BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA

FACULTAD DE CIENCIAS DE LA COMPUTACIÓN

Un modelo de solución de inteligencia de negocios para el estudio del CO_2 en el cambio climático

Tesis presentada para obtener el título de:
MAESTRO EN CIENCIAS DE LA COMPUTACIÓN

Presenta:
ELIZABETH SABINO MOXO

Asesor de tesis: DR. ABRAHAM SÁNCHEZ LÓPEZ

Coasesor de tesis: DRA. MARÍA TERESA TORRIJOS MUÑOZ

OCTUBRE 2015

Agradecimientos

Agradezco a mi familia por apoyarme en todo el camino de mi vida estudiantil en licenciatura y en maestría.
A mi asesor de tesis el Dr. Abraham Sánchez López y mi coasesora la Dra. María Teresa Torrijos Muoz que me guiaron en este camino motivándome, siempre en la mejor disposición, respeto, y más importante, su ejemplo.
A mis compaeros de generación de Computación Matemática por su grata convivencia.
Alejandro López López por su invaluable apoyo en mi vida universitaria como amigo y prometido.
Al Consejo Nacional de Ciencia y Tecnología por el apoyo económico que me brindó durante estos dos aos de estudios de la maestría.

Resumen

En los últimos años el cambio climático se ha vuelto cada vez más evidente debido al alcance mundial de sus consecuencias y su impacto potencial sobre el patrimonio cultural y natural de la humanidad. El cambio climático se estudiará en la obtención de modelos de predicción que permitan conocer el comportamiento a corto o largo plazo de un fenómeno, esta ardua tarea realizada por científicos de varias disciplinas no sólo requiere de la habilidad de usar datos en un modelo matemático, si no que depende en gran medida de un tratamiento previo de grandes cantidades de datos en crudo, que permiten al científico hacer observaciones puntuales que colaboran en la formulación de acciones que permitan probar o complementar su hipótesis, además del manejo de herramientas de software que requiere de cierto grado de destreza que sólo puede obtenerse durante largas horas de trabajo. El trabajo de investigación propone un modelo de solución de inteligencia de negocios que ayude a este problema originalmente planteado por científicos del NOAA, por medio de la realización de un modelo de predicción de concentraciones de CO_2 a partir de datos mediante árboles de decisión, además proponiendo una estructura de almacenamiento de datos que facilite su análisis.

Índice general

Agradecimientos III

RESUMEN V

Índice general VII

Índice de figuras IX

Índice de cuadros 1

1. Introducción 3

1.1. Antecedentes y trabajos relacionados 4

1.2. Justificación 7

1.3. Contribución 7

2. Estado del arte 11

2.1. El problema del cambio climático 11

2.2. Grandes almacenes de datos 12

2.3. Dato, información y conocimiento 14

2.4. Particularidades en la atmósfera 14

2.4.1. GEI 16

2.5. La importancia del estudio del dióxido de carbono 20

3. Inteligencia de negocios aplicada en los datos del NOAA 23

3.1. Data warehouse 24
3.2. ¿ Por qué utilizar BI? 26
3.3. Dominio de la aplicación 29
3.4. Selección de datos 29
3.5. Preproceso de datos 32
3.6. Transformación 34
3.7. Minería de datos 36
3.7.1. Técnicas de minería de datos 38
3.7.2. Árboles de decisión 39
3.8. Árboles autorregresivos 41
3.9. Marco de aprendizaje 42
3.10. La puntuación bayesiana de un modelo ART 44

4. Propuesta de los modelos de predicción 47

4.1. Metodología 47
4.2. Diagnóstico de un modelo de regresión 48
4.3. Construyendo un modelo de árboles de decisión 49
4.3.1. Método de puntuación 49
4.3.2. Establecer los regresores 51
4.3.3. Umbral de crecimiento del árbol de decisión 51
4.4. Modelos de predicción basado en árboles de decisión 51
4.4.1. Prediciendo con valores conocidos 60
4.4.2. Interpretación del modelo 62

5. Conclusiones y trabajo a futuro 67

Bibliografía 69

Índice de figuras

2.1. Centros de monitoreo pertenecientes la división global de monitoreo ESRL. 13

3.1. Arquitectura del proceso de inteligencia de negocios. 26
3.2. Proceso de generación del conocimiento. 28
3.3. Sitio Web del ESRL. 30
3.4. Ejemplo de un documento suministrado por el ESRL del monitoreo de CO_2. 34
3.5. Estructura que almacena los datos proporcionados por el ESRL. 35
3.6. Estructura general multidimensional resultado de la transformación de datos. 35
3.7. Estructura multidimensional resultado de la transformación de datos aplicado al caso de estudio. 36
3.8. Ejemplo de un árbol tipo AR. 42

4.1. Árbol de decisión correspondiente al modelo x. 52
4.2. Gráfica de correlación correspondiente al modelo x. 53
4.3. Diagramas de correlación pertenecientes a algunos modelos de árboles de decisión construidos 56
4.4. Gráfica de dispersión correspondiente al modelo e. 57
4.5. Gráfica de relación entre el residuo y el valor predicho. 57

4.6. Comportamiento normal de los datos predichos por el Modelo e. 59
4.7. Árbol correspondiente al Modelo e. 60
4.8. Predicciones para la concentración de CO_2 obtenidas mediante el Modelo e para el año 2013 61
4.9. Alemania(○), Canadá(x), China(●), E.U(◇), México(□), Reino Unido(△). 63
4.10. Alemania(○), Canadá(x), China(●), E. U(◇), México(□), Reino Unido(△). 64

