49
	No aplica	5	0.04	3.350669	1.68	0.84
	Total	421974	1,365.00	113905.9	56,952.94	28,476.47
Selva		557735	915.52	77902.03	38,951.02	19,475.51
Grand Total		979709	2,280.52	191805.9	95,902.94	47,951.47

Figura 70: Análisis de información agregada por comunidad vegetal y ecosistema, mediante el examinador del servicio de análisis de SQL Server.
Fuente: Elaboración propia para la investigación.

En la Figura 71 se muestra la información agregada a nivel de comunidad vegetal, luego a nivel de ecosistema y además filtrada sólo para el Estado de México.



Ecosistema	Comunidad Vegetal	N Arboles	Area	Volumen	Biomasa	Carbono
Bosque	Bosque bajo y abierto	16	0.04	18.66422	9.33	4.67
	Bosque cultivado	37	0.08	30.63342	15.32	7.66
	Bosque de abies	1631	3.84	1213.979	606.99	303.49
	Bosque de encino	3804	9.20	974.0956	487.05	243.52
	Bosque de encino-pino	1708	3.76	572.8185	286.41	143.20
	Bosque de pino	1683	6.12	1296.059	648.03	324.01
	Bosque de pino-encino	1154	3.24	661.7553	330.88	165.44
	Bosque de tascate	112	0.32	20.48728	10.24	5.12
	Bosque mesofilo de montaña	129	0.28	83.25448	41.63	20.81
	Total	10274	26.88	4871.748	2,435.87	1,217.94
Selva		477	3.00	81.81598	40.91	20.45
Grand Total		10751	29.88	4953.564	2,476.78	1,238.39

Figura 71: Análisis de información agregada por comunidad vegetal y ecosistema para el Estado de México, mediante el examinador del servicio de análisis de SQL Server.
Fuente: Elaboración propia para la investigación.

La ventaja poder intercambiar filas y columnas para realizar un análisis, se muestra en la Figura 72, donde la información se encuentra agregada por estados en el área de filas, y en el área de las columnas se encuentra agregada por ecosistemas.

Dimension	Hierarchy	Operator	Filter Expression
Region	Nom Estado	Equal	{ All }
<Select dimension>			

Drop Filter Fields Here

	Ecosistema														
	Bosque					Selva					Grand Total				
Nom Estado	N Arboles	Area	Volumen	Biomasa	Carbono	N Arboles	Area	Volumen	Biomasa	Carbono	N Arboles	Area	Volumen	Biomasa	Carbono
Aguascalientes	827	2.72	80.30193	40.15	20.08	160	0.48	7.742883	3.87	1.94	987	3.20	88.04481	44.02	22.01
Baja California	287	2.72	296.157	148.08	74.04						287	2.72	296.157	148.08	74.04
Baja California Sur	281	1.24	58.11024	29.06	14.53	1069	4.16	95.62499	47.81	23.91	1350	5.40	153.7352	76.87	38.43
Campeche						142222	179.36	17699.16	8,849.58	4,424.79	142222	179.36	17699.16	8,849.58	4,424.79
Chiapas	17609	59.48	5589.389	2,794.69	1,397.35	19469	52.08	5816.456	2,908.23	1,454.11	37078	111.56	11405.84	5,702.92	2,851.46
Chihuahua	93550	309.92	19045.38	9,522.69	4,761.35	1453	5.60	143.4715	71.74	35.87	95003	315.52	19188.86	9,594.43	4,797.21
Coahuila	3712	16.56	570.8999	285.45	142.72						3712	16.56	570.8999	285.45	142.72
Colima	605	1.56	252.3163	126.16	63.08	3357	7.44	388.7438	194.37	97.19	3962	9.00	641.0601	320.53	160.27
Distrito Federal	542	2.32	438.3289	219.16	109.58						542	2.32	438.3289	219.16	109.58
Durango	74667	226.48	19508.18	9,754.09	4,877.04	2237	6.24	274.6264	137.31	68.66	76904	232.72	19782.8	9,891.40	4,945.70
Guanajuato	5540	18.28	709.7272	354.86	177.43	317	1.84	26.42251	13.21	6.61	5857	20.12	736.1497	368.07	184.04
Guerrero	18198	83.96	8505.084	4,252.54	2,126.27	8323	29.28	1150.218	575.11	287.55	26521	113.24	9655.303	4,827.65	2,413.83
Hidalgo	6885	18.64	1644.8	822.40	411.20	1148	4.12	202.104	101.05	50.53	8033	22.76	1846.904	923.45	461.73
Jalisco	34354	114.04	10393.87	5,196.93	2,598.47	15514	41.80	2116.552	1,058.28	529.14	49868	155.84	12510.42	6,255.21	3,127.61
Mexico	10274	26.88	4871.748	2,435.87	1,217.94	477	3.00	81.81598	40.91	20.45	10751	29.88	4953.564	2,476.78	1,238.39
Michoacan	21876	62.40	8334.126	4,167.06	2,083.53	9168	26.72	1110.398	555.20	277.60	31044	89.12	9444.523	4,722.26	2,361.13

Figura 72: Análisis de información agregada por estados y por ecosistemas mediante el examinador de datos del servicio de análisis de SQL Server.
Fuente: Elaboración propia para la investigación.

El examinador es una interfaz gráfica para el rápido análisis de información que resulta de mucha ayuda cuando se requiere solamente consultar la información. Para poder tener una mejor apreciación de la información o visualizarla en alguna terminal sin acceso a la base de datos, se puede optar por migrar la información a un archivo de Microsoft Excel. Los datos de este archivo automáticamente se agrupan para su manejo mediante tablas dinámicas, en la siguiente sección se describe a mayor detalle el análisis de información mediante tablas dinámicas.

12.2 Análisis de datos mediante Microsoft Excel

En esta sección se lleva a cabo, el análisis de información mediante tablas dinámicas y gráficos dinámicos usando una hoja de cálculo de Microsoft Excel vinculada a los datos de un cubo OLAP generado con el servicio de análisis de SQL Server. Como y se había mencionado, la versión 2012 y posteriores, de SSAS integra las herramientas para crear una conexión entre los datos procesados mediante un cubo OLAP y una hoja de cálculo de Excel. Al manipularse una gran cantidad de datos, de más de un millón de registros, Excel presenta problemas de

procesamiento, asociados con el tamaño de la memoria de la computadora donde se procesa la información, pero al generar una conexión entre un cubo multidimensional creado con el servicio de análisis, Excel puede operar esta gran cantidad de datos sin presentar el problema en cuestión.

12.2.1 Análisis de datos mediante tablas dinámicas

El análisis de los datos comienza seleccionando del grupo de medida, los campos **Número de árboles**, **Área**, **Volumen**, **Biomasa** y **Carbono**; el orden de selección es importante, ya que será el orden en que se desplieguen la información en la tabla. En las versiones de office anteriores a 2013 es necesario arrastrar estos campos hasta la categoría de valores, de la misma forma como se hace con el examinador de datos. En la Figura 73 (lado derecho) se muestra la sección de campos de tabla dinámica; para obtener la información agregada a nivel de ecosistemas sólo se arrastra el campo **Ecosistema** agrupado en la jerarquía **Región**, hasta el área de filas. La información que se mostrará como valores de tabla dinámica se encuentran en la sección campo de valores y se agrega automáticamente al ser seleccionadas. Los resultados se muestran del lado derecho de la Figura 73.

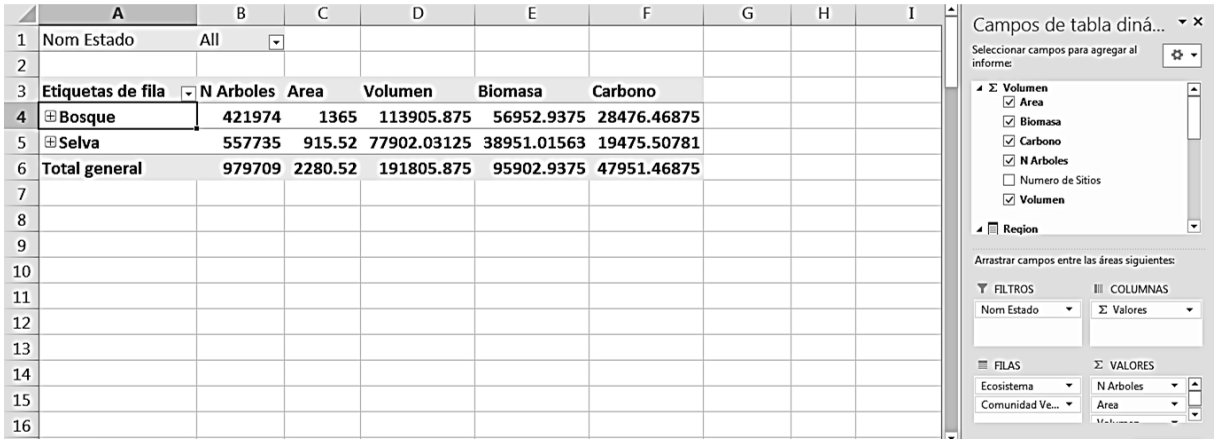
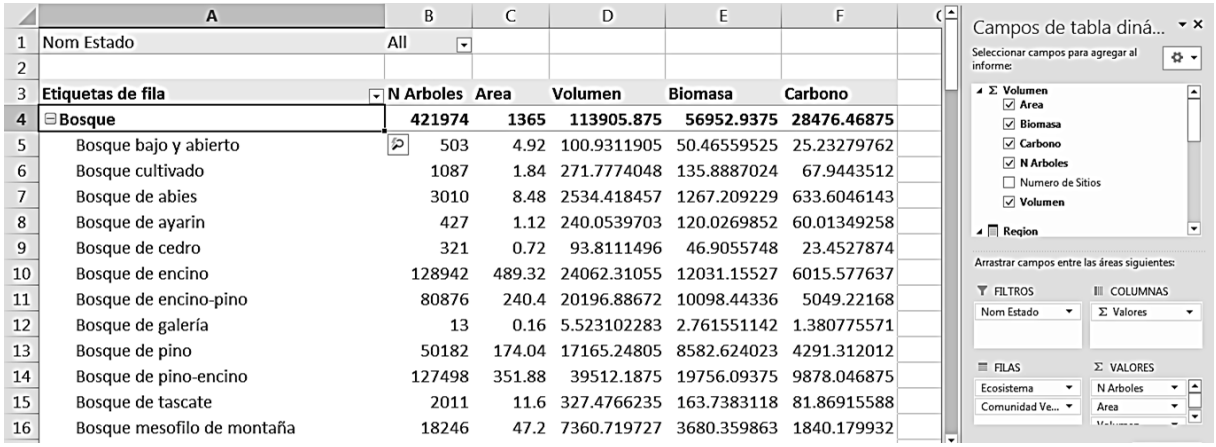


Figura 73: Análisis de información agregada por ecosistema usando tablas dinámicas
Fuente: Elaboración propia para la investigación.

Para realizar un análisis con mayor nivel de detalle, arrastre el campo **Comunidad vegetal** hasta el campo de filas, después del campo **Ecosistema**, la tabla resultante contiene la información

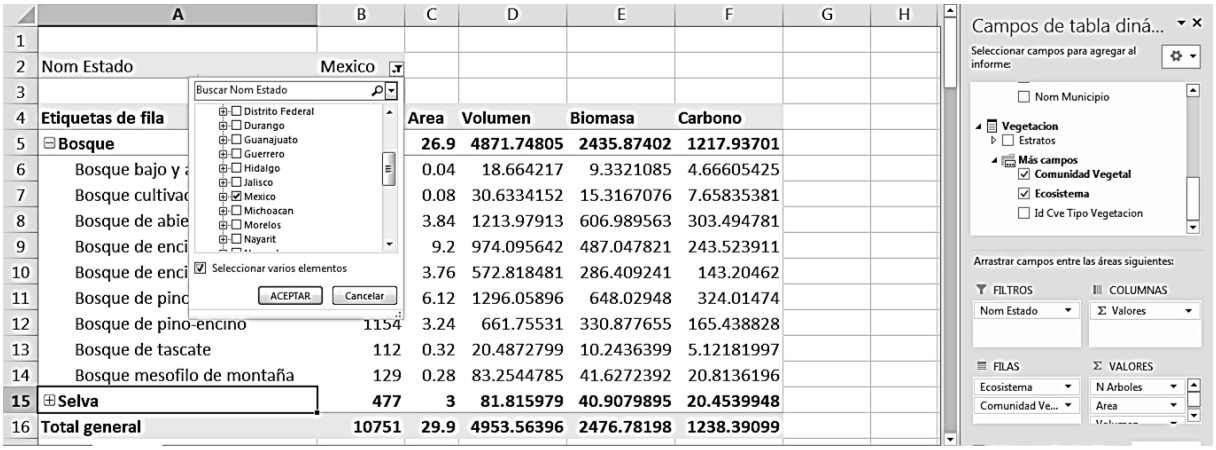
agregada, primero a nivel de comunidad vegetal y después a nivel de ecosistema; el resultado se muestra en la Figura 74.



Nom Estado	All				
Etiquetas de fila	**N Arboles**	**Area**	**Volumen**	**Biomasa**	**Carbono**
Bosque	**421974**	**1365**	**113905.875**	**56952.9375**	**28476.46875**
Bosque bajo y abierto	503	4.92	100.9311905	50.46559525	25.23279762
Bosque cultivado	1087	1.84	271.7774048	135.8887024	67.9443512
Bosque de abies	3010	8.48	2534.418457	1267.209229	633.6046143
Bosque de ayarin	427	1.12	240.0539703	120.0269852	60.01349258
Bosque de cedro	321	0.72	93.8111496	46.9055748	23.4527874
Bosque de encino	128942	489.32	24062.31055	12031.15527	6015.577637
Bosque de encino-pino	80876	240.4	20196.88672	10098.44336	5049.22168
Bosque de galería	13	0.16	5.523102283	2.761551142	1.380775571
Bosque de pino	50182	174.04	17165.24805	8582.624023	4291.312012
Bosque de pino-encino	127498	351.88	39512.1875	19756.09375	9878.046875
Bosque de tascate	2011	11.6	327.4766235	163.7383118	81.86915588
Bosque mesofilo de montaña	18246	47.2	7360.719727	3680.359863	1840.179932

Figura 74: Análisis de información agregada por comunidad vegetal y ecosistema usando tablas dinámicas.
Fuente: Elaboración propia para la investigación.

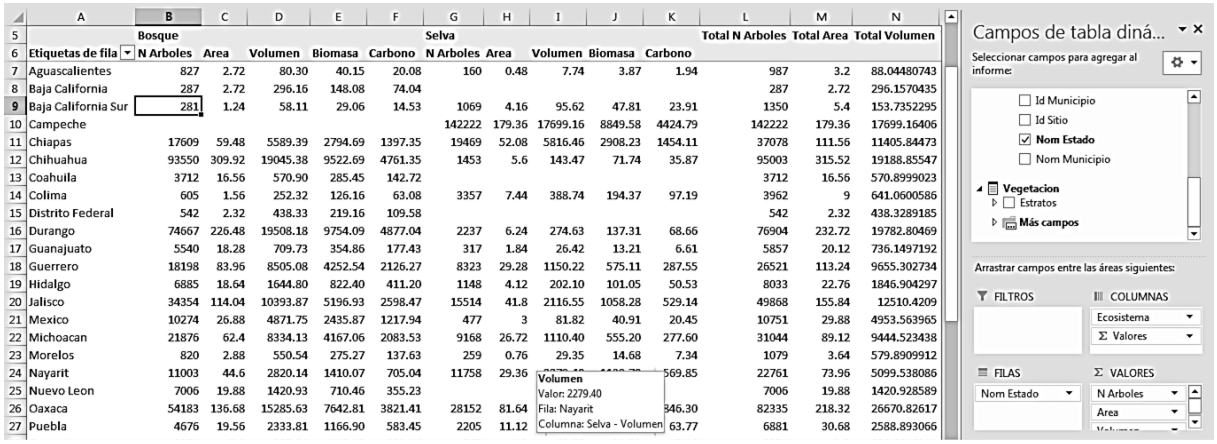
En la Figura 75 se muestra la información agregada por **comunidad vegetal** y por **ecosistema**, pero sólo para el Estado de México. Para obtener el resultado anterior, arrastre el campo **Nom Estado** hasta el área de filtro y cuando aparezca en la parte superior de la tabla dinámica, podrá ser usado para filtrar la información por estados. Marca la casilla etiquetada como seleccionar varios elementos y se activan las casillas para seleccionar los estados que se quieran agregar al análisis, para el ejemplo se seleccionó el Estado de México.