Índice de cuadros

4.1. Variaciones y resultados de las variaciones en el algoritmo de árboles de decisión para obtener el modelo x. 52
4.2. Variaciones y resultados de las variaciones en el algoritmo de árboles de decisión en la generación de modelos. 55
4.3. Error estándar de estimación para los modelos de predición. 58

1 Introducción

Existen pocas dudas sobre el cambio climático global y sus consecuencias. Los gases de efecto invernadero(GEI) son asociados con la mayoría del aumento de temperatura observada desde mediados del siglo XX, y los fenómenos naturales tales como la variación solar ha tenido un pequeño efecto de calentamiento y un pequeño efecto de enfriamiento posterior a 1950[23]. El cambio climático no es un fenómeno que pueda ser estudiado de manera aislada ya que su impacto es global, es necesario contar con la mayor cantidad de datos acerca de concentraciones de sustancias o contaminantes presentes en la atmósfera, para este efecto se han creado redes de monitoreo atmosférico globales como los pertenecientes a la ESRL(Global Monitoring Division) que alberga gran cantidad de datos que son compartidos con toda la comunidad científica. El cambio climático se ha convertido en un tema de estudio científico de enfoque multidisciplinario debido a sus consecuencias, cuyo éxito depende en gran medida de un tratamiento previo de grandes cantidades de datos en crudo, que permiten a un científico hacer observaciones puntuales que colaboran en la formulación de acciones que permitan probar, refutar o complementar su hipótesis de investigación. Sin embargo, en búsqueda de la realización de estos trabajos de origen científico es crucial el manejo de grandes cantidades de datos, la estructura actual de ellos dificulta el trabajo de investigación del cambio climático en las disciplinas de interés, lo que hace necesario que el investigador posea habilidades en el manejo de datos que van más allá de la destreza en el uso de herramientas de software que a su vez es una parte básica y por lo tanto esencial en el tratamiento de datos, con la finalidad de realizar un análisis que permita generar información y conoci-

miento. Adquirir las habilidades necesarias para todo lo referido al tratamiento de datos no es una labor imposible, sin embargo implica un costo de tiempo en la obtención de dichas habilidades que bien podrían ser empleados en el tema principal del trabajo de investigación.

1.1. Antecedentes y trabajos relacionados

El cambio climático es uno de los grandes problemas científicos de origen antropogénico de nuestra era. El cambio climático es un proceso natural que hasta antes del abuso de combustibles de origen fósil[40] se había llevado a cabo paulatinamente, permitiendo a los seres vivos que habitamos en el planeta adaptarnos de forma natural, sin embargo las emisiones de GEI han incrementado considerablemente y son éstas las que junto con el abuso de aerosoles han provocado que el cambio climático se acelere. Existen cuatro factores determinantes en el cambio climático; la temperatura, el cambio en la precipitación, el aumento en el nivel del mar y los sucesos extremos, lo cual genera impacto y deja vulnerables los ecosistemas, los recursos hídricos, la seguridad alimentaria y la salud humana, frenando el desarrollo económico de una región o nación que necesita apoyarse en el uso de nuevas tecnologías, generar nuevas pautas de producción y consumo así como de comercio y socioculturales[2].
Reportes de alcance mundial y nacional[2] dicen que el cambio climático modificara la disponibilidad del agua intensificando el ciclo hidrológico. Las sequías y las inundaciones serán más intensas en numerosas zonas. Habrá más lluvias en latitudes altas, menos precipitaciones en el are tropical seca y cambios inciertos en las regiones tropicales. La elevación del nivel del mar provocará inundaciones así como la perdida de pantanos y humedales. La erosión de las costas permitirá acceso del agua salada en la superficie de terreno y en las aguas subterráneas, además, las diferencias de disponibilidad hídrica entre regiones se harán cada vez más pronunciadas[11]. La comunidad científica está trabajando en la construcción de modelos del sistema climático de la tierra, adquiriendo datos de la observación, así como de manera remota mediante sensores, lo que facilita la toma de decisiones mediante la construcción de modelos para predecir el cambio climático. Los científicos analizan los datos para fórmular y probar hipótesis

en los fenómenos presentes en el ambiente. Las hipótesis robustas llegan a formar parte la base de conocimiento para fórmular modelos, mismos que llegan a ser probados mediante la experimentación para entender las relaciones causales y eventualmente, para así construir modelos climáticos globales.

Hoy en día, las organizaciones manejan un flujo de información que era inimaginable apenas unos años atrás, debido al incremento en el interés de datos climatológicos que se recopilan de diversas fuentes ya que se pueden organizar y almacenar en repositorios que permiten generar estructuras dinámicas, que a su vez pueden permitir el proceso analítico en línea para generar conocimiento que funciona como insumo para las herramientas de uso táctico y estratégico. Un ejemplo de este tipo de repositorio que nace para hacer frente a una necesidad creciente de acceso remoto a altos volúmenes de modelos y datos es NOMADS(NOAA Operational Model Archive and Distribution System) iniciada por el Centro Nacional de Datos Climáticos (NCDC), el Centro Nacional para la Predicción del Medio Ambiente (NCEP) y el Laboratorio Geofísico de Dinámica de Fluidos (GFDL) para abordar las necesidades de acceso a modelos de datos como se indica en el Programa de Estados Unidos de Investigación Meteorológica (USWRP). NOMADS[3] fue desarrollado para facilitar modelos climáticos y datos observados. NOMADS se está desarrollando como "Un clima Unificado y Archivo de Clima"para que los usuarios puedan tomar decisiones acerca de sus necesidades específicas en escalas de tiempo de días (clima), al mes, a décadas (calentamiento global).