Nom Estado	Mexico				
Etiquetas de fila		**Area**	**Volumen**	**Biomasa**	**Carbono**
Bosque		**26.9**	**4871.74805**	**2435.87402**	**1217.93701**
Bosque bajo y a		0.04	18.664217	9.3321085	4.66605425
Bosque cultivad		0.08	30.6334152	15.3167076	7.65835381
Bosque de abie		3.84	1213.97913	606.989563	303.494781
Bosque de enci		9.2	974.095642	487.047821	243.523911
Bosque de enci		3.76	572.818481	286.409241	143.20462
Bosque de pino		6.12	1296.05896	648.02948	324.01474
Bosque de pino-encino	1154	3.24	661.75531	330.877655	165.438828
Bosque de tascate	112	0.32	20.4872799	10.2436399	5.12181997
Bosque mesofilo de montaña	129	0.28	83.2544785	41.6272392	20.8136196
Selva	**477**	**3**	**81.815979**	**40.9079895**	**20.4539948**
Total general	**10751**	**29.9**	**4953.56396**	**2476.78198**	**1238.39099**

Figura 75: Análisis de información para el Estado de México agregada por ecosistemas y comunidad vegetal usando tablas dinámicas.
Fuente: Elaboración propia para la investigación.

El siguiente análisis corresponde a la información agregada por filas a nivel de estados y por columnas a nivel de ecosistemas. Para obtener estos resultados, intercambie el campo **ecosistema** del área de filas al área de columnas, elimine el campo **comunidad vegetal** del área de filas y arrastre el campo **Nom Estado** hasta el área de filas. En la Figura 76 se muestra los resultados obtenidos, del lado izquierdo, y la configuración de los campos, del lado derecho.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
5		Bosque					Selva					Total N Arboles	Total Area	Total Volumen
6	Etiquetas de fila	N Arboles	Area	Volumen	Biomasa	Carbono	N Arboles	Area	Volumen	Biomasa	Carbono			
7	Aguascalientes	827	2.72	80.30	40.15	20.08	160	0.48	7.74	3.87	1.94	987	3.2	88.04480743
8	Baja California	287	2.72	296.16	148.08	74.04						287	2.72	296.1570435
9	Baja California Sur	281	1.24	58.11	29.06	14.53	1069	4.16	95.62	47.81	23.91	1350	5.4	153.7352295
10	Campeche						142222	179.36	17699.16	8849.58	4424.79	142222	179.36	17699.16406
11	Chiapas	17609	59.48	5589.39	2794.69	1397.35	19469	52.08	5816.46	2908.23	1454.11	37078	111.56	11405.84473
12	Chihuahua	93550	309.92	19045.38	9522.69	4761.35	1453	5.6	143.47	71.74	35.87	95003	315.52	19188.85547
13	Coahuila	3712	16.56	570.90	285.45	142.72						3712	16.56	570.8999023
14	Colima	605	1.56	252.32	126.16	63.08	3357	7.44	388.74	194.37	97.19	3962	9	641.0600586
15	Distrito Federal	542	2.32	438.33	219.16	109.58						542	2.32	438.3289185
16	Durango	74667	226.48	19508.18	9754.09	4877.04	2237	6.24	274.63	137.31	68.66	76904	232.72	19782.80469
17	Guanajuato	5540	18.28	709.73	354.86	177.43	317	1.84	26.42	13.21	6.61	5857	20.12	736.1497192
18	Guerrero	18198	83.96	8505.08	4252.54	2126.27	8323	29.28	1150.22	575.11	287.55	26521	113.24	9655.302734
19	Hidalgo	6885	18.64	1644.80	822.40	411.20	1148	4.12	202.10	101.05	50.53	8033	22.76	1846.904297
20	Jalisco	34354	114.04	10393.87	5196.93	2598.47	15514	41.8	2116.55	1058.28	529.14	49868	155.84	12510.4209
21	Mexico	10274	26.88	4871.75	2435.87	1217.94	477	3	81.82	40.91	20.45	10751	29.88	4953.563965
22	Michoacan	21876	62.4	8334.13	4167.06	2083.53	9168	26.72	1110.40	555.20	277.60	31044	89.12	9444.523438
23	Morelos	820	2.88	550.54	275.27	137.63	259	0.76	29.35	14.68	7.34	1079	3.64	579.8909912
24	Nayarit	11003	44.6	2820.14	1410.07	705.04	11758	29.36			569.85	22761	73.96	5099.538086
25	Nuevo Leon	7006	19.88	1420.93	710.46	355.23						7006	19.88	1420.928589
26	Oaxaca	54183	136.68	15285.63	7642.81	3821.41	28152	81.64			846.30	82335	218.32	26670.82617
27	Puebla	4676	19.56	2333.81	1166.90	583.45	2205	11.12			63.77	6881	30.68	2588.893066

Figura 76: Análisis de información, mediante el examinador de datos de datos del servicio de análisis de SQL Server, agregada por estados y ecosistema.
Fuente: Elaboración propia para la investigación a partir del cubo OLAP para análisis de volumen, biomasa y carbono.

12.2.2 Análisis de datos mediante gráficos dinámicos

La forma más sencilla de apreciar la información es usando una herramienta visual, como por ejemplo las gráficas. Para emplear las herramientas de gráficos dinámicos proporcionadas por Excel, primero se debe crear un vínculo entre los datos procesados mediante el paquete de análisis y una hoja de cálculo. Los pasos siguientes detallan como re realiza este procedimiento.

En la pestaña insertar de Microsoft Excel, seleccione el icono de gráfico dinámico como se muestra en la Figura 77.

Figura 77: Inserción de un gráfico dinámico para el análisis de información.

Fuente: Elaboración propia para la investigación.

En la ventana emergente selecciones la opción Utilice un fuente de datos externa y hacer clic en elegir conexión (lado izquierdo de la Figura 78). En la siguiente ventana emergente selecciones localhost TablaVolumenes y haga clic en el botón abrir (lado derecho de la Figura 78).

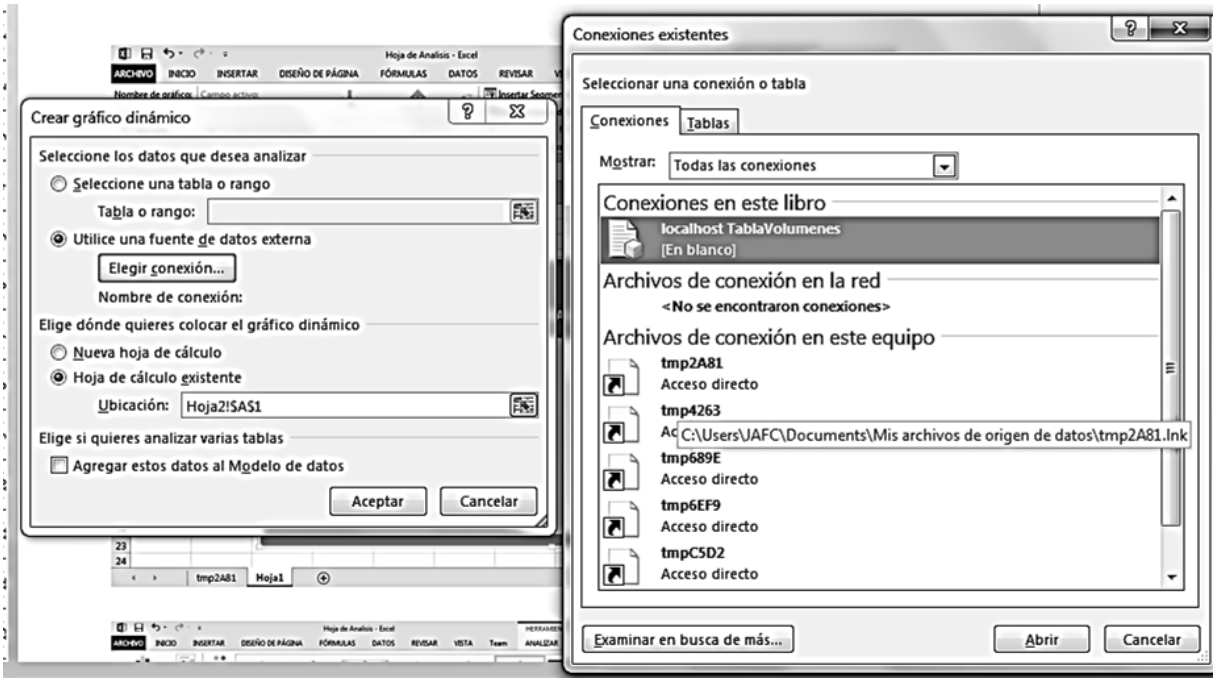


Figura 78: Establecer conexión para gráfico dinámico.
Fuente: Elaboración propia para la investigación.

Finalmente, dar en aceptar en la ventana **crear gráfico dinámico** para establecer la conexión. Al instante aparece las métricas y dimensiones con sus respectivas jerarquías en la sección **Campos de grafico dinámico**. La forma de realizar los análisis es similar al usado para tablas dinámicas, sólo que dependiendo del tipo de gráfico es el número de variables o campos que se pueden analizar al mismo tiempo; para el ejemplo sólo se analizará la información correspondiente al volumen, para ello arrastre el campo **volumen** hasta el área de valores, el campo **ecosistema** hasta el área de ejes y el campo **Nom Estado** hasta el área de filtro. El gráfico que se genera por default es un gráfico de barras que puede modificarse de manera similar a un gráfico tradicional de Excel. La opción más recomendable para la visualización de la información de totales sería un gráfico circular que se puede obtener siguiendo los pasos que se describen a continuación.

1. Seleccionar el gráfico generado y después seleccionar el icono **Cambiar tipo de gráfico**, en el lado superior derecho de la pantalla y selecciones circular el menú de gráficos.
2. Ir a la pestaña **Herramientas del gráfico dinámico** y luego **diseño**, para configurar la forma en que se presenta el gráfico, en el menú que se despliega hay diferentes estilos para un mismo tipo de gráfico.

El gráfico obtenido se muestra del lado derecho de la Figura 79 y la configuración de los campos en el lado izquierdo de la misma Figura.

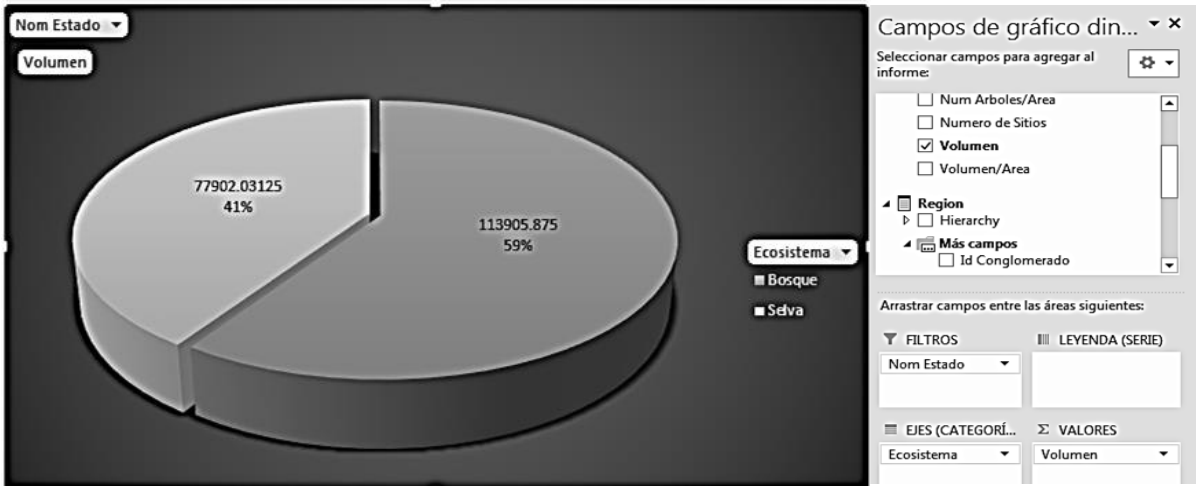


Figura 79: Análisis de volumen de madera mediante gráficos dinámicos agregado por ecosistema.
Fuente: Elaboración propia para la investigación.

Información con mayor detalle a nivel de comunidad vegetal, se obtiene al eliminar el campo **ecosistema** del área de ejes y arrastrar el campo **comunidad vegetal** hasta ésta área. El detalle de una correcta visualización se centra en la cantidad de comunidades diferentes, ya que al ser muchas, algunas no se logran apreciar adecuadamente en el gráfico, por lo cual lo ideal es seleccionar sólo algunas comunidades e ir analizando en pequeños grupos. Para realizar esta tarea se debe establecer un filtro en el área de eje de categorías, primero se mueve el campo **comunidad vegetal** al área de filtro, después se selecciona el grupo de comunidades a analizar y finalmente se regresa campo **comunidad vegetal** al área de valores. En la Figura 80 se puede observar un análisis realizado siguiendo los pasos descritos en éste párrafo.

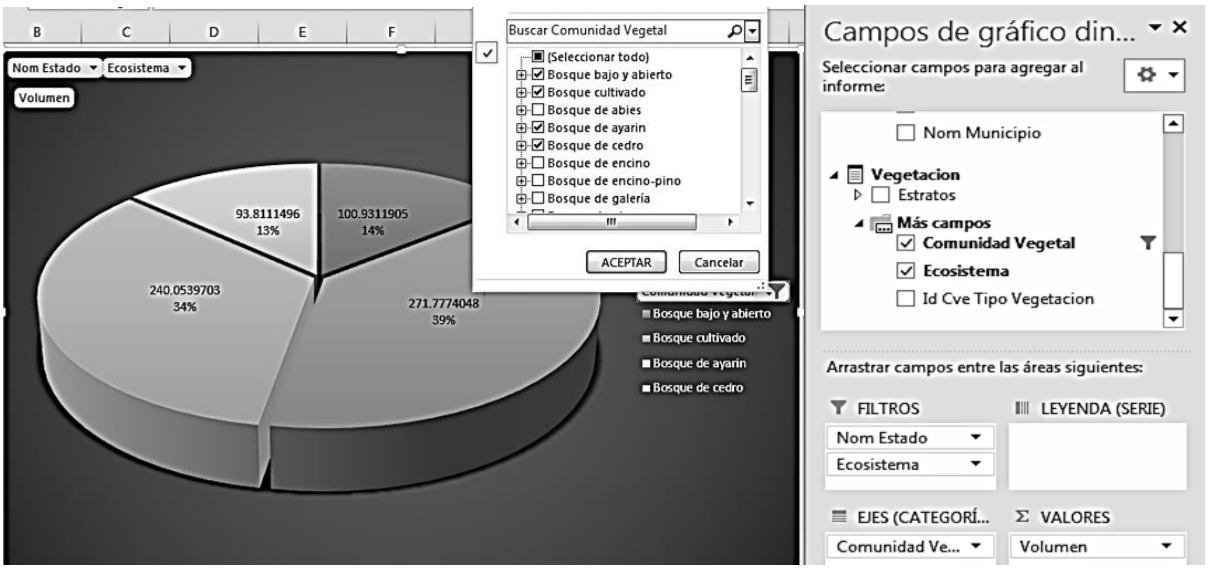


Figura 80: Análisis de volumen de madera mediante gráficos dinámicos agregado por comunidad vegetal.
Fuente: Elaboración propia para la investigación a partir del cubo OLAP para análisis de volumen, biomasa y carbono.

Al seguir la metodología de análisis propuesta desde un principio, en la Figura 81 se muestra el volumen madera agregado a nivel de ecosistemas, pero sólo la información que corresponde al Estado de México.

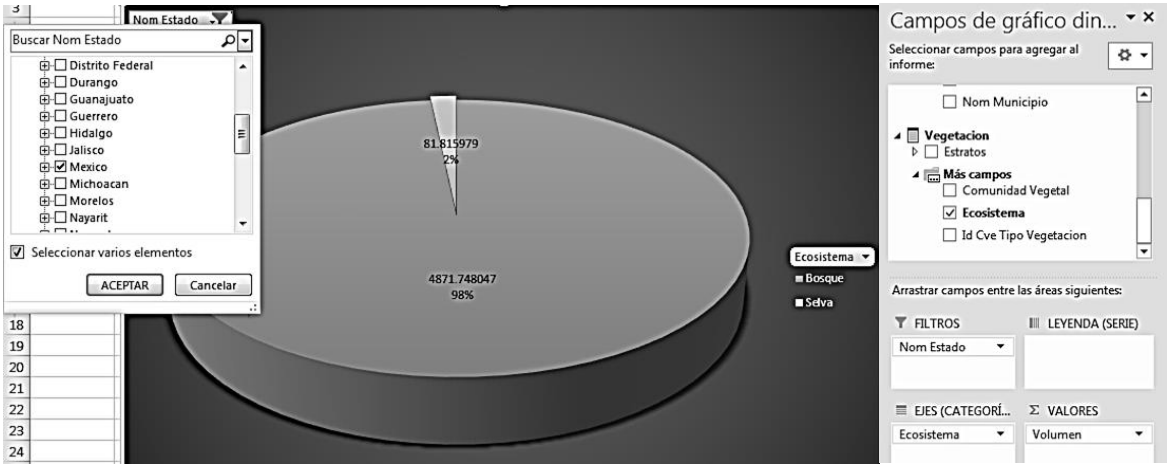


Figura 81: Análisis de volumen de madera mediante gráficos dinámicos, para el Estado de México, agregado por ecosistema.
Fuente: Elaboración propia para la investigación a partir del cubo OLAP para análisis de volumen, biomasa y carbono.