NOMADS es una red de servidores de datos que usan tecnologías para acceder e integrar modelos y otros datos almacenados en repositorios distribuidos geográficamente en formatos heterogéneos. NOMADS permite el intercambio y la comparación entre los modelos resultado y es un esfuerzo importante en colaboración de múltiples agencias gubernamentales y las instituciones académicas. Los datos disponibles en el marco NOMADS incluyen el modelo de entrada y de predicción numérica del tiempo (NWP), y Modelos Climáticos Globales (GCM) y simulaciones de GFDL y otras instituciones líderes de todo el mundo.

Las ventajas que ofrece el gran cúmulo de datos son enormes, sin embargo adolece de una manera de buscar información de forma precisa, lo suficientemente clara como para poder realizar deducciones que nos lleven a realizar acciones de mitigación o adaptación

en aras del cambio climático, se trata entonces de convertir los datos en información, y a su vez la información en conocimiento. En la búsqueda de construir modelos de predicción de datos que permitan conocer el comportamiento, los investigadores requieren de datos que estén bajo el método científico. PDS (Planetary Data System) [10] es un ejemplo de herramienta que puede entregar datos que pueden ser trabajados directamente con el método científico, sin embargo no existe una herramienta que entregue la misma calidad de datos de tipo climatológico. En la búsqueda de comprender el origen de los fenómenos relacionados con el ambiente está implícito el análisis de cada una de las variables contenidas en ella, puesto que el fenómeno del cambio climático es un problema complejo es necesario descomponer en factores, el presente trabajo afronta la comprensión de las emisiones de CO_2 que es uno de los factores que repercuten de manera drástica en el cambio climático, mediante la construcción de un modelo de predicción de CO_2 basado en datos previamente ajustados al método científico para funcionar como una herramientas de inteligencia de negocios.

El área de estudio más común dentro del cambio climático es sin duda el entendimiento del CO_2 debido a que es el gas de mayor injerencia en los cambios ambientales, además está estrictamente ligado a la desproporcionada actividad humana[2]. Existen diversos trabajos relacionados con la elaboración de modelos de predicción para el CO_2([41], [4], [9]), en su mayoría consideran series de tiempo, y trabajan con algoritmos como ARIMA, el trabajo de investigación plantea la realización de un modelo de predicción construido a partir de árboles de decisión, utilizando datos recopilados alrededor del mundo, puesto que la mayoría de los existentes solamente se enfoca en las emisiones de un país, siendo que el cambio climático es un problema global, las emisiones de CO_2 generadas en un país o estado no hacen exclusivos los problemas ambientales en dicha zona geográfica, y la obtención de un algoritmo de predicción hace posible comprender el fenómeno y así tomar medidas a corto o largo plazo para la mitigación y/o adaptación ante tal problemática.

1.2. Justificación

En la era de la Predicción Numérica Climática (Numerical Weather Prediction, NWP) de archivos de datos, la comunidad científica interesada en el problema del cambio climático expone:

"La necesidad de información sobre el cambio climático nunca había sido tan grande. Los urbanistas, empresarios e investigadores demandan tal información climática adaptada a su propósito. Al mismo tiempo el crecimiento de los usuarios de datos y el aumento exponencial en el volumen de datos que se ha producido está limitando el acceso a dicha información[38]".

La necesidad de recursos computacionales y servicios orientados al usuario nunca ha sido más evidente. El problema se expone ante la sociedad para comenzar a desarrollar los estudios entre ciencias interdisciplinarias que vinculen datos a través de bases de datos masivas para proporcionar productos que generen conocimientos y que no sólo resuelvan el acceso a los datos en bruto.

1.3. Contribución

Se abordará un modelo de solución de inteligencia de negocios, que aproveche las bondades de la tecnología de cómputo para resolver un problema de administración del conocimiento. Las aplicaciones de la inteligencia de negocios tienden a:

"Crear sistemas especializados en una función específica de la empresa, que contribuya a ser eficiente el diagnóstico de una situación y tomar la decisión adecuada para su solución; mediante la sistematización del manejo de datos, refinamiento de la información, representación del conocimiento[18]".

Es necesario crear un modelo de solución de inteligencia de negocios para permitir la implementación y desarrollo de la inteligencia de negocios aplicada a los datos de origen atmosférico obtenidos de los centros de monitoreo pertenecientes al NOAA. Se propone una nueva estructura de almacenamiento de datos que contenga en ella todos los datos proporcionados por todos los centros de monitoreo existentes dentro del NOAA, que apoyan el acceso de datos, con el fin de aplicar tecnologías que hagan efectiva la toma de decisiones en torno a acciones que mejoren la calidad ambiental, de manera sucinta se hace especial enfoque en la realización de un modelo de predicción para la concentración de CO_2 basado en la estructura propuesta, que puntualmente va de la mano con la misión del NOAA: *"Entender y predecir los cambios en el ambiente de la tierrra[1]"* así como atender la problemática que da origen al trabajo de tesis.