El tipo de gráfico, utilizado como ejemplo, no permite realizar un análisis multidimensional de los datos, como ya se había visto en los ejemplos anteriores. Como sólo se puede analizar una variable a la vez, agregada por un criterio, entonces para mostrar la información de los cálculos

de volumen de madera por estados, sólo se arrastra el campo **Nom Estado** hasta el área de ejes y se eliminan los campos restantes en esa área. Nuevamente, al haber muchos estados, la información no se aprecia con detalle, por lo cual se recomienda analizar un grupo pequeño de entidades federativas, filtrando la información y seleccionando sólo aquellos que se quieran analizar.

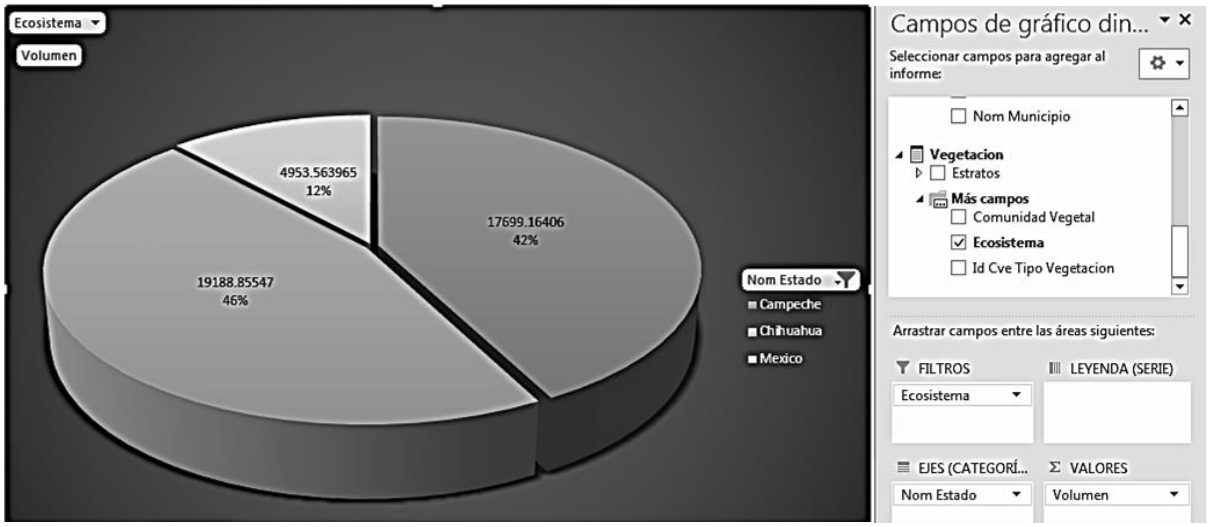


Figura 82. Análisis de información a diferente niveles mediante el examinador.
Fuente: Elaboración propia para la investigación a partir del cubo OLAP para análisis de volumen, biomasa y carbono.

El análisis de volumen madera agregada por estados, se muestra en la Figura 82, donde nuevamente sólo se seleccionó uno cuantos estados para poder observar adecuadamente la información.

Existen técnicas para realizar un análisis multidimensional de los datos usando gráficas circulares o de pastel, mediante la creación de segmentos de datos, pero ese tipo de análisis no se contempla en la presente investigación.

12.3 Análisis de datos mediante interfaces de consultas

Uno de los objetivos de utilizar sistemas OLAP para el análisis de información, es que ésta se encuentre disponible para los usuarios finales, analistas, investigadores, o cualquier categoría

de usuarios. La ejemplificación del uso de herramientas OLAP, para el análisis de la información procesada mediante cubos multidimensionales, no estaría completa si no se desarrollaran las interfaces necesarias para que los usuarios, no expertos en bases de datos, puedan acceder a ésta información.

La plataforma de desarrollo de Visual Studio 2010, proporciona un conjunto de herramientas para visualizar y analizar la información. Como productos para esta investigación se desarrollaron dos interfaces para interactuar con la información contenida en los cubos procesados. Tanto la interfaz de escritorio como la interfaz web consideran las ventajas de realizar un análisis de los datos de la misma forma como se realiza mediante tablas dinámicas con MS Excel.

12.3.1 Análisis de datos mediante la interfaz web

La interfaz web fue desarrollada con la intensión de que cualquier usuario con una conexión de internet pueda acceder a la información en línea. Ésta interfaz cuenta con cuatro áreas: la de filtro, la de filas, la de columnas y la de gráficos. En la Figura 83 se muestra la interfaz web en la que se presenta la información a un nivel de agregación muy bajo, a nivel de ecosistema, mediante gráficos y tablas. Lo interesante de ésta interfaz es que permite realizar un análisis dinámico muy similar a la de una tabla dinámica de MS Excel, para la información contenida en la tabla.

Ejemplos de operaciones que se pueden realizan con la interfaz son: filtros de información, operaciones de Drill Down y Roll up e intercambiar filas por columnas (pivotaje), la única información que es fija es la correspondiente a los hechos como número de árboles, número de hectáreas, volumen de madera, biomasa y carbono. La información presentada en los gráficos sólo trabaja con los totales a nivel de ecosistema, es decir, no se puede llevar a cabo un análisis dinámico de la información mediante gráficas.

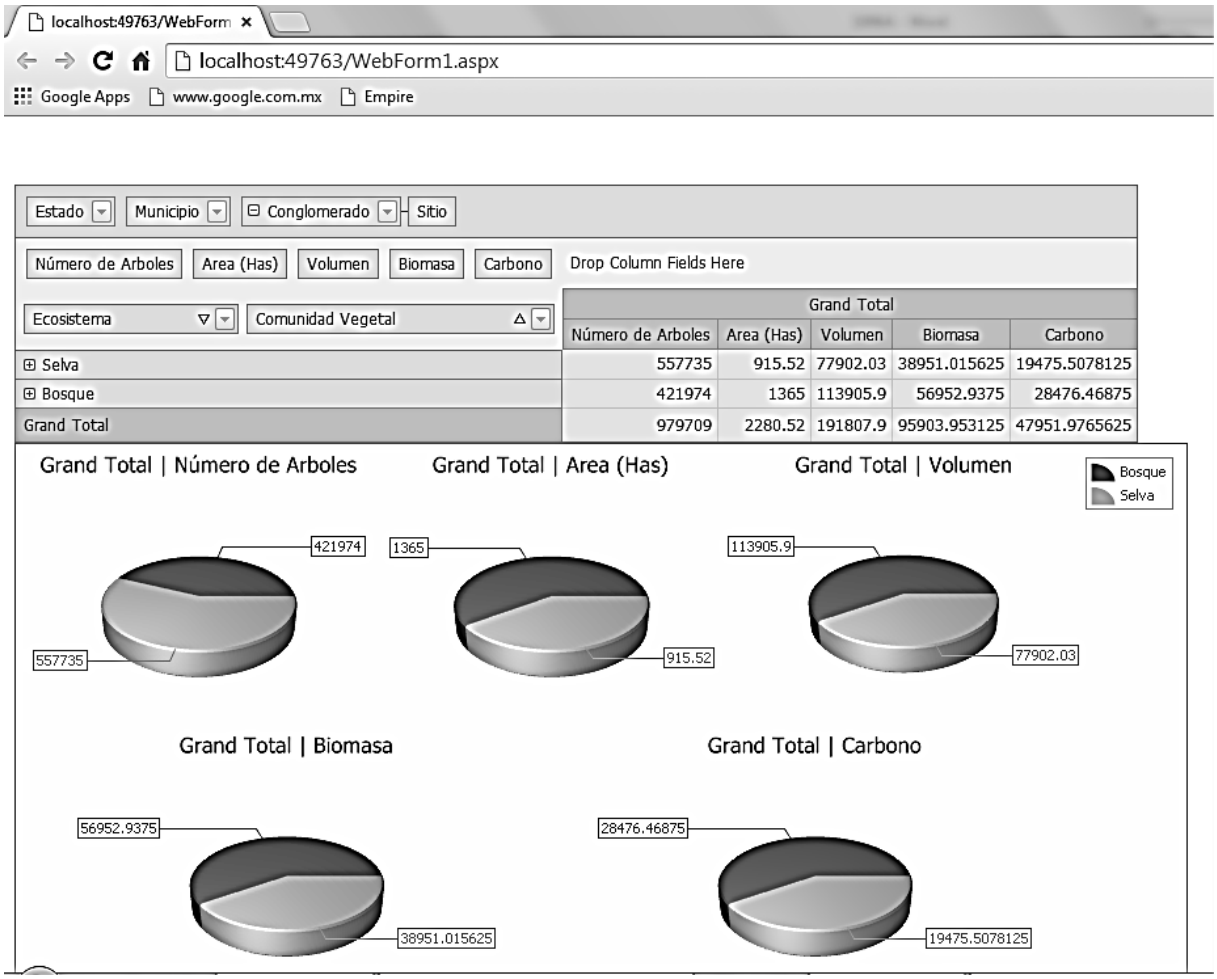


Figura 83: Análisis de información agregada a nivel de ecosistema usando la interfaz de consulta vía web.
Fuente: Elaboración propia para la investigación.

En la Figura 84 se ejemplifica la operación Drill Down para mostrar información con mayor nivel de detalle, hasta el nivel comunidad vegetal. Esta operación se puede realizar seleccionando el recuadro con signo más antes de la etiqueta en la columna ecosistema; de igual forma se realiza la operación inversa para ver la información nuevamente a nivel de ecosistema.

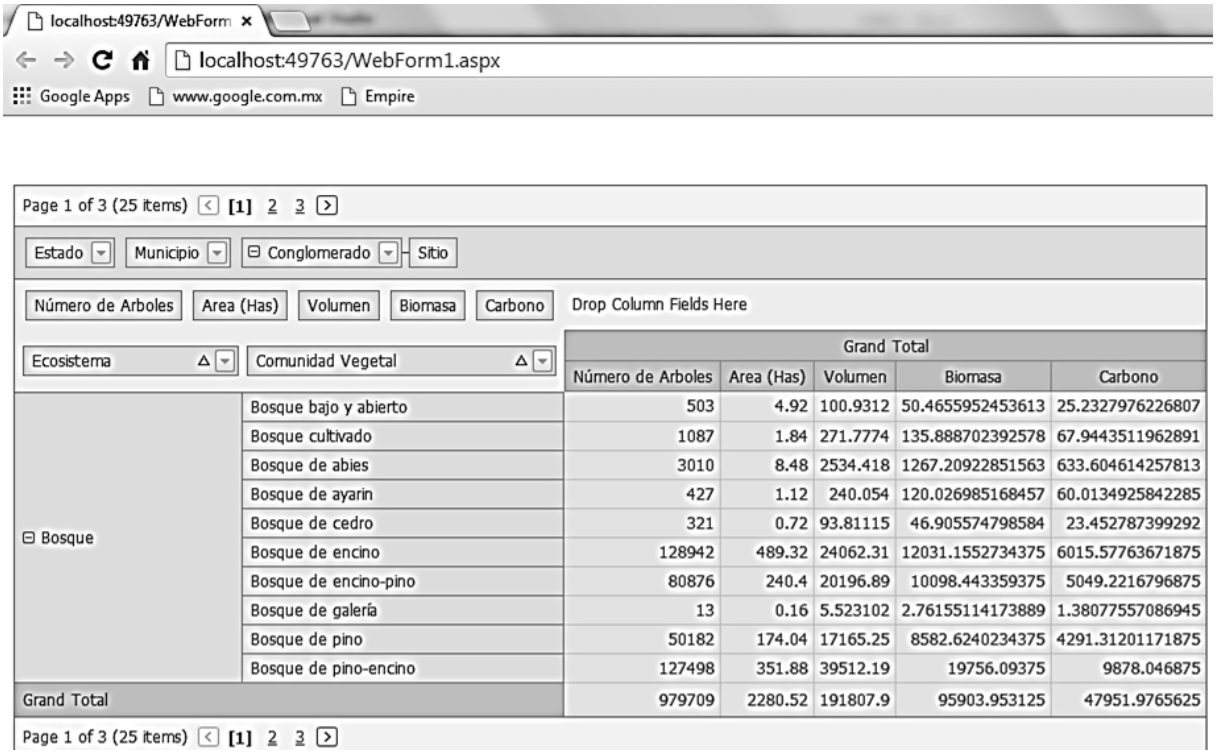


Figura 84: Análisis de información agregada a nivel de ecosistema y comunidad vegetal usando la interfaz de consulta vía web.
Fuente: Elaboración propia para la investigación.

En la Figura 85 se muestra la información agregada a nivel de ecosistema y a nivel de comunidad vegetal con el mismo grupo de métricas definidas desde un principio, pero mostrando sólo la información del estado de México.

En la Figura 86 se muestra un análisis de la información de los bosques y selvas a nivel de Estados, la principal ventaja de utilizar esta interfaz consiste en que es posible intercambiar la información entre filas, columnas, y área de filtro, pudiendo de ésta manera, observar los mismos datos desde diferentes ángulo y tener la certeza de que no se ha cometido algún error durante el proceso.

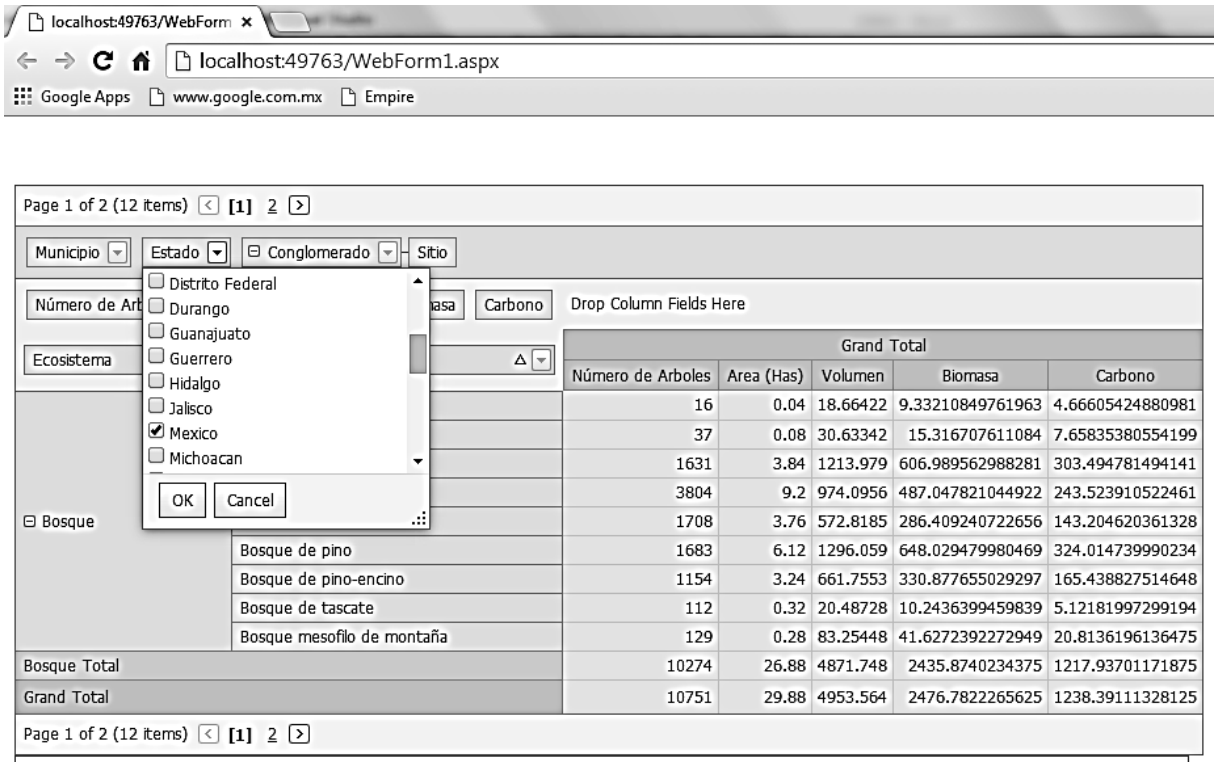


Figura 85: Análisis de información para el Estado de México agregada a nivel de ecosistema y comunidad vegetal, usando la interfaz de consulta vía web.
Fuente: Elaboración propia para la investigación.