En el capítulo 2 se da una descripción del estado del arte, se aborda de manera general el contexto del trabajo de investigación, ¿cómo surge la problemática?, la situación actual del problema, los recursos disponibles, así como algunos conceptos relacionados con el problema y que actúan como líneas de apoyo en el desarrollo del trabajo de investigación, también enmarca algunas particularidades presentes en la atmósfera, en general a aquellos componentes que se hacen presentes en ellas, en su mayoría gases, se hablará de manera sintetizada de algunos de los efectos de los gases en el ambiente así como su origen, para finalmente enunciar al CO_2 como el gas objeto de estudio central dentro del trabajo de tesis, enfatizando el por qué de su investigación. Los capítulos anteriores dan paso al capítulo 3, se detalla a groso modo la metodología para llevar a cabo el estudio centrándose en el proceso de generación del conocimiento, mismo que es descrito en cada paso, destacando los algoritmos usados y la teoría necesaria para la compresión de los mismos, con el fin de desarrollar de modelos que predigan la concentración de CO_2. El capítulo 4 está dedicado exclusivamente a mostrar la construcción de todos los modelos realizados haciendo hincapié en las bases utilizadas para su desarrollo así como la evaluación de los mismos, para después predecir las concentraciones de CO_2 y compararlas con las que se tienen disponibles y así determinar su efectividad, estas tareas conllevan a la predicción la concentración del CO_2 para finalizar con el capítulo

con una reseña general de la interpretación de las predicciones del gas estudiado desde el punto de vista ambiental, el capítulo final enuncia algunas conclusiones y trabajos a futuro.

2

Estado del arte

Durante este capítulo se explica la importancia del estudio de los datos así como la vía de obtención de los mismos para el desarrollo de la investigación.

2.1. El problema del cambio climático

Las actitudes entorno al cambio climático y el comportamiento que se debe asumir entorno a él son problemas que actualmente aquejan a la sociedad. Desde 1990, el inminente crecimiento del problema del medio ambiente ha generado una crisis sin precedentes [47]. En el cuarto reporte publicado en el IPCC (Intergovernmental Panel on Climate Change) publicado el 2 de febrero de 2007, claramente se indica que por 100 años (1906-2005) la convergencia lineal de la temperatura global es 0.74, además el informe también indica que la mayor parte del calentamiento observado en el sistema climático en los últimos 50 años es debido a las actividades humanas [34], se puede decir entonces que el cambio climático es un problema de orden mundial debido a su aumento aelerado, en especial por que:

« La influencia humana en el cambio climático es clara, y las recientes emisiones antropogénico de gases de efecto invernadero son las más altas de la historia. Los recientes cambios climáticos han extendido su impacto sobre humanos y sistemas naturales[2]. »

Actualmente se sabe que cada una de las últimas tres décadas han sido sucesivamente más cálidas en la superficie de la tierra que en alguna otra década precedente desde 1850. El periodo de 1983 a 2012 es más cálido en un periodo de 30 años de los últimos 1400 años en el Hemisferio Norte[2].
La problemática del cambio climático genera consecuencias adversas sobre todo el mundo, el aumento de la temperatura es quizá el más notable y desencadena una serie de efectos que preocupan a la sociedad, el problema va de la mano con las actividades humanas de tipo industrial, que se derivan en los altos niveles de concentraciones de compuestos químicos que hacen presa fácil a la sociedad de enfermedades respiratorias y cardiovasculares.

2.2. Grandes almacenes de datos

Las desproporcionadas actividades humanas han acelerado de manera preocupante el deterioro del medio ambiente. El concepto de conciencia ambiental empezó a sonar en 1960[20], es por eso que el cambio climático es un tema amplio de investigación que ha sido abordado desde diferentes ángulos por la diversidad de sus consecuencias sobre el mundo, una tarea clave en su estudio es el análisis de datos, mismos que son obtenidos a través del monitoreo continuo, con el fin de conocer su comportamiento, existen sistemas de monitoreo a nivel mundial para mejorar su alcance, respondiendo a esa necesidad existe la División global de Monitoreo ESRL que posee centros de monitoreo alrededor del mundo (figura 2.1) que son capaces de medir temperatura, $C0_2$, presión atmosférica, etcétera.
Puesto que el monitoreo debe ser continuo y constante los datos obtenidos son medidos durante las 24 horas de cada día, en un mes de 30 o 31 días, durante un año, los datos son albergados en grandes almacenes ya que su tamaño aumenta con el pasar de los años, los datos no pueden ser desechados por su importancia debido a que el análisis de este problema no brinda resultados fiables si sólo se analiza un pequeño grupo de datos, el tamaño aumenta si consideramos que cada estación de monitoreo mide un número considerable de elementos en el ambiente, que no es constante para los restantes centros