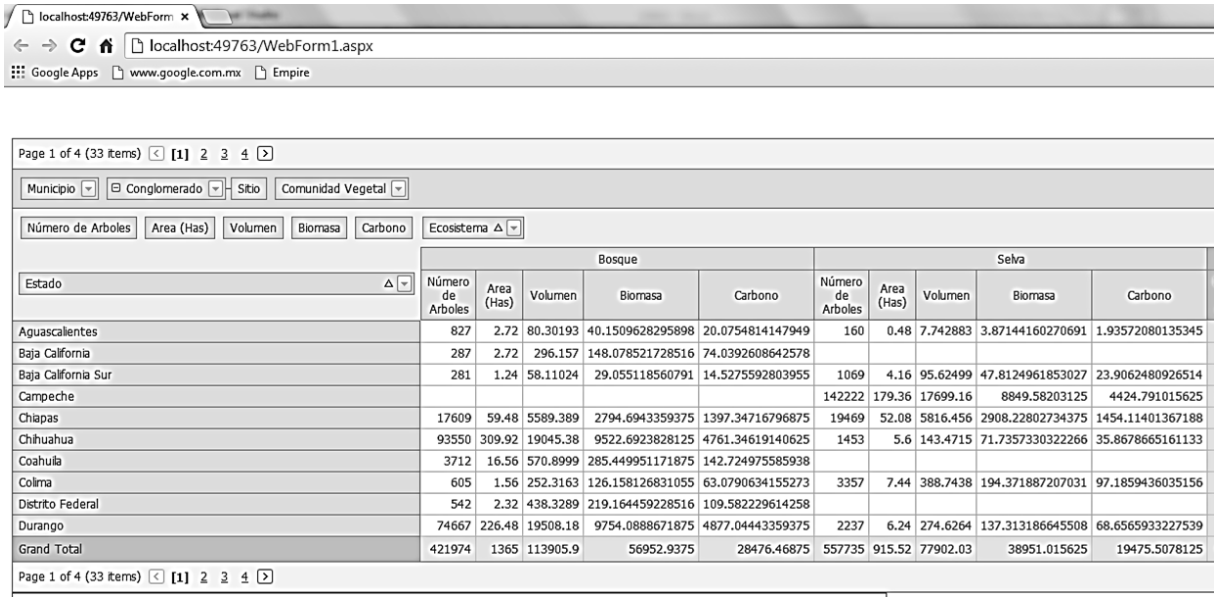


Figura 86: Análisis de información agregada, por columnas a nivel de ecosistema y por filas a nivel de estados, usando la interfaz de consulta vía web.
Fuente: Elaboración propia para la investigación.

12.3.2 Análisis de datos mediante la interfaz de escritorio

La interfaz de escritorio fue desarrollada con la intensión de los usuarios pertenecientes a la institución puedan acceder a la información mediante una terminal en red. Ésta interfaz es muy parecida a la interfaz web en cuanto a las tareas de filtrado, drill Down, roll up e intercambio de filas por columnas y viceversa, la diferencia entre esta interfaces es que la de escritorio no presenta gráficos de los totales por ecosistema. Para realizar las operaciones mencionadas anteriormente, se cuenta con tres áreas: la de filtro, la de filas y las de columnas; de forma predeterminada la información aparece con los campos **Ecosistema** y **Comunidad vegetal** en el área de filas, pero se puede realizar cualquier modificación en la dimensiones, adecuándolas al análisis que se pretenda. En la Figura 87 se muestra la interfaz de consulta resumida con bajo nivel de detalle, por ecosistema.

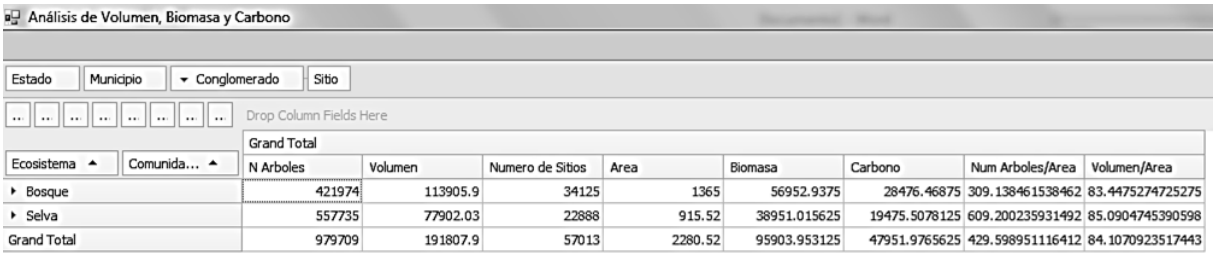
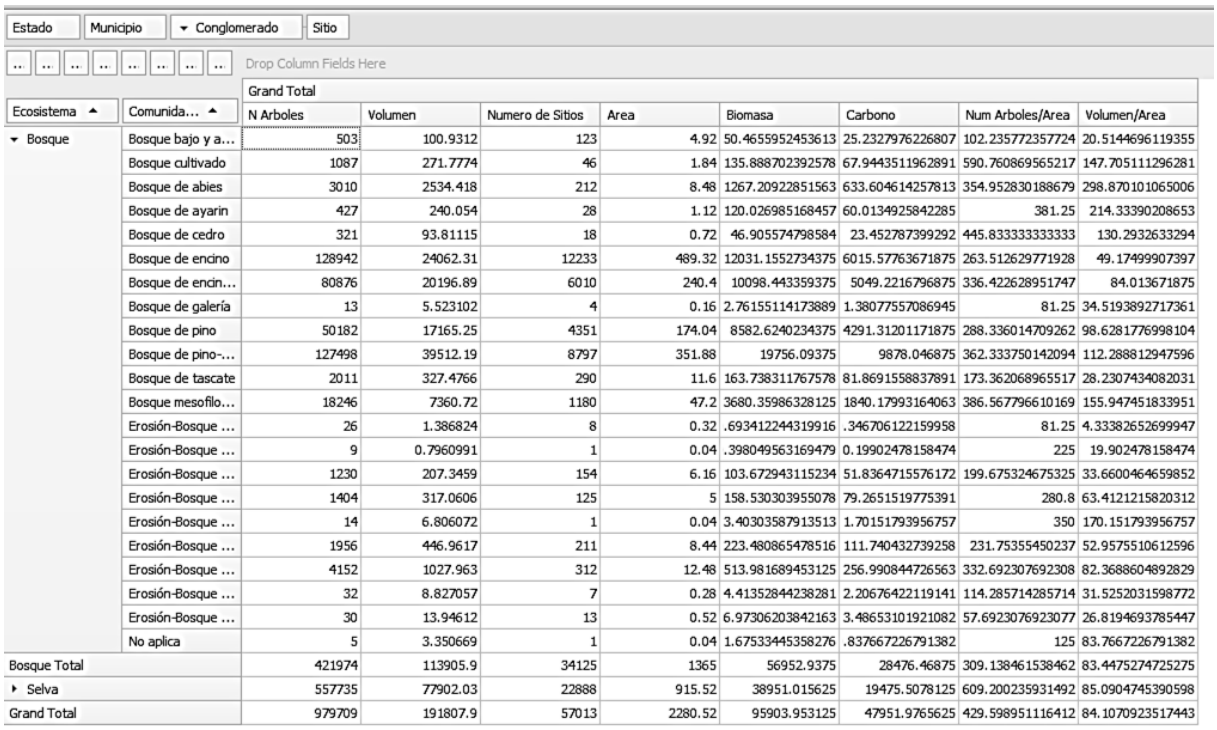


Figura 87: Análisis de información agregada a nivel de ecosistema, usando la interfaz de consulta de escritorio.
Fuente: Elaboración propia para la investigación.

En la Figura 88 se muestra la información agregada por comunidad vegetal y después por ecosistema. La forma de desplazarse por los diferentes niveles de agregación es similar al de la interfaz web, seleccionando la etiqueta en la columna ecosistema.



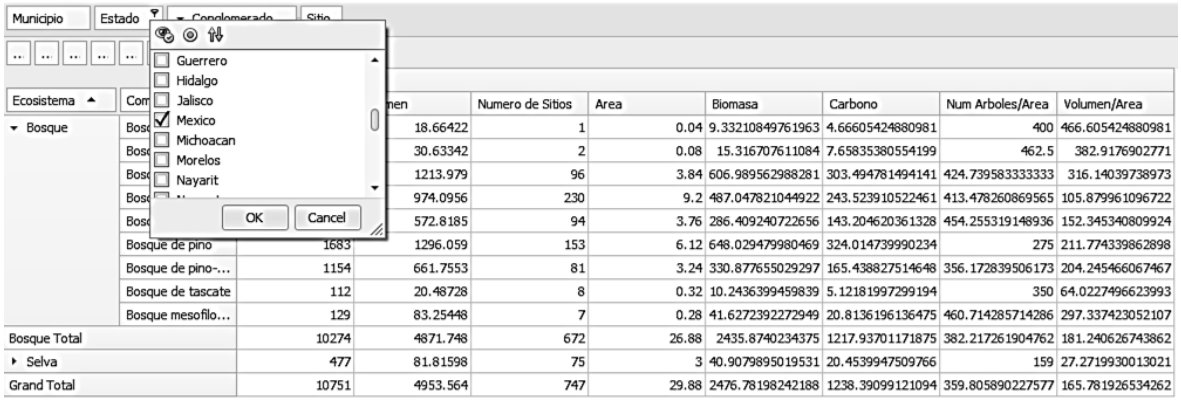
Estado | Municipio | Conglomerado | Sitio

Drop Column Fields Here

Ecosistema	Comunida...	Grand Total N Arboles	Volumen	Numero de Sitios	Area	Biomasa	Carbono	Num Arboles/Area	Volumen/Area
Bosque	Bosque bajo y a...	503	100.9312	123	4.92	50.4655952453613	25.2327976226807	102.235772357724	20.5144696119355
	Bosque cultivado	1087	271.7774	46	1.84	135.888702392578	67.9443511962891	590.760869565217	147.705111296281
	Bosque de abies	3010	2534.418	212	8.48	1267.20922851563	633.604614257813	354.952830188679	298.870101065006
	Bosque de ayarin	427	240.054	28	1.12	120.026985168457	60.0134925842285	381.25	214.33390208653
	Bosque de cedro	321	93.81115	18	0.72	46.905574798584	23.452787399292	445.833333333333	130.2932633294
	Bosque de encino	128942	24062.31	12233	489.32	12031.1552734375	6015.57763671875	263.512629771928	49.17499907397
	Bosque de encin...	80876	20196.89	6010	240.4	10098.443359375	5049.2216796875	336.422628951747	84.013671875
	Bosque de galería	13	5.523102	4	0.16	2.76155114173889	1.38077557086945	81.25	34.5193892717361
	Bosque de pino	50182	17165.25	4351	174.04	8582.6240234375	4291.31201171875	288.336014709262	98.6281776998104
	Bosque de pino-...	127498	39512.19	8797	351.88	19756.09375	9878.046875	362.333750142094	112.288812947596
	Bosque de tascate	2011	327.4766	290	11.6	163.738311767578	81.8691558837891	173.362068965517	28.2307434082031
	Bosque mesofilo...	18246	7360.72	1180	47.2	3680.35986328125	1840.17993164063	386.567796610169	155.947451833951
	Erosión-Bosque ...	26	1.386824	8	0.32	.693412244319916	.346706122159958	81.25	4.33382652699947
	Erosión-Bosque ...	9	0.7960991	1	0.04	.398049563169479	0.19902478158474	225	19.902478158474
	Erosión-Bosque ...	1230	207.3459	154	6.16	103.672943115234	51.8364715576172	199.675324675325	33.6600464659852
	Erosión-Bosque ...	1404	317.0606	125	5	158.530303955078	79.2651519775391	280.8	63.4121215820312
	Erosión-Bosque ...	14	6.806072	1	0.04	3.40303587913513	1.70151793956757	350	170.151793956757
	Erosión-Bosque ...	1956	446.9617	211	8.44	223.480865478516	111.740432739258	231.75355450237	52.9575510612596
	Erosión-Bosque ...	4152	1027.963	312	12.48	513.981689453125	256.990844726563	332.692307692308	82.3688604892829
	Erosión-Bosque ...	32	8.827057	7	0.28	4.41352844238281	2.20676422119141	114.285714285714	31.5252031598772
	Erosión-Bosque ...	30	13.94612	13	0.52	6.97306203842163	3.48653101921082	57.6923076923077	26.8194693785447
	No aplica	5	3.350669	1	0.04	1.67533445358276	.837667226791382	125	83.7667226791382
Bosque Total		421974	113905.9	34125	1365	56952.9375	28476.46875	309.138461538462	83.4475274725275
Selva		557735	77902.03	22888	915.52	38951.015625	19475.5078125	609.200235931492	85.0904745390598
Grand Total		979709	191807.9	57013	2280.52	95903.953125	47951.9765625	429.598951116412	84.1070923517443

Figura 88: Análisis de información agregada a nivel de ecosistema y comunidad vegetal, usando la interfaz de consulta de escritorio.
Fuente: Elaboración propia para la investigación.

En la Figura 89 se muestra la información para el Estado de México, agregada por comunidad vegetal y por ecosistema. La manera de obtener estos resultados es seleccionando el campo Estado en el área de filtro y marcar solamente el recuadro del Estado de México.



Municipio | Estado | Conglomerado | Sitio

Guerrero
Hidalgo
Jalisco
Mexico
Michoacan
Morelos
Nayarit
OK | Cancel

Ecosistema	Com...	...men	Numero de Sitios	Area	Biomasa	Carbono	Num Arboles/Area	Volumen/Area
Bosque	Bosq...	18.66422	1	0.04	9.33210849761963	4.66605424880981	400	466.605424880981
	Bosq...	30.63342	2	0.08	15.316707611084	7.65835380554199	462.5	382.9176902771
	Bosq...	1213.979	96	3.84	606.989562988281	303.494781494141	424.739583333333	316.14039738973
	Bosq...	974.0956	230	9.2	487.047821044922	243.523910522461	413.478260869565	105.879961096722
	Bosq...	572.8185	94	3.76	286.409240722656	143.204620361328	454.255319148936	152.345340809924
	Bosque de pino	1683	1296.059	153	6.12	648.029479980469	324.014739990234	275	211.774339862898
	Bosque de pino-...	1154	661.7553	81	3.24	330.877655029297	165.438827514648	356.172839506173	204.245466067467
	Bosque de tascate	112	20.48728	8	0.32	10.2436399459839	5.12181997299194	350	64.0227496623993
	Bosque mesofilo...	129	83.25448	7	0.28	41.6272392272949	20.8136196136475	460.714285714286	297.337423052107
Bosque Total		10274	4871.748	672	26.88	2435.8740234375	1217.93701171875	382.217261904762	181.240626743862
Selva		477	81.81598	75	3	40.9079895019531	20.4539947509766	159	27.2719930013021
Grand Total		10751	4953.564	747	29.88	2476.78198242188	1238.39099121094	359.805890227577	165.781926534262

Figura 89: Análisis de información agregada, por columnas a nivel de ecosistema y por filas a nivel de estados, usando la interfaz de consulta vía web.
Fuente: Elaboración propia para la investigación.

En la Figura 90 se muestra la información agregada a nivel de estado, por filas, y a nivel de ecosistema, por columna. La forma de obtener éstas consultas es desplazando el campo **Estado** al área de filas y el campo **Ecosistema** al área de columnas, también es necesario quitar los campos que no se necesitan del área de filas y pasarlos al área de filtro, para éste ejemplo el campo el campo que no se necesita es **comunidad vegetal**.

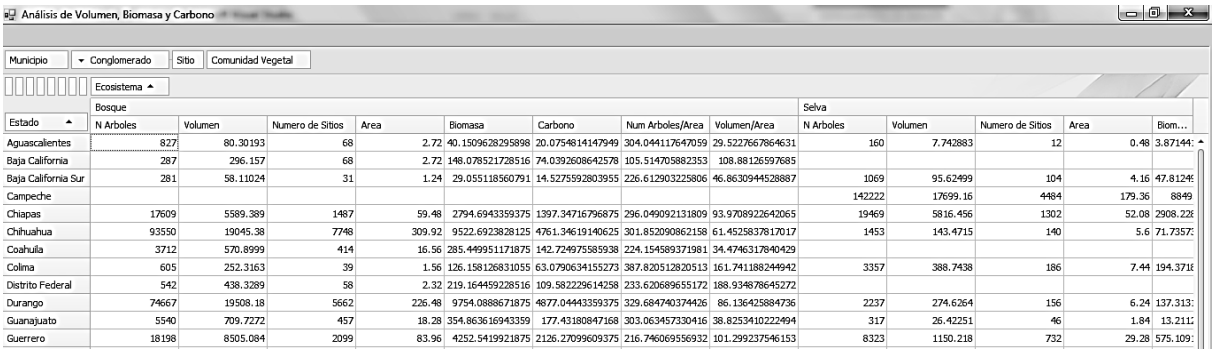


Figura 90: Análisis de información agregada, por columnas a nivel de ecosistema y por filas a nivel de estados, usando la interfaz de consulta vía web.
Fuente: Elaboración propia para la investigación.

12.4 Resumen

En el presente capítulo se ejemplificó el uso de las diferentes interfaces con que cuenta el servicio de análisis de SQL Server y las interfaces desarrollada, para el análisis y visualización de la información procesada mediante el cubo OLAP.

La interfaz principal corresponde al examinador de datos que es una herramienta gráfica de consultas, integrado al servicio de análisis de SQL Server (SSAS).

La segunda interfaz es un archivo de MS Excel, generado a partir del cubo, con el objetivo de explotar el potencial de las tablas y gráficos dinámicos.

Las dos interfaces restantes fueron desarrolladas para servir como intermediario entre los usuarios finales y la información del cubo; se desarrollaron una interfaz de escritorio y una interfaz web que emulan al examinador de datos de SSAS.

Con cada una de las interfaces se realizaron una serie de análisis para ejemplificar el funcionamiento de cada una; los análisis realizados con cada interfaz fueron iguales con el objetivo de obtener los mismos resultados. La información analizada correspondió a los cálculos de volumen de madera, biomasa y carbono, siguiendo el orden siguiente:

- En primer lugar se analizó la información agregada a nivel de ecosistemas.
- En segundo lugar se analizó la información agregada por Comunidad Vegetal y por Ecosistema.
- En tercer lugar, se analizó la información correspondiente al Estado de México.
- En último lugar se analizó la información agregada por entidad federativa, a nivel de filas y pro ecosistemas, a nivel de columnas.

13. ANÁLISIS DE RÚBRICAS Y CONTRASTE DE HIPÓTESIS

13.1 Rúbricas

En esta sección se pone una autoevaluación del software a través de la rúbrica diseñada (Anexo 2).

La evaluación realizada es la siguiente:

Categoría	1	2	3	4	Valor
Facilidad de uso	Se necesita ser un experto en el uso de computadoras para usar las interfaces	Se necesita algún entrenamiento para usar las interfaces	No se necesitan grandes conocimientos de computación para usar las interfaces.	Cualquier persona puede usar las interfaces	**4**
Capacidad de personalización	Las interfaces no pueden ser adaptadas de acuerdo a las necesidades de análisis.	Algunas partes de las interfaces pueden ser adaptadas de acuerdo a las necesidades de análisis.	La mayor parte de las interfaces pueden ser adaptadas de acuerdo a las necesidades de análisis.	Cualquier parte de las interfaces puede ser adaptada de acuerdo a las necesidades de análisis.	**3**
Accesibilidad	Sólo se puede acceder a las interfaces instalándolas en la computadora.	Se puede acceder a las interfaces desde internet o instaladas a partir de un CD.	Se puede acceder a las interfaces desde internet.	Las interfaces son accesibles desde internet. No se requiere su descarga.	**4**
Lectura de resultados	Resultados difíciles de leer.	Los resultados presentan alguna dificultad en la lectura.	Los resultados no presentan dificultad en su lectura.	Los resultados son muy fáciles de leer.	**4**
Originalidad con respecto al diseño	Las interface no presentan un diseño original.	Las interfaces presentan un diseño ligeramente original.	Las interfaces presentan un diseño original	Las interfaces presentan un diseño muy original.	**2**
Uso de tecnología avanzada	Las interfaces no hacen uso de tecnología avanzada.	Las interfaces hacen ligero uso de tecnología avanzada.	Las interfaces hacen uso de tecnología avanzada.	Las interfaces hacen uso de tecnología muy avanzada.	**4**
Utilidad para el análisis de datos	Las interfaces no son útiles para el análisis de datos.	Las interfaces resultan poco útiles para el análisis de datos.	Las interfaces son útiles para el análisis de datos.	Las interfaces son muy útiles para el análisis de datos.	**4**

De lo anterior se puede deducir que el software implementado presenta algunos aspectos positivos y negativos a resaltar, estos son:

Aspectos Positivos:

- Útil para la obtención de datos forestales de volumen, carbono y biomasa, importantes para medir el cambio climático y volumen de carbono, entre otros.
- Original, no se detectó un software realizado que utilice un Data Warehouse en el área forestal.
- Facilidad de uso.

Aspectos negativos:

- Se tiene que tener instalar el SQL Server 2008 Enterprise o versiones superiores en la computadora donde se va a utilizar esta herramienta.
- El SQL Server debe tener acceso vía Internet para las consultas web.
- Faltó tiempo para experimentar más con la obtención de gráficos dinámicos, el software solo lo hace con datos totales.

13.2 Contraste de hipótesis

La *hipótesis general* de la investigación es: A partir de la base de datos del Inventario Nacional Forestal y de Suelos 2004-2009, se pueden aplicar modelos de minería de datos y diseñar un Data Warehouse para el análisis de información forestal específica. La hipótesis planteada no se rechaza porque fue posible aplicar cuatro modelos de minería de datos usando la información de las tablas **TblArboladoBosqueSelva**, **TblSitio** y **TblConglomerado**, de la base datos del Inventario Nacional Forestal y de Suelos 2004-2009, para la clasificación del género arbóreo *Quercus* y usando la información de las tablas **TblArboladoBosqueSelva**, **TblSitio** y **TblConglomerado**, **CatVegetacionInegiGeneral**, **CatEstado** y **CatMunicipio**, se diseñó un Data Warehouse para el análisis de volumen de madera, biomasa y carbono.

Hipótesis específicas

- La reconstrucción del diagrama de la base de datos del Inventario Nacional y de Suelos 2004-2009 permite la identificación de su diseño.

La hipótesis no se rechaza porque al diseñar el diagrama de la base de datos del Inventario Nacional Forestal y de Suelos 2004-2009 se fueron identificando cada una de las entidades que la conforman, así como sus relaciones y cardinalidades, encontrándose que su diseño está basado en un modelo relacional.

- La información de la base de datos del Inventario Nacional y de Suelos 2004-2009 se utiliza para elegir modelos de minería de datos para la clasificación del género arbóreo *Quercus*.

La hipótesis no se rechaza porque, usando la información de las tablas **TblArboladoBosqueSelva**, **TblSitio** y **TblConglomerado**, de la base datos del Inventario Nacional Forestal y de Suelos 2004-2009, se utilizaron los modelos de minería de datos basados en árboles de decisión, análisis clúster, red neuronal y regresión logística, seleccionándose los modelos de árboles de decisión y análisis clúster como los mejores modelos que clasificaron al género arbóreo *Quercus*.

- La elaboración de un Data Warehouse a partir de la base de datos del Inventario Nacional Forestal y de Suelos 2004-200,9 permite el análisis de datos y la generación de reportes.

La hipótesis no se rechaza porque la tabla de hechos del Data Warehouse contiene información del volumen de madera, agregado a nivel sitios, lo que lo hace ideal para realizar diferentes tipos de análisis, en concordancia con el diseño de muestreo utilizado en la recolección de datos, y las dimensiones fueron específicamente diseñadas para la generación de reportes, basándose en el formato que utiliza la CONAFOR, tanto para reportes nacionales como internacionales.

- Mediante la tecnología de cubos dimensionales a partir de un Data Warehouse se analizan datos de volumen de madera, biomasa y carbono.

La hipótesis no se rechaza porque el cubo multidimensional, diseñado a partir del Data Warehouse, contiene la información correspondiente al volumen de madera, biomasa y carbono, así como el número de hectáreas utilizada como variable auxiliar para llevar a cabo estimaciones espaciales de estos indicadores. Las dimensiones del cubo permiten realizar análisis del volumen de madera, biomasa y carbono de forma jerarquizada por ecosistema y comunidad vegetal o por estados, municipios y conglomerados.

- La plataforma de desarrollo Visual Studio 2010 proporciona las herramientas necesarias para el desarrollo de aplicaciones para la visualización y análisis de datos procesados mediante un cubo multidimensional generado a partir de un Data Warehouse.

La hipótesis no se rechaza porque usando ésta plataforma de desarrollo y el lenguaje de programación .Net se diseñaron dos interfaces, una de escritorio y otra web, para el análisis de la información procesada mediante el cubo multidimensional. Las interfaces se conectan al cubo y permiten visualizar la información y llevar a cabo un análisis de ésta desde diferentes perspectivas, intercambiando filas, columnas o filtros.

14. CONCLUSIONES Y RECOMENDACIONES

14.1 Conclusiones

Generales

- El objetivo general de la investigación se cumplió considerando que, con la información de la base de datos del Inventario Nacional Forestal y de Suelos 2004-2009, se pudieron aplicar cuatro algoritmos de minería de datos para la clasificación del género arbóreo Quercus, y se diseñó un Data Warehouse que fue usado para el análisis de información forestal como volumen de madera, biomasa y carbono, mediante la construcción de cubos multidimensionales.
- La hipótesis planteada no se rechazó con lo cual se establece que si es posible aplicar modelos de minería de datos y diseñar un Data Warehouse a partir de la base de datos del Inventario Nacional Forestal y de Suelos 2004-2009.

Base de datos del Inventario Nacional Forestal y de Suelos 2004-2009

- Se identificó la estructura de la base de datos del Inventario Nacional Forestal y de Suelos 2004-2009. La conclusión a la que se llegó fue que la base de datos está basado en un modelo relacional pero que no se apega en un 100% a este modelo, por lo que se puede decir que es semi-relacional. Para llevar a cabo esta tarea primero se analizaron las tablas presentes en la base de datos, encontrándose que está conformado por 67 tablas distribuidas de la siguiente manera: 24 tablas con información del medio físico y 43 tablas del tipo catálogo. Para la investigación que se presenta la tabla más importante es TblArboladoBosqueSelva que contiene la información dasométrica del arbolado, y para la CONAFOR las tablas más importantes son: Tblconglomerado y TblSitio que registran la información de los conglomerados o unidades primarias de muestreo y los sitios o unidades secundarias de muestreo, respectivamente. Encontrar las relaciones entre cada una de las tablas no fue una tarea sencilla, ya que algunos nombre de los atributos que

relacionan a dos entidades no son iguales en las dos entidades. El principal método para la identificar las relaciones y las cardinalidades fue usar las herramientas de base de datos de Microsoft Access, donde en el apartado de relaciones se eliminaron todas las tablas, se seleccionó una en particular y a partir de esta se fueron agregando las tablas relacionadas. Un comentario importante es que para determinar el esquema de una base de datos que contiene un gran cantidad de tablas resulta muy complicado.

- Se verificó si la base de datos se apega a las reglas de Codd para bases de datos relacionales, encontrándose que ésta forma de análisis no es adecuada debido a que sólo 4 de 12 reglas pueden ser aplicadas al diseño de la base de datos y las restantes 8 reglas se aplican a los sistemas gestores de bases de datos. De las 4 reglas aplicables, sólo tres reglas se cumplieron satisfactoriamente por lo cual se concluyó que el modelo de datos no es completamente relacional.
- La base datos del Inventario Nacional Forestal y de Suelos 2004-2009 contiene mucho ruido en sus datos, los más comunes son los datos atípicos y los datos perdidos, lo que implica un largo proceso de depuración de la información para poder utilizarla en informes e investigaciones.
- Uno de los principales resultados obtenido por los Inventarios Forestales Nacionales fue la integración de una base de datos geoespacial, reportada por la Gerencia de Inventario y Geomática de la CONAFOR como estructurada bajo un modelo relacional, orientado hacia el análisis y la obtención de los principales indicadores que se establecen en la Ley General de Desarrollo Forestal Sustentable y su Reglamento. No obstante que el modelo usado pueda proporcionar las condiciones anteriormente mencionadas, la teoría de base de datos sugiere que antes de elegir un modelo de datos, es necesario analizar los objetivos de la integración de una base de dato. Diferentes modelos de datos pueden ser, más o menos adecuados, para ciertos propósitos; en algunos casos, como se mencionó en el marco teórico, las bases de datos relacionales no son la mejor opción para almacenar información que va ser sujeta a análisis porque algunas consultas requieren de la información contenida en diferentes tablas lo que implica realizar operaciones de reunión natural o externa que son muy costosas a nivel de recursos máquina. La mejor

manera de solucionar el problema anteriores es quitando la normalidad de las tablas y usar en su lugar un diagrama de estrella o copo de nieve.

Minería de Datos

- Se probaron cuatro modelos de minería de datos, para la clasificación del género arbóreo *Quercus,* del cual el mejor fue el de árbol de decisión con una precisión del 74%, seguido por el de análisis de conglomerados con una precisión 71% y los modelos de redes neuronales y regresión logística con una precisión del 69% y 67%, respectivamente. Para la generación de los modelos se utilizó el servicio de análisis de SQL Server 2008. Como datos de entrenamiento y validación se tomó un muestreo simple aleatorio de aproximadamente el 10% de la información dasométrica de los géneros maderables *Quercus, Pinus, Bursera, Lysiloma y Piscidia,* a nivel nacional, el tamaño de la muestra tomada se justifica en el Teorema de Límite Central y la experiencia de MacLennan *et al.* (2009). La decisión sobre el mejor modelo se tomó con base en dos criterios, la gráfica de elevación y la matriz de clasificación, ambos obtenidos al momento de procesar los modelos.
- En un principio se contempló la posibilidad de aplicar minería de datos a la información de la tabla de hechos del Data Warehouse, pero el alto grado de agregación, a nivel de sitios, no permitió utilizarla para llevar a cabo un análisis de datos mediante algoritmos de minería. Para analizar los datos con algún algoritmo de minería de datos se necesita los datos dasométricos de cada árbol, por lo cual, el análisis se realizó usando los datos de la tabla TblArboladoBosqueSelva contenida en la base datos original del Inventario Nacional Forestal y de Suelos 2004-2009.
- La creación de un modelo de minería de datos es un proceso dinámico e iterativo; es decir, una vez que han explorado los datos, se puede descubrir que son insuficientes para crear los modelos de minería de datos adecuados y que, por tanto, se deben considerar más datos. O bien, se pueden generar varios modelos y descubrir entonces que no responden adecuadamente al problema planteado cuando los definió y que, por tanto, debe volver a definir el problema. Puede que haya que repetir cada paso del proceso varias veces para crear un modelo adecuado. Cuando ya se tienen definidos uno o varios

modelos, es posible que deba actualizarlos debido a la posibilidad de tener más datos disponibles.

Data Warehouse

- Diversos autores hacen las mismas referencias a alguno de los dos enfoque principales de la Teoría Data Warehouse, propuestos por William Inmon (Data Warehouse Empresarial) y Ralph Kimball (Data Warehouse Bus).
- Se diseñó un Data Warehouse a partir la base de datos del Inventario Nacional Forestal y de Suelos 2004-2009 que fue utilizado para el análisis de datos y la elaboración de reportes, mediante la construcción de cubos multidimensionales. La metodología seguida en el diseño del Data Warehouse fue seleccionada después de analizar los dos enfoques principales de la teoría Data Warehouse, siendo la metodología de Kimball la más apropiada, considerando el tiempo que toma el desarrollo de un Data Warehouse con ambas metodologías y los objetivos que se persiguen en el presente trabajo.
- La tabla de hechos se conformó por información agregada a nivel de sitios de los cálculos de volumen, biomasa y carbono, así como la información del número de sitios y el número de hectáreas muestreadas.
- Se definieron dos dimensiones para llevar a cabo los análisis y generar los reportes. La primera dimensión etiquetada como ESTRATOS consistió de dos niveles cuya jerarquía fue Ecosistemas y luego Comunidad Vegetal, la definición de estos niveles se basó en la definición del diseño de muestreo y la forma en que la CONAFOR presenta sus reportes. La segunda dimensión consistió de tres niveles cuya jerarquía es Estados, Municipio y Conglomerados; estos niveles, debido a la intensidad de muestreo con la cual se tomaron los datos, no se utilidad para realizar una estimación de las variables anteriormente mencionadas, pero puede servir como punto de referencia para el análisis de la información en inventarios futuros, considerando que la perspectiva que se tiene para el Inventario Nacional Forestal de que sea integrado a partir de Inventarios Forestales Estatales.
- Se analizó el volumen de madera, biomasa y carbono, para el Estado de México, mediante el procesamiento de cubos multidimensionales, a partir del Data de un Data

Warehouse. La información de la tabla de hechos fueron tomadas como el grupo de medida para la construcción de un cubo multidimensional cuyas aristas corresponden a las dos dimensiones del Data Warehouse. Aunque el cubo contiene información sobre volumen de madera, biomasa y carbono a nivel nacional, el principal análisis se centró en el Estado de México para el cual el volumen de madera fue calculado utilizando ecuaciones de volumen específicas para cada género arbóreo; las ecuaciones de volumen utilizadas fueron obtenidas del trabajo desarrollado por Méndez y De los Santos (2011).