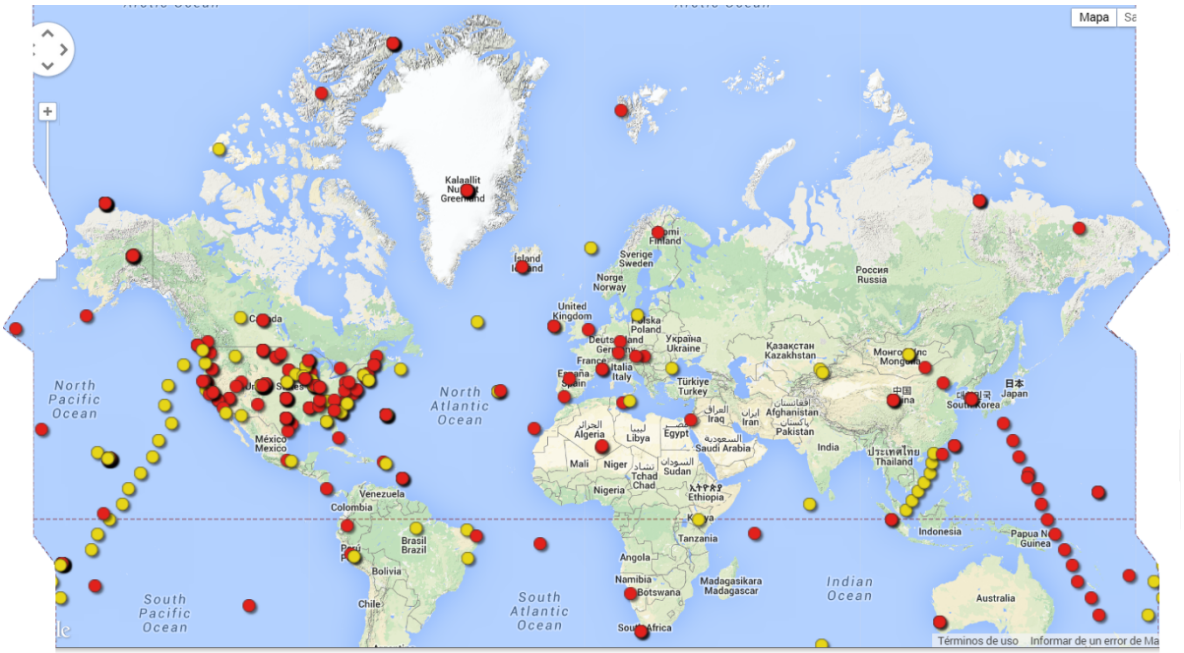


Figura 2.1: Centros de monitoreo pertenecientes la división global de monitoreo ESRL.

de monitoreo, existen 215 sitios que funcionan como centros de monitoreo repartidos en 51 ciudades, todos esos datos provenientes de los centros constituyen bases de datos de tamaño considerable, en total 3106 bases de datos, se habla del orden de Petabytes de datos en total.

De acuerdo al aumento del cambio climático se ha despertado el interés de varias áreas de estudio con el afán de contribuir en la elaboración de opciones de adaptación y mitigación para poder direccionar el cambio climático, tal efecto desata una serie de problemas a numerosos investigadores, con una gran constante para un Matemático, Químico, Físico, etcétera, cada uno debe enfrentarse al reto de la manipulación de los datos independientemente de su hipótesis, el problema con exactitud es que cada uno de ellos debe aprender a manejar herramientas para el posterior análisis de los datos que involucra conocer la sintaxis de un lenguaje, buscar, aprender, modificar y codificar algoritmos, además probar que algoritmo funciona mejor de acuerdo a sus necesidades, todas las actividades enlistadas anteriormente y otras requieren de un alto nivel en el manejo de herramientas de software que solo se obtiene mediante la experiencia y que

solo dificultan la investigación.
El problema ya no está en la obtención de datos, ya que para el caso de estudio antes citado es público, si no que da un giro importante, los datos son conjuntos discretos de factores objetivos sobre un hecho real y no tienen la capacidad de decir el porqué de los fenómenos, y que por sí mismo tiene poca o ninguna relevancia o propósito.

2.3. Dato, información y conocimiento

A diferencia de los datos, la información tiene significado (relevancia y propósito). La información es algo más apelativa ya que requiere la aceptación del receptor y por lo general el emisor ha intentado estructurar ese conjunto de datos para que impacte, para que verdaderamente informe. Pero la información todavía no es suficiente para concretar la acción, ni siquiera la indica, la información describe un estado por ello es efímera, esa información es válida mientras ese estado no varíe.
El conocimiento es un conjunto más complejo que consta de datos, información, conceptos y experiencias. El conocimiento se logra al interpretar y analizar un conjunto de informaciones referentes a un hecho puntual para decidir cómo actuar sobre él.
Una aproximación que reuniría a las tres podría ser la siguiente: los datos están localizados en el mundo y el conocimiento está localizado en agentes (personas, organizaciones,...), mientras que la información adopta un papel mediador entre ambos conceptos. Retomando los grandes almacenes de datos de datos en la atmósfera debemos de entender que se tiene que construir una manera de hacer la transición de datos a conocimiento de interés para un investigador.

2.4. Particularidades en la atmósfera

Nos adentraremos al trabajo de tesis haciendo uso de conceptos relacionados dentro del ámbito del cambio climático, se incluye la situación actual de la atmósfera, misma que involucra forzosamente estudiar los componentes presentes dentro de ella y de los que se hace mención, ya que en altas cantidades son capaces de desencadenar diversos fenómenos en el ambiente, para después centrarse en el gas que será estudiado indicando

la razón de su análisis.
Es necesario llevar a cabo un análisis de los componentes encontrados en la atmósfera y que son medidos con sistemas de monitoreo atmosférico, para poder acotar el dominio de desarrollo se optó por el estudio del $C0_2$. A continuación se presentan algunas particularidades de la presencia de dichos componentes y la justificación en la elección del CO_2.
Con el actual y creciente problema del cambio climático es necesario tomar en cuenta las condiciones ambientales presentes y que pueden ser obtenidas mediante el monitoreo, a partir de ellas es realizado un análisis para un trabajo de investigación. Algunos parámetros a los que podemos tener acceso en el repositorio del NOAA son:

- Monóxido de carbono
- Dióxido de carbono
- Metano
- Etano
- Propano
- Aerosoles
- Ozono
- Oxido nitroso
- Vapor de agua
- Sulfuro de carbonilo
- ...