- El proceso extracción, transformación y carga, para la presente investigación, se llevó a cabo, sin necesidad de usar un lenguaje diferente a SQL, mediante un paquete desarrollado con las herramientas presentes en el servicio de integración del Sistema Gestor de Bases de Datos SQL Server 2008.
- La dimensión tiempo que es considerada como principal en ambos enfoques de la teoría Data Warehouse, no fue considerada en la presente investigación, debido a que no se contó con la información de los inventarios anteriores.
- Los Data Warehouse representan un conjunto de técnicas y tecnología utilizados por las grande empresas para incrementar sus niveles de competencia en el mercado; este mismo conjunto de metodologías y herramientas pueden ser utilizadas por el sector gubernamental y académico para ser más eficientes el procesamiento y análisis de datos, con diferentes naturalezas, mediante una adecuación a la teoría original de los Data Warehouse.
- Un Data Warehouse es la herramienta ideal para el almacenamiento y análisis de los datos del Inventario Nacional Forestal y de Suelos ya que proporciona la flexibilidad de utilizar repositorios de información no relacionados directamente con la base de datos. Además de que en las etapas previa a la carga de datos es necesario la aplicación de técnicas de extracción, transformación y carga (ETL) para la unificación de los dominios y la eliminación de ruido en los datos (datos atípicos y datos faltantes).

Interfaces de Análisis

- Se utilizaron cuatro interfaces para la visualización y el análisis de datos procesados mediante un cubo multidimensional generado a partir de un Data Warehouse; una de las

interfaces forma parte de las herramientas del Servicio de Análisis de SQL Server, la segunda interface corresponde a una conexión hacia una hoja de cálculo de Microsoft Excel, la cual permite realizar un análisis mediante tabla y gráficos dinámicos y por último se desarrollaron dos interfaces, una de escritorio y otra web.

- El Software utilizado para el desarrollo de las interfaces de escritorio y web, fue Visual Studio 2010 y algunos complementos de DevExpress, el lenguaje utilizado fue el Visual Basic .Net.
- La principal característica de las dos interfaces desarrolladas es su facilidad de uso, permitiendo llevar a cabo de manera muy sencilla las principales operaciones permitidas en un cubo multidimensional como son: Roll up, Drill down, Slice, Dice y Pivotaje o rotación; su funcionamiento es similar al de una tabla dinámica de Microsoft Excel, permitiendo llevar a cabo un análisis dinámico, desde diferentes perspectivas, al intercambiar los diferentes criterios entre filas, columnas o utilizarlos como filtro.
- El cubo OLAP desarrollado para el análisis de volumen de madera, biomasa y carbono permitió analizar la información de diferentes maneras, usando la información de sus dimensiones como fuente principal de consulta. La información pudo analizarse por regiones a nivel de estados, municipios, conglomerados y sitios y a nivel de estratos por ecosistemas y comunidad vegetal.
- Las interfaces desarrolladas permitieron realizar un análisis eficiente del volumen de madera, biomasa y carbono.

Softwares utilizados

- La principal tarea de Microsoft SQL Server Management Studio es administrar objetos de las bases de datos y configurar objetos existentes del servicio de análisis; mientras que Business Intelligence Development Studio su principal función es para el desarrollo de aplicaciones de inteligencia de negocios.
- Las herramientas de Integration Services facilitan realizar tareas como la integración de información proveniente de diferentes fuentes electrónicas, considerando que la estructura de los datos no es la misma (formatos o sistemas). Otra tarea, que convierte a

Integration Services en una herramienta muy potente, es el poder realizar cálculos complejos y programación durante la etapa de integración.

- Las herramientas del servicio de análisis de SQL Server permiten realizar análisis de grandes volúmenes de datos de manera eficiente, comparada contra consultas tradicionales en una base de datos, mediante el procesamiento de cubos OLAP.
- La definición de un cubo OLAP mediante el servicio de análisis de SQL Server involucra crear una conexión a una fuente de datos definir las dimensiones y los grupos de medidas. Una vez procesado el cubo, se puede utilizar la herramienta de visualización y exploración de datos proporcionada por el mismo módulo, y cuyo comportamiento es similar al de una tabla dinámica de MS Excel, para llevar a cabo los análisis correspondientes; la versión 2012 de MS SQL Server integra la opción de crear una conexión hacia un hoja de cálculo de MS Excel y utilizar sus herramientas de análisis como son tablas dinámicas y gráficos dinámicos.
- Las herramientas de minería de datos del servicio de análisis de SQL Server ayudan a identificar patrones en los datos, para determinar las razones por las que suceden las cosas, o crear reglas y recomendaciones. Con el servicio de análisis de SQL Server no es necesario crear un almacén de datos para realizar tareas de minería de datos, ya que se pueden usar datos tabulares de proveedores externos, hojas de cálculo e incluso archivos de texto. También se pueden minar cubos OLAP creados con la misma herramienta.
- Visual Studio 2010 proporcionan las herramientas necesarias para desarrollar interfaces para la visualización y análisis de la información contenida en los cubos multidimensionales desarrollados con del servicio de análisis de SQL Server. Se pueden desarrollar interfaces para escritorio e interfaces web con comportamientos similares a las tablas dinámicas de Microsoft Excel.

14.2 Recomendaciones y Trabajos Futuros

La realización de este trabajo permite adelantar una serie de sugerencias y recomendaciones a los desarrolladores de software y a los generadores de la información, para trabajos futuros que permitan explotar realmente la riqueza de la gran cantidad de datos que se colectan en el ámbito forestal en México.

Recomendaciones a corto plazo

- Evaluar este software con un grupo de expertos.
- Mejorar el software desarrollado, integrando más procesos, la dimensión tiempo y las interfaces desarrolladas, para poder visualizar la información mediante gráficos dinámicos.
- Proponer cursos de actualización para los analistas de datos forestales, con respecto a los temas de Bases de Datos y Data Warehouse, para que puedan desarrollar soluciones de análisis para los diferentes procesos que se llevan a cabo en sus instituciones.
- Verificar la normalidad de las tablas de la base de datos del Inventario Nacional Forestal.
- Realizar una evaluación, de diversas características, de los softwares actuales más comunes para realizar minería de datos.
- Realizar una investigación de las instituciones que actualmente hacen uso de los Data Warehouse como una herramienta para el almacenamiento, análisis de información y toma de decisiones.

Recomendaciones a mediano plazo

- Recomendar a la CONAFOR que identifique y defina cada uno de los procesos (cálculos y estimaciones) que se quieren llevar a cabo con la información levantada en campo, con la finalidad de recolectar sólo la información necesaria y evitar incongruencias en sus datos como datos atípicos y perdidos.
- Recomendar a la CONAFOR que rediseñe la base de datos del Inventario Nacional Forestal para eliminar atributos no relacionados con algún proceso de análisis.

Recomendaciones a largo plazo

- Recomendar a la CONAFOR que se actualicen las herramientas de análisis por otras más apropiadas para manejar y analizar grandes volúmenes de datos.
- Sugerir a la CONAFOR que utilicen Data Warehouse y Minería de datos para hacer más eficientes sus consultas y métodos de análisis.
- Diseñar un Data Warehouse institucional para modelar todos los procesos requeridos, cálculo y estimaciones, de los principales parámetros e indicadores que se actualizan de manera temporal y que son usados en reportes. Aplicar el análisis multidimensional para presentar información actualizada y accesible a todos los usuarios de la información forestal y desarrollar interfaces web dinámicas que permitan, analizar los datos en línea, de manera eficiente.
- Proponer un proyecto de investigación que evalué los métodos utilizados en México para la obtención de los principales indicadores que se reportan a nivel nacional e internacional y comprarlos contra otros países latinoamericanos.

REFERENCIAS DOCUMENTALES

Aldana, B. R. (2012). Memoria documental: Inventario Nacional Forestal y de Suelos. [Disponible en: http://www.conafor.gob.mx:8080/documentos/docs/8/4125CNF-24_INFyS.pdf fecha de consulta 26/05/2014]

Bertino, E. y Martino, L. (1995). *Sistemas de bases de datos orientadas a objetos: Conceptos y arquitectura*. Massachusetts: Addison-Wesley Iberoamericana S.A.

Berry, M. J. A. y Linoff, G. S. (2004). *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*. Indianapolis: Wiley Publishing, Inc.

Caballero, D. M. (1998). El inventario forestal en México: evolución y perspectivas. North American Science Symposium. Guadalajara, México.

Carmona Mota, V. L. (2006). *Minería de datos usando SAS Entreprise Miner; una aplicación en datos forestales*. Tesis de Maestría no publicada, Colegio de Postgraduados, Programa de Socioeconomía, Estadística e Informática orientado a la Estadística.

Carpani, F. (2000). *CMDM: Un Modelo Conceptual para la Especificación de Bases Multidimensionales*. Tesis de Maestría no publicada, Universidad de la República, Instituto de Computación –Facultad de Ingeniería.

Clark, P. y Boswell, R. (2000). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann Publishers.

Colledge, R. (2010). *SQL Server 2008: Administration in Action*. Greenwick: Manning Publications Co.

CONAFOR. (2010). Visión de México sobre REDD+: Hacia una estrategia nacional. Zapopan Jalisco México.

CONAFOR. (2011). Inventario Nacional Forestal y de Suelos: Manual y procedimientos para el remuestreo de campo. [Disponible en: http://www.climateactionreserve.org/wp-content/uploads/2011/03/Sampling_Manual-_Remuestreo-_Conafor_INFyS.pdf fecha de consulta 29/05/2014]

CONAFOR. (2012). *Inventario Nacional Forestal y de Suelos. Informe de resultados 2004-2009*. Coordinación General de Planeación e Información- Gerencia de Inventario Forestal y Geomática. Zapopan, Jalisco, México.

Date, C. J. (2001). *Introducción a los Sistemas de Bases de Datos*. (7ma ed.) México: Pearson Educación S.A.

De Miguel, A., Martínez, P., Castro, E., Cavero, J. M., Cuadra, D., Iglesias, A. M. y Nieto, C. (2000). *Diseño de bases de datos: problemas resueltos*. Madrid: Editorial Ra-Ma.

De la Fuente Fernández, S. (2011). Regresión Logística. Universidad Autónoma de Madrid (UAM). Facultad de Ciencias Económicas y Empresariales. [Disponible en http://www.fuenterrebollo.com/Economicas/ECONOMETRIA/CUALITATIVAS/LOGISTICA/regresion-logistica.pdf fecha de consulta 23/09/2014].

Departamento de Lenguajes y Sistemas Informáticos. (2004). Bases de Datos: Modelo Relacional de Codd: Estructuras y restricciones. [Disponible en http://www.lsi.us.es/docencia/get.php?id=3183 fecha de consulta 04/06/2014].

Elmasri, R. A. y Navathe, S. B. (2007). *Fundamentos de Sistemas de Bases de Datos.* (5ta ed.) Madrid: Pearson Educación S.A.

FAO (2002). Evaluación de los recursos forestales mundiales 2000 - Informe principal. Estudio FAO. Roma, Italia.

Fayyad, U., Piatetsk-Shapiro, G., y Smyth. (1996). Knowledge Discovery in Databases. *Al Magazine* 17(3):37-54

Frawley, W. J., Piatetski-Shapiro, G. y Matheus, C. J. (1992). Knowledge Discovery in Databases: An Overview. *Al Magazine* 13(3):57-70

Gómez Pino, F. (2004). Los Sistemas Gestores de Bases de Datos Actuales [Disponible en http://www.slideshare.net/fsgpino/los-sistemas-gestores-de-bases-de-datos-actuales-5439890 fecha de consulta 29/08/2014].

Hansen, G. W., y Hansen, J. V. (2002). *Diseño y Administración de Bases de Datos.* (2da ed.) Prentice Hall.

Havemann, T., Negra, C. y Ashton, R. (2009). Measuring and monitoring terrestrial carbon as part of "REDD+" MRV systems. The State of the Sciencie and Implications for Policy Makers. Prepared for The Terrestrial Carbon Group, The Heinz Center, and UN-REDD programme.

Hernández Orallo, J., Ramírez Quintana, M. J. y Ferri Ramírez, C. (Eds). (2004). *Introducción a la minería de datos*. Madrid: Pearson Educación S.A.

Hernández Sampieri, R., Fernández Collado, C. y Baptista Lucio, P. (2010). *Metodología de la investigación* (5ta ed.). México: McGraw-Hill.

Hilera, J. R. y Martínez V. J**.** (1995). *Redes Neuronales Artificiales: Fundamentos, Modelos y Aplicaciones*. Madrid: Ra-Ma.

Holmgren, P. y Persson, R. (2003). Evolución y perspectiva de las evaluaciones forestales mundiales. [Disponible en: http://www.fao.org/docrep/005/y4001s/Y4001S02.html fecha de consulta 29/05/2014]

Imhoff, C., Galemmo, N., y Geiger, J. G. (2003). *Mastering Data Warehouse Design: Relational* and Dimensional Techniques. Indianapolis: Wiley Publishing Inc.

Inmon, H. W. (2005). Building the Data Warehouse. (4ta. Ed.). Indianapolis: Wiley Publishing Inc.

INE. (2006). Inventario Nacional de Gases de Efecto Invernadero 1990-2002. [Disponible en http://www.inecc.gob.mx/descargas/cclimatico/mexico_nghgi_2002.pdf fecha de consulta 11/06/2014].

INE. (2014).México y la participación de países en desarrollo en el régimen climático. [Disponible en: http://www2.inecc.gob.mx/publicaciones/libros/437/tudela.html fecha de consulta 14/06/2014].

INIFAP–FAO. (1961-1964). Inventario Forestal de México. Informe Técnico de Trabajos realizados. Vol. I. México.

INIFAP. (1984). Memoria del primer encuentro nacional sobre inventarios forestales. publicación especial núm. 45. México.

Jiawei Han, y Micheline Kamber. (2006). *Data Mining: Concepts and Techniques*. (2da. Ed.) San Francisco: Morgan Kaufmann.

Jiawei Han, Micheline Kamber y Jian Pei. (2012). *Data Mining: Concepts and Techniques*. (3ra. Ed.) San Francisco: Morgan Kaufmann.

Kimball, R. y Ross, M. (2002). *The Data Warehouse Toolkits: The Complete Guide to Dimensional Modeling*. (2da. Ed.). New York: Wiley & Sons.

Krzysztof, J. C., Witold, P. Roman, W. S. y Lukasz, A. K. (2007). *Data Mining: A Knowledge Discovery Approach*. New York: Springern Cience.

Kumar, P. (2011). Crop yield forecasting by adaptive neuro fuzzy inference system. *Mathematical Theory and Modeling* 1(3): 1-7.

Larose, D. T. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining*. New Jersey: John Wiley & Sons, Inc.

Macías Rodríguez, M. (2008). Técnicas de Minería de Datos para la Retención de Clientes en el Sector Asegurador. Comisión Nacional de Seguros y Fianzas. [Disponible en: http://www.cnsf.gob.mx/Eventos/Premios/2008%20Seguros/ANIVDELAREV.pdf fecha de consulta 16/08/2014].

MacLennan, J., Tang, Z. y Crivat, B. (2009). *Data Mining with Microsoft SQL Server 2008*. Indianapolis, Indiana: Wiley Publishing, Inc.

Maniatis D., Mollicone D. (2010). Options for sampling and stratification for national forest inventories to implement REDD+ under the UNFCCC. *Carbon Balance and Management* 5 (9): 1-14.

Mannino, M. V. (2007). *Administración de bases de datos: Diseño y desarrollo de aplicaciones*. (3ra ed.) México: McGraw Hill.

Markov, Z. y Larose, D.T. (2007). *Data Mining the Web: Uncovering patterns in web content, structure and usage.* New Jersey: John Wiley & Sons.

Martín, Q. y Rosario de Paz Santana, Y. (2007). *Aplicación de las redes neuronales a la regresión*. Madrid: Editorial la Muralla S.A.

Méndez, L. B., y De los Santos, P. H. (2011). Diagnóstico del potencial de reducción de emisiones de GEI derivadas de actividades REDD+ como un insumo a la elaboración de la Estrategia Nacional REDD+. Documento no publicado.

Mood, A. M., Graybill, F. A., y Boes, D. C. (1974). *Introduction to the Theory of Statistics*. McGraw-Hill.

Morteo, F. A., Bocalandro, N. L. E., Cascón, C. A., Cascón, H. G., Descalzo, C. D., De la Rosa, K. M., y Krauthmer, D. (2007). *Fundamentos de diseño y modelado de datos*. Buenos Aires: Ediciones Cooperativas.

Muñoz, R. M., Valdez, L. J. R, De los Santos, P. H. y Ángeles, P. G. (2012). Estimación de variables dasométricas en el bosque templado de Hidalgo, México mediante datos espectrales y del inventario nacional forestal.

Naciones Unidas. (1998). Protocolo de kyoto de la convención marco de las naciones unidas sobre el cambio climático. [Disponible en: http://unfccc.int/resource/docs/convkp/kpspan.pdf fecha de consulta 04/06/2014]

Nagabhushana, S. (2006). *Data Warehousing: OLAP and Data Mining*. New Delhi: New Age International Publishers.