La presencia de estos compuestos contribuye de manera importante al cambio climático, y son parte fundamental en su estudio para lograr generar políticas que ayuden en la mitigación del problema. A continuación se hará una descripción de algunos componentes.

2.4.1. GEI

La vida en la Tierra depende de la energía que recibe del sol, cerca de la mitad de la luz que llega a la atmósfera terrestre pasa a través del aire y las nubes para llegar a la superficie donde se absorbe y luego es irradiado nuevamente en forma de calor (ondas infrarojas). De este calor el 90 % es absorbido por los gases de efecto invernadero y devuelta hacia la superficie que la ayuda a calentar hasta una temperatura promedio de 15 grados Celcius perfecto para la vida, es conocido como el efecto invernadero. Los GEI principales son:

- El vapor de agua, el más abundante y funciona como un gas que actúa en retroalimentación con el clima, a mayor temperatura de la atmósfera, más vapor, más nubes y más precipitaciones.
- Dióxido de carbono (CO_2), un componente menor, pero muy importante de la atmósfera. Se libera en procesos naturales como la respiración y en erupciones volcánicas y a través de actividades humanas como la deforestación, cambio en el uso de suelos y la quema de combustibles fósiles. Desde el inicio de la Revolución Industrial (aproximadamente 1760) la concentración de CO_2 ha aumentado en un 43 % (para el 2013).
- Metano, un gas hidrocarburo que tiene origen natural y resultado de actividades humanas, que incluyen la descomposición de rellenos sanitarios, la agricultura (en especial el cultivo de arroz), la digestión de rumiantes y el manejo de desechos de ganado y animales de producción. Es un gas más activo que el dióxido de carbono, aunque menos abundante.
- Óxido nitroso, gas invernadero muy poderoso que se produce principalmente a través del uso de fertilizantes comerciales y orgánicos, la quema de combustibles fósiles, la producción de ácido nítrico y la quema de biomasa.
- Los clorofluorocarbones (CFCs), son compuestos sintéticos de origen industrial que fueron utilizados en varias aplicaciones, ahora ampliamente regulados en su producción y liberación a la atmósfera para evitar la destrucción de la capa de ozono.

Los GEI constituyen un elemento esencial para la vida: sin ellos, el planeta sería un bloque de hielo. Si en un invernadero la cobertura plástica evita la pérdida del calor y conserva una temperatura estable, en la Tierra estos gases consiguen un efecto similar. Su presencia en la atmósfera permite beneficiarse de parte del calor que envía el Sol. De ahí su nombre. Los principales GEI son de origen natural. El problema surge cuando la cantidad de estos gases aumenta porque se altera el equilibrio natural y el clima se comporta de manera distinta. La industrialización, con el uso masivo de combustibles fósiles (petróleo, carbón y gas) y todas las actividades humanas derivadas, como el transporte o el uso intensivo de la agricultura y la ganadería, contribuyen desde el siglo XIX a incrementar estos gases. El aumento de los GEI se asocia también a otros problemas antropogénicos (causados por el ser humano) para el medio ambiente. La deforestación ha limitado la capacidad regenerativa de la atmósfera para eliminar el dióxido de carbono (CO_2), uno de los principales GEI. La gran mayoría de la comunidad científica internacional está de acuerdo en la importancia de reducir la emisión de estos gases[2]. Para ello, se proponen diversas medidas: sustituir los combustibles fósiles por energías renovables, asumir de forma plena un mercado de emisiones de GEI, aplicar medidas de eficiencia energética, aumentar la reforestación y, en definitiva, introducir en la sociedad prácticas de desarrollo sostenible en todas las actividades.

Dióxido de carbono

El CO_2 es uno de los gases de efecto invernadero más importantes que causan el calentamiento global y cambio climático. La emisión anual ha crecido entre 1970 y 2004 cerca de un 80 % [24], es un gas incoloro, denso y poco reactivo, que forma parte de nuestra atmósfera; se genera a través de la respiración animal y vegetal para los procesos naturales de fotosíntesis. La actividad humana también produce CO_2 mediante el sector industrial. Dado que en el último siglo ha aumentado significativamente su concentración en la atmósfera, el CO_2 es reconocido por la ONU como uno de los principales causantes del Calentamiento Global, junto con otros cinco gases[2].

Monóxido de carbono

El monóxido de carbono (CO) es un gas incoloro e inodoro que en concentraciones elevadas puede ser letal. Se forma en la naturaleza mediante la oxidación del metano (CH_4), gas común producido por la descomposición de la materia orgánica. La fuente antropogénica del CO es la quema incompleta de combustibles (gasolina, gas, carbón, madera y combustóleo). En este sentido, para tener menos emisiones de CO es necesario tener procesos de combustión más completos, lo que requiere de una cantidad adecuada de oxígeno; cuando éste es insuficiente, se forma del CO. Los vehículos son la principal fuente de CO en áreas urbanas, por lo que en la actualidad los automóviles nuevos cuentan con un convertidor catalítico que permite reducir las emisiones de CO a la atmósfera, así como de otros gases contaminantes [46]. El CO tiene una acción tóxica en el ser humano por su afinidad con la hemoglobina, a la que se enlaza disminuyendo el transporte del oxígeno en el cuerpo. Entre los efectos en la salud que se asocian con la ex-posición al CO se encuentran una menor coordinación motora, agravamiento de enfermedades cardiovasculares, fatiga, dolores de cabeza, confusión, náuseas y mareos. En casos extremos, con exposiciones muy elevadas que superan por mucho los niveles que se presentan en la atmósfera de zonas urbanas, puede causar la muerte [33].