Nielsen, P., White, M. y Parui, U. (2009). *Microsoft SQL Server 2008 Bible*. Indianapolis: Wiley Publishing Inc.

Olson, D. L. y Denle, D. (2008). *Advanced Data Mining Techniques*. Berlin: Springer.

Organización de las Naciones Unidas para la Agricultura y la alimentación, FAO. (2010) *Evaluación de los recursos forestales mundiales (FRA) 2010, Informe Nacional México*. Roma, Italia.

Pajares Martinsanz, G. y De la Cruz García, J. M. (Eds.) (2010). *Aprendizaje automático: un enfoque práctico.* Madrid: RA-MA Editorial.

Palmer, P. A. Montaño, J. J. (1999). ¿Que son las redes neuronales artificiales? Aplicaciones realizadas en el ámbito de las adicciones. *Revista Adicciones* 11(3): 243-255.

Ramos Martín. M. J., Ramos Martín. A. y Montero Rodríguez. F. (2006).*Sistemas Gestores de Bases de Datos.* Madrid: McGraw Hill Interamericana de España, S. A. U.

Real Academia Española de la lengua [Disponible en: http://lema.rae.es/drae/]

Reinosa, E. J., Maldonado, C. A., Muñoz, R., Damiano, L. E. y Abrutsky, M. A. (2012). *Bases de datos*. Buenos Aire Argentina: Alfaomega Grupo Editor Argentino.

Ricardo, C. M. (2004). *Bases de Datos*. México: McGraw Hill Interamericana Editores, S. A de C.V.

Ruíz, O., Jiménez, M. y Bauz, S. (2008). Disminución de Tiempo y costo en la Obtención de Información Biológica de Campo con Valoración Estadística. *Revista Tecnológica ESPOL* 21: 91-98.

Ruíz Torres, M. K. (2007). *Data Warehouse y Minería de Datos*. Manuscrito no publicado. México: Universidad Nacional Autónoma de México.

Sánchez Cañizares, S. M., Ayuso Muñoz, M. A. y Caridad y Ocerin, J. M. (2005). Software de minería de datos: Análisis de características. Conferencia IADIS Ibero-Americana WWW/Internet 2005.

Santizo Rincón, J. A. (2001). *Evolución y perspectivas en la metodología de la enseñanza de los recursos de servicio de estadística en el Colegio de Postgraduados*. Tesis de Doctorado no publicada, Colegio de Postgraduados, Instituto de Socioeconomía, Estadística e Informática, Especialidad en Estadística.

Savin, I., Stathakis, D., Negre, T. y Isaev, V. A. (2007). Prediction of crop yields with the use of neural networks. *Russian Agric. Sci*. 33(6): 361-363.

SEMARNAT. (2002). Inventarios forestales y tasas de deforestación. México. [Disponible en: http://app1.semarnat.gob.mx/dgeia/informe_04/02_vegetacion/recuadros/c_rec3_02.htm fecha de consulta 29/05/2014]

SEMARNAT. (2002). Informe de la situación del medio ambiente en México. Compendio de estadísticas ambientales. México. [Disponible en http://www.ibiologia.unam.mx/pdf/directorio/z/introduccion/com_estad_amb.pdf fecha de consulta 30/05/2014]

SEMARNAT. (2005). Informe de la situación del medio ambiente en México. Compendio de estadísticas ambientales. México.

Sheinbaum C., Masera O. (2000). Mitigating carbon emissions while advancing national development priorities: the case of Mexico, climatic change [Disponible en https://www.google.com/url?q=http://unfccc.int/files/meetings/workshops/other_meetings/application/vnd.ms-powerpoint/jm_mexico.ppt&sa=U&ei=Ty0UUqXaOtLo2gXM84G4Aw&ved=0CAcQFjAA&client=internal-uds-cse&usg=AFQjCNH6lWXMekd_HsWaHl49psZ8WX9xaw fecha de consulta 30/05/2014]

Silberschatz, A., Korth, H. F. y Sudarshan, S. (2002). *Fundamentos de Bases de Datos*. (4ta. Ed.). Madrid: Mcgraw-Hill/Interamericana de España, S. A.

Stastny, J., Konecny, V., y Trenz, O. (2011). Agricultural data prediction by means of neural network. *Agric. Econ.–Czech*. 57(7): 356-361.

Sumathi, S. y Sivanandam, S. N. (2006). *Introduction to Data Mining and its Applications*. New York: Springer.

Untaru, M., Rotarescu, V., y Dorneanu, L. (2012). Artificial neural networks for sustainable agribusiness: a case study of five energetic crops. Ioan Slavici University, 144, Dr. A. Paunescu-Podeanu Street, 300587, Timisoara, Romania. *Agrociencia* 46(5): 507-518.

Veerman, E., Lachev, T. y Sarka, D. (2009). *Exam 70-448: TS: Microsoft SQL Server 2008-Business Intelligence Development and Maintenance*. Washington: Microsoft Press

Velazco, B. E., Ramírez, M. H., Moreno, S. F. y De la Rosa, V. A. (2003). Estimadores de razón para el Inventario Forestal Nacional de México. Revista Mexicana de *Ciencias Forestales* 28(94):23-43.

Wang, J. (Ed.) (2009). *Enclyclopedia of Data Warehousing and Mining* (2da. Ed.). New York: Information Science Reference.

Witten, I. H., y Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. (2da. Ed.) San Francisco: Morgan Kaufmann Publisher

Yelitza, J. M. y Talavera, P. R. (2007). Minería de datos como soporte a la toma de desiciones empresariales. *Opcion año 23* 52: 104-118.

Zhu, D. (2009). Analytical Competition for Managing Customer Relations. En Wang, J. (Ed.), *Enclyclopedia of Data Warehousing and Mining* (pp. 25-30) (2da. Ed.). New York: Information Science Reference.

Zon, R. (1910). The forest resources of the world. United States Department of Agriculture, Forest Service Bulletin No. 83. Washington, D.C., Estados Unidos de América, Government Printing Office.

ANEXOS

Anexo 1. Guion de entrevistas a profesores investigadores del Colegio de Postgraduados.

COLEGIO DE POSTGRADUADOS

INSTITUCION DE ENSEÑANZA E INVESTIGACION EN CIENCIAS AGRÍCOLAS
CAMPUS MONTECILLO
POSTGRADO DE SOCIOECONOMÍA, ESTADÍSTICA E INFORMATICA
COMPUTO APLICADO

1. ¿Considera que la información de la base de datos del Inventario Nacional Forestal y de Suelos 2004-2009 es fácilmente accesible? Si | No

   ¿Por qué?

2. ¿Qué información del Inventario Nacional Forestal es la que más utiliza?
3. ¿Qué problemas se presentan frecuentemente al procesar la información del Inventario Nacional Forestal?
4. ¿Qué softwares utiliza para analizar los datos del Inventario Nacional Forestal?
5. ¿Qué propondría usted para mejorar la calidad de los datos del Inventario Nacional Forestal?
6. ¿Considera que se puede mejorar la estructura de la base de datos del Inventario Forestal Nacional para obtener información de manera más eficiente y confiable?
7. ¿Qué propondría usted para mejorar el acceso a los datos del Inventario Nacional Forestal?
8. ¿Ha escuchado sobre las base de datos multidimensionales?
9. ¿Conoce usted qué son los Data Warehouse? Si | No
10. ¿Ha escuchado sobre el tema de minería de datos? Si | No

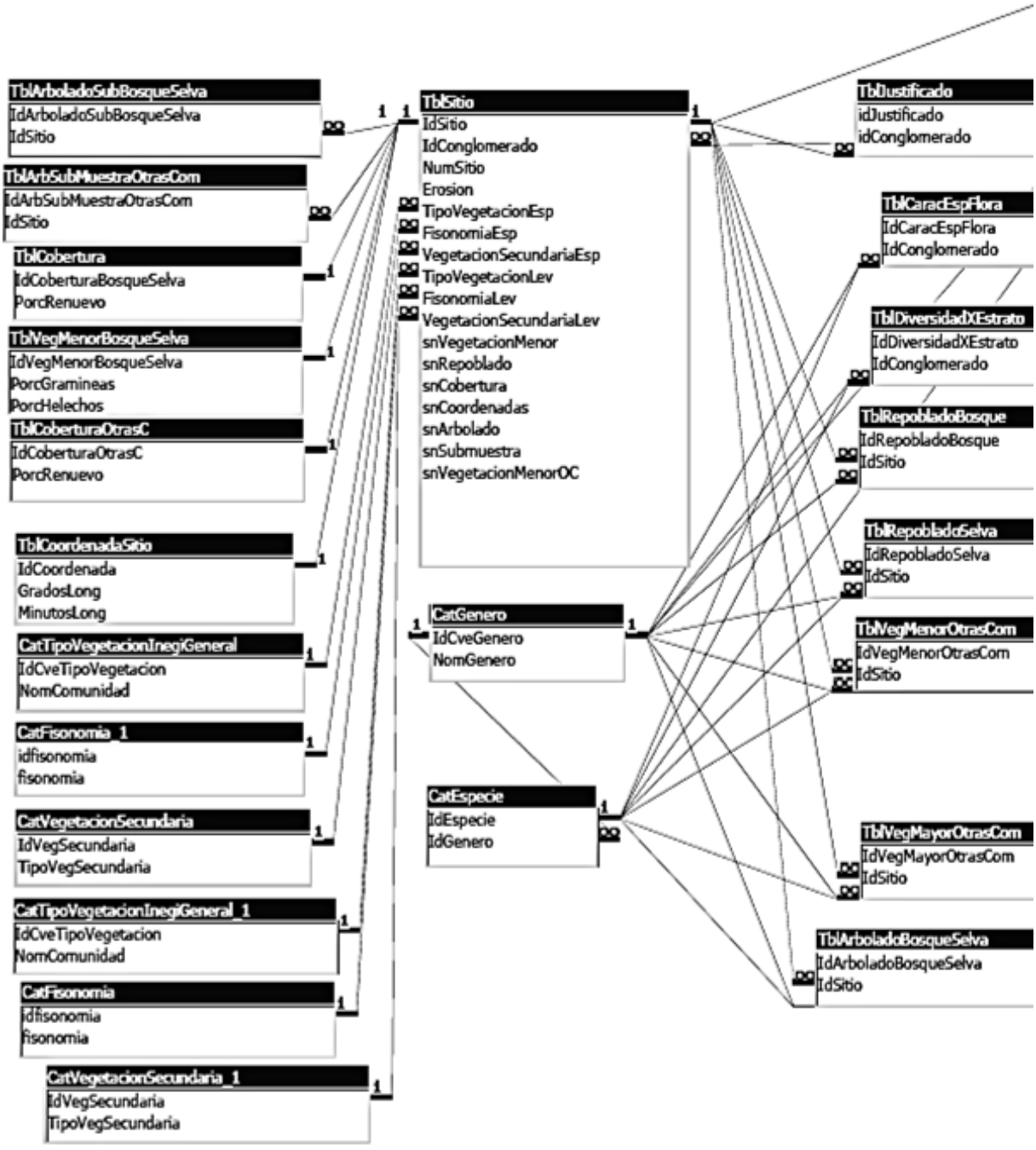
Anexo 2. Rúbricas para evaluación de las interfaces de análisis.

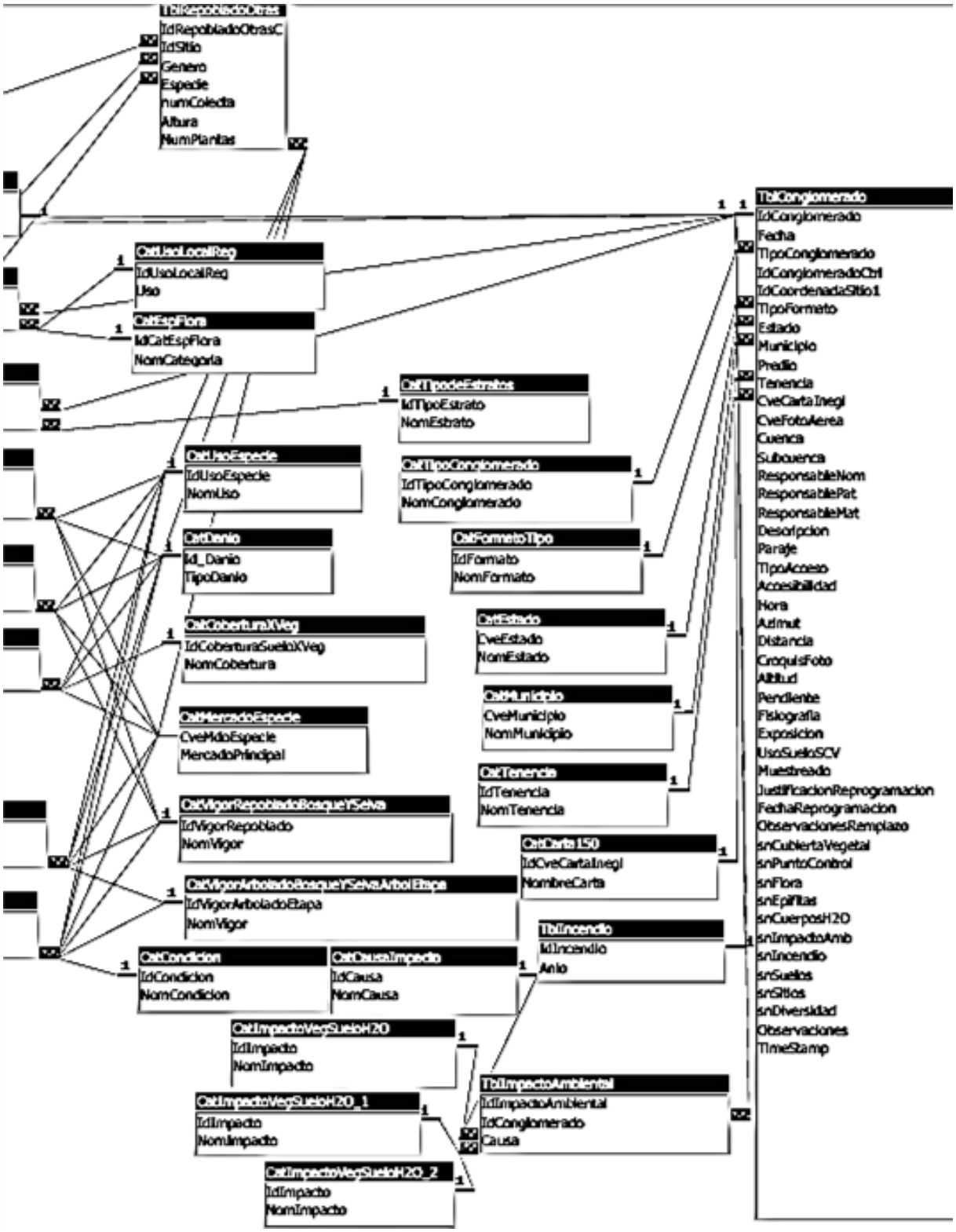
COLEGIO DE POSTGRADUADOS

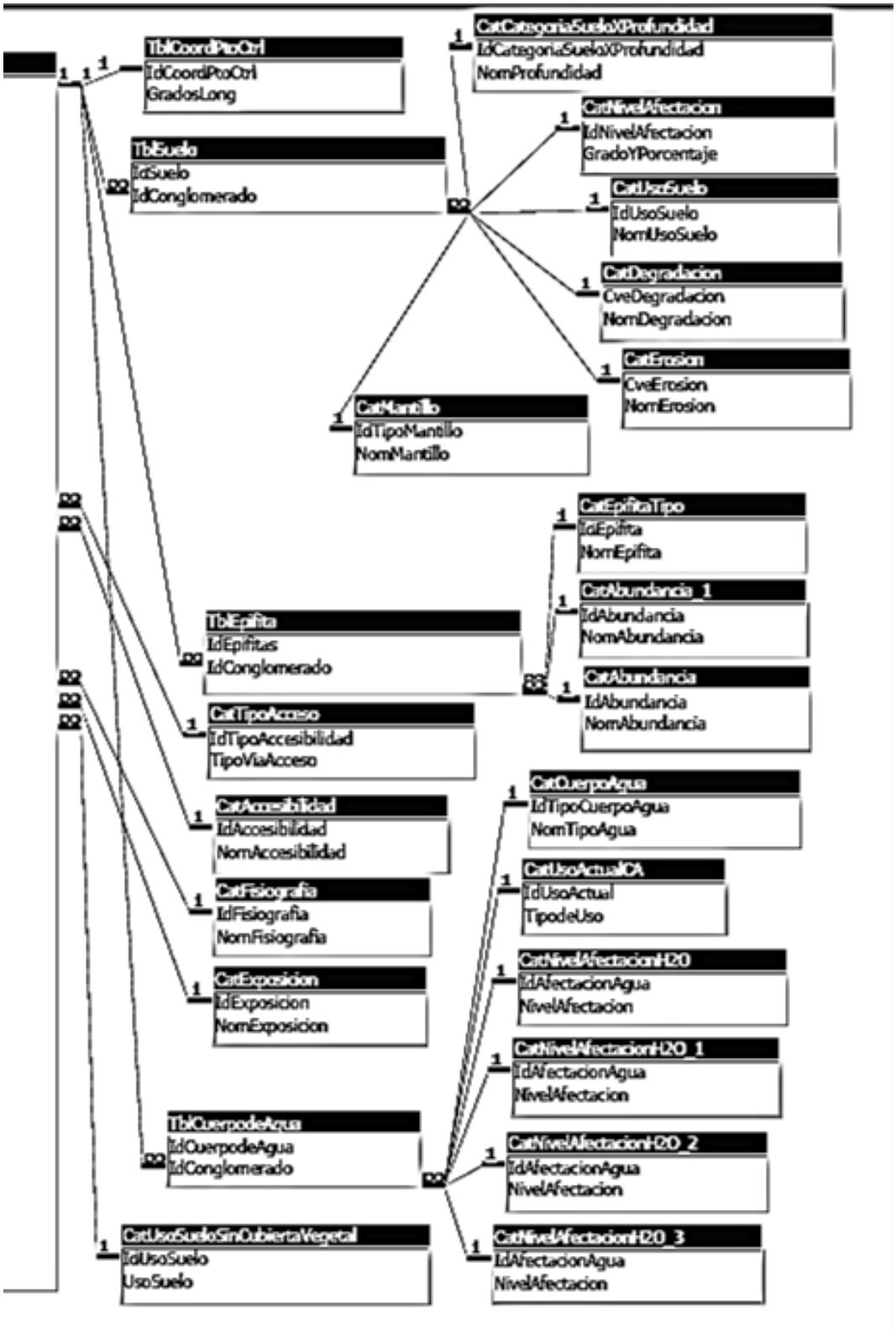
INSTITUCION DE ENSEÑANZA E INVESTIGACION EN CIENCIAS AGRÍCOLAS
CAMPUS MONTECILLO
POSTGRADO DE SOCIOECONOMÍA, ESTADÍSTICA E INFORMATICA
COMPUTO APLICADO