Ozono

El ozono (O_3) es un gas que se forma en la atmósfera cuando se combinan 3 átomos de oxígeno; se caracteriza además por ser incoloro, con un olor particular, y por poseer la capacidad de oxidar materiales. El O_3 es muy reactivo e incluso a bajas concentraciones es irritante y tóxico. El O_3 se encuentra de manera natural en la estratósfera; sin embargo, en la tropósfera es un contaminante secundario que se forma mediante la reacción química de óxidos de nitrógeno (NO_X) y compuestos orgánicos volátiles (COV) en presencia de la luz solar. La exposición al O_3 puede causar daños en los humanos, tanto sanos como enfermos, siendo los grupos más vulnerables los niños y las personas de edad avanzada; es capaz de ocasionar inflamación pulmonar, depresión del sistema inmunológico, cambios agudos en la función, estructura y metabolismo pulmonar, que pueden resultar en síntomas y enfermedades respiratorios [12], e incluso existe evidencia

sobre una asociación con mortalidad prematura [6].

Bióxido de azufre

Los gases de la familia de los óxidos de azufre (SO_X), entre los que se encuentra el bióxido de azufre (SO_2), son incoloros y de olor irritante; se forman al quemar combustibles con azufre, y tienden a disolverse fácilmente en agua. La fuente primaria de emisiones de SO_2 es la quema de combustibles fósiles que contienen azufre, tales como combustóleo, diesel y carbón. Las fuentes naturales de SO_2 incluyen erupciones volcánicas, decaimiento biológico e incendios forestales. El SO_2 es, además, precursor de otros contaminantes, como el trióxido de azufre (SO_3), el ácido sulfórico ($H_2SO_4SO_2$) y los sulfatos, que contribuyen a la formación de partículas finas en la atmósfera y de la lluvia ácida. La exposición al SO_2 se ha asociado con daños respiratorios temporales en niños y adultos asmáticos que realizan actividades al aire libre. Se ha observado que la exposición aguda de los individuos asmáticos a magnitudes elevadas de SO_2 al realizar ejercicio moderado, puede causar reducción de la función pulmonar, estado que se puede acompañar de síntomas como estornudos, opresión en el pecho y falta de aire. Por otra parte, los efectos que se han asociado con exposiciones crónicas aunadas a concentraciones elevadas de partículas en el ambiente, incluyen enfermedades respiratorias, alteraciones en las defensas pulmonares y agravación de enfermedades cardiovasculares preexistentes. En la Ciudad de México se encontró que en niños menores de 16 años una exposición a 50 ppb de SO_2 está asociada con un incremento del 5 % en el nómero de visitas a salas de emergencias por sintomatología de asma [36].

Bióxido de nitrógeno

El bióxido de nitrógeno (NO_2) es un gas que pertenece a los óxidos de nitrógeno (NO_X), término genérico comónmente empleado para referirse a un grupo de gases sumamente reactivos que contienen diferentes cantidades de oxígeno y nitrógeno, como el óxido nítrico (NO) y el NO_2 . El NO_2 es un gas incoloro, inodoro y muy corrosivo, precursor del ozono troposférico y de partículas suspendidas, tales como nitratos. El NO_2 se forma cuando se quema combustible a altas temperaturas y dicho combustible contiene compuestos nitrogenados, como la gasolina o el keroseno. Las principales fuentes antro-

pogénicas de NO_2 son los vehículos automotores, plantas de generación de electricidad y otras fuentes industriales, comerciales y residenciales que queman combustibles. El NO_2 puede formarse también de manera natural por la descomposición bacteriana de nitratos [17]. La exposición al NO_2 puede causar irritación pulmonar, problemas de percepción olfativa, molestias respiratorias, dolores agudos en vías respiratorias y edema pulmonar; además, la exposición prolongada a este gas se asocia con daños a la percepción sensorial, agravamiento de los síntomas en asmáticos y de los síntomas relacionados con enfermedades pulmonares crónicas. Los resultados de los estudios llevados a cabo en la ZMVM son congruentes con la literatura internacional, ya que encontraron una asociación entre un incremento del nómero de visitas a hospitales por síntomas de infecciones respiratorias agudas, y la exposición al NO_2 [42].

2.5. La importancia del estudio del dióxido de carbono

Aunque la variedad de componentes que podemos medir mediante los redes de monitoreo atmosférico son limitadas, su selección está fuertemente ligada a los estragos en el ambiente de manera directa o indirecta. Para el presente trabajo de tesis se aborda el estudio del dióxido de carbono, ya que forma parte de los GEI. Su concentración en la atmósfera se ha mantenido constante desde el final del Precámbrico hasta la Revolución Industrial[28], pero debido al crecimiento desmesurado de la combustión de combustibles fósiles la concentración está aumentando, incrementando el calentamiento global[2]. La decisión de estudiar uno de los GEI radica en que contribuyen al calentamiento global y es un problema que encausa la desertificación, cambia los patrones y el aumento de las precipitaciones, aumento en el nivel del mar, entre otros que no sólo afectan las condiciones climatológicas, si no que repercute en la flora y fauna silvestre. La presencia de estos compuestos contribuye de manera importante al cambio climático, y son parte fundamental en su estudio para lograr generar políticas que ayuden en la mitigación del problema[2].