Categoría	1	2	3	4	Valor
Facilidad de uso	Se necesita ser un experto en el uso de computadoras para usar las interfaces	Se necesita algún entrenamiento para usar las interfaces	No se necesitan grandes conocimientos de computación para usar las interfaces.	Cualquier persona puede usar las interfaces	
Capacidad de personalización	Las interfaces no pueden ser adaptadas de acuerdo a las necesidades de análisis.	Algunas partes de las interfaces pueden ser adaptadas de acuerdo a las necesidades de análisis.	La mayor parte de las interfaces pueden ser adaptadas de acuerdo a las necesidades de análisis.	Cualquier parte de las interfaces puede ser adaptada de acuerdo a las necesidades de análisis.	
Accesibilidad	Sólo se puede acceder a las interfaces instalándolas en la computadora.	Se puede acceder a las interfaces desde internet o instaladas a partir de un CD.	Se puede acceder a las interfaces desde internet.	Las interfaces son accesibles desde internet. No se requiere su descarga.	
Lectura de resultados	Resultados difíciles de leer.	Los resultados presentan alguna dificultad en la lectura.	Los resultados no presentan dificultad en su lectura.	Los resultados son muy fáciles de leer.	
Originalidad	Las interface no presentan un diseño original.	Las interfaces presentan un diseño ligeramente original.	Las interfaces presentan un diseño original	Las interfaces presentan un diseño muy original.	
Uso de tecnología avanzada	Las interfaces no hacen uso de tecnología avanzada.	Las interfaces hacen ligero uso de tecnología avanzada.	Las interfaces hacen uso de tecnología avanzada.	Las interfaces hacen uso de tecnología muy avanzada.	
Utilidad para el análisis de datos	Las interfaces no son útiles para el análisis de datos.	Las interfaces resultan poco útiles para el análisis de datos.	Las interfaces son útiles para el análisis de datos.	Las interfaces son muy útiles para el análisis de datos.	

Anexo 3. Base de Datos del Inventario Nacional Forestal y de Suelos 2004-2009.




IdRepobladoObrasC
IdSitio
Genero
Especie
numColecta
Altura
NumPlantas
IdUsoLocalReg
Uso
IdCatEspFlora
NomCategoria
IdTipoEstrato
NomEsbrato
IdUsoEspecie
NomUso
IdTipoConglomerado
NomConglomerado
Id_Danio
TipoDanio
IdFormato
NomFormato
IdCoberturaSueloXVeg
NomCobertura
CveEstado
NomEstado
CveMunicipio
NomMunicipio
CveMdoEspecie
MercadoPrincipal
IdTenencia
NomTenencia
IdVigorRepoblado
NomVigor
IdCveCartaInegi
NombreCarta
IdVigorArboladoEtapa
NomVigor
IdIncendio
Anio
IdCondicion
NomCondicion
IdCausa
NomCausa
IdImpacto
NomImpacto
IdImpacto
NomImpacto
IdImpacto
NomImpacto
IdImpactoAmbiental
IdConglomerado
Causa
IdConglomerado
Fecha
TipoConglomerado
IdConglomeradoCbtl
IdCoordenadaSitio1
TipoFormato
Estado
Municipio
Predio
Tenencia
CveCartaInegi
CveFotoAerea
Cuenca
Subcuenca
ResponsableNom
ResponsablePat
ResponsableMat
Descripcion
Paraje
TipoAcceso
Accesibilidad
Hora
Azimut
Distancia
CroquisFoto
AltitUd
Pendiente
Fisiografia
Exposicion
UsoSueloSCV
Muestreado
JustificacionReprogramacion
FechaReprogramacion
ObservacionesRemplazo
snCubiertaVegetal
snPuntoControl
snFlora
snEpifitas
snCuerposH2O
snImpactoAmb
snIncendio
snSuelos
snSitios
snDiversidad
Observaciones
TimeStamp


TblCoordPtoCtrl
IdCoordPtoCtrl
GradosLong
CatCategoriaSueloXProfundidad
IdCategoriaSueloXProfundidad
NomProfundidad
CatNivelAfectacion
IdNivelAfectacion
GradoYPorcentaje
TblSuelo
IdSuelo
IdConglomerado
CatUsoSuelo
IdUsoSuelo
NomUsoSuelo
CatDegradacion
CveDegradacion
NomDegradacion
CatErosion
CveErosion
NomErosion
CatMantillo
IdTipoMantillo
NomMantillo
CatEpifitaTipo
IdEpifita
NomEpifita
CatAbundancia_1
IdAbundancia
NomAbundancia
TblEpifita
IdEpifitas
IdConglomerado
CatAbundancia
IdAbundancia
NomAbundancia
CatTipoAcceso
IdTipoAccesibilidad
TipoViaAcceso
CatQuerpoAgua
IdTipoQuerpoAgua
NomTipoAgua
CatAccesibilidad
IdAccesibilidad
NomAccesibilidad
CatUsoActualCA
IdUsoActual
TipodeUso
CatFisiografia
IdFisiografia
NomFisiografia
CatNivelAfectacionH20
IdAfectacionAgua
NivelAfectacion
CatExposicion
IdExposicion
NomExposicion
CatNivelAfectacionH20_1
IdAfectacionAgua
NivelAfectacion
TblQuerpodeAgua
IdQuerpodeAgua
IdConglomerado
CatNivelAfectacionH20_2
IdAfectacionAgua
NivelAfectacion
CatUsoSueloSinCubiertaVegetal
IdUsoSuelo
UsoSuelo
CatNivelAfectacionH20_3
IdAfectacionAgua
NivelAfectacion

Anexo 4. Relevancia de la investigación como soporte en la Estrategia Nacional REDD+.

Los efectos del cambio climático actualmente causan estragos alrededor del mundo, mediante las alteraciones en la temperatura, los regímenes de precipitación y aumento de la frecuencia de eventos extremos, que afectan de manera significativa a las sociedades humanas; otro tipo de afectaciones se refleja en la regulación de los ciclos hidrológicos, promoviendo de ésta forma la degradación de los bosques y la estabilidad de las reservas de carbono forestal. Los bosques capturan alrededor de 5, 000 millones de toneladas de dióxido de carbono de los 32, 000 millones que se emiten anualmente como resultado de las actividades humanas. Los bosques tropicales contienen aproximadamente el 40% del carbono acumulado en la biomasa terrestre por lo que cualquier perturbación de estos ecosistemas podría causar un cambio significativo en el ciclo de carbono mundial; lo que hace realmente importante evitar la deforestación y degradación de los bosques y aumentar el acervo de carbono como una medida para combatir el calentamiento global. En el contexto internacional, enfocándose a la mayor parte de los países en desarrollo, el proceso de degradación y deforestación contribuye a contrarrestar, con casi 20% del total de emisiones de gases de efecto invernadero (GEI). En México, estimaciones preliminares de las emisiones de GEI, provenientes del uso de suelo, cambio de uso de suelo y silvicultura, llevados a cabo para el periodo 1990-2002, arrojó que contribuían en un 14% de las emisiones totales y para el 2006 se estimó en 9.9% del total nacional (CONAFOR, 2010).

Los bosques representan un recurso vulnerable cuya conservación, manejo y restauración ofrecen una gran oportunidad para contrarrestar el cambio climático, mediante la reducción de emisiones derivadas de la deforestación y degradación, es por ello que en la Convención Marco de la Naciones Unidad sobre el Cambio Climático (CMNUCC) celebrado en Kioto, Japón, el 11 de diciembre de 1997 se firmó el Protocolo de Kioto que representa un acuerdo internacional con el objetivo de reducir las emisiones de seis gases de efectos invernadero que causan el calentamiento global. En esta convención se estableció el compromiso obligatorio de cumplimiento, cuando los países industrializados responsables de, al menos, un 55 % de las emisiones de CO_2 lo ratificaran, Naciones Unidas (1998). Los gases que están contemplado dentro de este documento son:

- Dióxido de carbono (CO_2)
- Gas metano(CH_4)
- Óxido nitroso (N_2O)
- Hidrofluorocarbonos (HFC)
- Perfluorocarbonos (PFC)
- Hexafluoruro de azufre (SF_6)

En la décima tercera Conferencia de las partes (COP 13), en 2007, se aprobó el plan de acción de Bali, en el cual se tomó la iniciativa de establecer un esquema de la Reducción de las emisiones ocasionadas por Deforestación y Degradación de los bosques (REDD), posteriormente se incluyeron los temas de la conservación, el manejo sustentable de los bosques y el mejoramiento de inventarios de carbono convirtiéndose en (REDD+) (CONAFOR, 2010).

El programa REDD+ tiene como propósito que la sociedad haga conciencia de que los bosques y selvas pueden tener un alto valor ecológico y financiero cuando se conservan en pie debido al carbono almacenado en los árboles y otros organismos vivos que contienen, así como en el material orgánico que subyace en el suelo. La valoración de los bosques y selvas por el carbono que conservan en sus diferentes reservorios (biomasa aérea, biomasa subterránea, carbono mineral) necesariamente implica su cuantificación (Havemann *et al.*, 2011). En una fase avanzada, el mecanismo REDD+ incluye el pago de compensaciones a los países en desarrollo por el carbono almacenado en sus bosques. Por otro lado, se contempla el establecimiento de un equilibrio económico que incentive una gestión sostenible de los bosques y selvas para que los bienes y servicios económicos, medioambientales y sociales que puedan beneficiar a países, comunidades, biodiversidades y usuarios de los bosques, mientras contribuyen en la reducción de emisiones de gas de efecto invernadero (UN-REDD, 2009).

En la décimo quinta Conferencia de las Partes (COP 15), en 2009, se decisión que los países que deseen participar en el mecanismo de mitigación REDD+ deberían establecer un sistema confiable de monitoreo forestal, el cual soportaría el requerimiento de Monitoreo, Reporte y Verificación (MRV) (Maniatis y Mollicone, 2010). Por lo anterior, los sistemas MRV se han convertido en un elemento crítico para la implementación exitosa de un mecanismo REDD+ en

cualquier nación. A pesar del esfuerzo realizado por varios países para desarrollar un sistema de esta naturaleza que cumpla con las especificaciones requeridas, a la fecha no existe un sistema que pueda considerarse genérico y de utilidad para la mayoría de los países con posibilidades de obtener recursos de REDD+ (MRV México, 2014).

México es miembro de la Conferencia de las Partes de la UNFCCC desde 1992 –país en vía de desarrollo con obligación de reportar emisiones de GEI a la CMNUCC– y firmante del protocolo de Kyoto desde junio 9 de 1998, con aprobación del senado mexicano el día 29 de abril de 2000 (Sheinbaum y Masera, 2000). El protocolo de Kioto fue el punto de partida para que muchos países volcarán la vista hacia el cuidado y protección del medio ambiente y emprendieran una serie de programas y evaluaciones para cumplir con los resultados propuestos. Por ello es necesario reportar la existencia y los cambios temporales en los niveles de carbono de diversos reservorios contenidos en sus bosques y selvas. De acuerdo con INE (2014) para realizar las tareas anteriores se emprendió una serie de acciones a través de políticas públicas e involucrando a un gran número de instituciones como la Secretaría de Medio ambiente y Recursos Naturales (SEMARNAT) y sus dependencias como la Comisión Nacional Forestal (CONAFOR), la Comisión Nacional de Áreas Naturales Protegidas (CONANP), la Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (COBABIO), el Instituto Nacional de Ecología (INE), la Procuraduría de Protección al Ambiente (PROFEPA), la Comisión Nacional del Agua (CONAGUA) y la Secretaría de Agricultura, Ganadería, Desarrollo Rural, Pesca y Alimentación (SAGARPA). Alguno documentos que se pueden consultar al respecto son "La Visión de México sobre REDD+", "Propuesta de preparación REDD (R-PP), "Inventario Nacional de Emisiones de Gases Efecto Invernadero 1990 - 2006" y la Estrategia Nacional REDD+.

En CONAFOR (2012) se plantea el camino que México sigue con la Estrategia Nacional REDD+, en ésta se define cinco líneas de acción las cuales son:

- Arreglos institucionales y políticas públicas.
- Esquemas de financiamiento.
- Nivel de referencia forestal y Sistema de medición, reporte y verificación.

- Desarrollo de capacidades.
- Comunicación, participación social y transparencia.

En este sentido la presente investigación se centra en aportar ideas y metodologías útiles para la línea estratégica relacionada con el nivel de referencia forestal y el sistema MRV. La construcción del nivel de referencia forestal requiere de una combinación de datos históricos y recientes sobre emisiones asociadas a la deforestación y/o degradación de los bosques, y a la de otros usos de suelo relevante, así como la estimación de emisiones y capturas futuras que existirían en el país si no se contara con incentivos adicionales para REDD+. Uno de los resultados esperados permite evaluar la tendencia histórica de deforestación y degradación de los bosques y los consiguientes cambios en la densidad del carbono. El seguimiento al desempeño de las medidas y su comparación con el nivel de referencia requiere que México establezca, opere y mantenga un sistema MRV para REDD+. Este sistema producirá información para el diseño de políticas de uso de suelo, permitirá la obtención de incentivos con base en resultados y facilitará el cumplimiento de los compromisos de reporte de datos del país ante foros internacionales (CONAFOR, 2010).

Un tema relevante para el establecimiento del sistema MRV, es el balance entre el grado de precisión en las mediciones y el monitoreo y el costo asociado, lo que implica probar nuevas tecnologías para lograr un equilibrio entre estas dos restricciones, es aquí donde radica la importancia de la presente investigación que propone el uso de un Data Warehouse diseñado a partir del Inventario Nacional Forestal y de Suelos 2004-2009, para el almacenamiento de la información histórica y reciente y garantice un acceso rápido de ésta, para su análisis y que además permita realizar usar cálculos complejos y elaborar reportes.

Anexo 5. Requisitos de hardware y software para instalar SQL Server 2008 Enterprise

Requisitos de Hardware para instalar SQL Server 2008 Enterprise

Requisitos mínimos para instalar SQL Server 2008 Enterprise en un Sistema Operativo Windows de 32 bits

Una computadora Pentium III con procesador de 1 GHz (Se recomienda un procesador de 2 GHz o superior).
512 MB de memoria RAM (se recomienda 2 GB o más)
2.1 GB espacio libre en disco duro

Requisitos de Software para instalar SQL Server 2008 Enterprise

SQL Server 2008 puede ser instalado en muchas versiones de Windows Server y Sistemas Operativos de escritorio entre los que se incluyen: Windows XP (con Service Pack 2 [SP2] o posteriores), Windows Server 2003 (con SP2), Windows Vista, y Windows Server 2008.