La preocupación por el estado del ambiente despierta interés en la sociedad científica que busca primordialmente conocer el comportamiento de los fenómenos que puedan

repercutir en la atmósfera, de manera tal que pueden explicarse la existencia de organizaciones como el NOAA, que posee gran injerencia a nivel mundial en materia ambiental y que en su preocupación por los efectos sobre el ambiente comparte de manera indistinta las concentraciones de los gases monitoreados por sus diferentes centros atmosféricos, en búsqueda siempre de una sociedad informada, en el capítulo siguiente se detalla de manera explícita como son trabajados los datos proporcionados por el NOAA en el desarrollo de una herramienta que facilite la toma de decisiones en búsqueda de reducir las emisiones de CO_2.

3

Inteligencia de negocios aplicada en los datos del NOAA

El actual estado de la atmósfera incita creciente preocupación dentro de la comunidad, las organizaciones comprometidas con el medio ambiente como el NOAA hacen posible tratar de comprender en números los fenómenos presentes en la atmósfera. En el capítulo 1 se expone la necesidad actual de crear modelos de predicción climática considerando el estado de los datos, en específico, la estructura existente aumenta cada año y además es primitiva, lo que retarda el acceso de los datos en la elaboración de trabajos de investigación, que motivan el análisis del estado presente y futuro de la atmósfera. El problema es propuesto por científicos pertenecientes al NOAA, mismo que va de la mano con la misión de su organización, enfocada en el monitoreo, análisis y cuidado del medio ambiente[1]. Dos puntos importantes dentro de la misión del NOAA consiste en:

- Entender y predecir cambios en el clima, el agua, los océanos y las costas.
- Compartir y entender conocimiento con la sociedad.

El NOAA persigue una sociedad además de informada que aproveche el entendimiento de cada factor participante en las costas, los océanos y la atmósfera[1], con el fin de tomar las mejores decisiones sociales y económicas que hagan posible la conservación del ambiente, es posible abordar el tema mediante herramientas de inteligencia de negocios, ya que hace posible soportar la toma de decisiones basadas en información precisa y oportuna, que garantiza la generación de conocimiento necesario para para tomar la mejor alternativa para el éxito en la organización [18].

Asociaciones como el NOAA plantean el uso de datos para ser convertidos en información de carácter climático para poder ser explotada. El problema planteado puede abordarse mediante BI(Business Intelligence, inteligencia de negocios), BI es un amplio conjunto de habilidades, procesos, tecnologías, aplicaciones y prácticas que apoyan a los miembros de una organización para tomar mejores decisiones y más eficientes [25], se transforman datos de los sistemas transaccionales e información "desestructurada", interna y externa a la organización, en información estructurada, para hacer posible la explotación directa o para su análisis y conversión en conocimiento, mismo que los miembros de la organización pueden utilizar para tomar mejores decisiones [22]. Los principales componentes de orígenes de datos en la inteligencia de negocios que existen en la actualidad son datamart(DM) y data warehouse(DW), de las que se hablará con mayor detalle en la sección siguiente.

3.1. Data warehouse

Una vez que se han identificado los datos que se tienen, es necesario crear una versión única, consolidada, estandarizada y confiable en una base de datos especial, conocida como data warehouse. En ésta se conservarán los datos históricos bajo una estructura optimizada para búsqueda, no para actualizaciones y que sea administrable, escalable y segura. Data warehouse es una colección centralizada de información corporativa, histórica y transformada, proveniente de sistemas transaccionales heterogéneos y externos para atender requerimientos que apoyen tecnológicamente en el proceso de toma de decisiones gerenciales[35]. Las técnicas de minería de datos y OLAP se aplican ya sea sobre un data warehouse o sobre un repositorio orientado a un subconjunto específico (data mart).

La técnica OLAP, llamada de análisis multidimensional y contrapuesta la de procesamiento transaccional en línea (OLTP) actúa sobre un DW organizando en "hipercubos.º cubos multidimensionales sobre los llamados elementos de análisis o "hechos"(por ejem-

plo, número de elementos defectuosos, máximo número de ventas, promedio de inasistencias), y también bajo ciertas dimensiones(por ejemplo, productos, centro de costo, maquina, año), OLAP trata cuestionamientos tales como ¿A cuánto ascienden las ventas netas (hechos) por producto, por plaza, por mes(dimensiones)?.

La presentación de datos analizados con OLAP se denomina visualización y utiliza herramientas y formas de representación gráfica como: tridimensionalidad, histogramas, regresiones, etcétera. La visualización permite efectuar análisis de tendencias, puntos de equilibrio y sensibilidad. Es posible usar tres tipos de Olap, cuya variación radica en la implementación del DW, a saber:

- Rolap. OLAP sobre DWs implementados como tablas relacionales (filas y columnas) usando modelos estrella y copo de nieve.

- Molap. OLAP sobre DWs implementados sobre bases de datos multidimensionales.

- Holap. Enfoque híbrido de ROLAP y MOLAP.

De manera complementaria o independiente al OLAP y otro tipo de requerimientos y consultas complejas de información se puede aplicar el DM, que es el proceso de encontrar patrones y correlaciones nuevas, ocultas o inesperadas. Sus principales herramientas son la Inteligencia Artificial (IA) (sistemas expertos, bases de conocimiento, redes neuronales, algoritmos genéticos, lógica difusa) y la Teoría de Decisiones. Son cuatro las técnicas que permiten descubrir y analizar tales patrones y correlaciones: clasificación (reglas Si-Entonces), asociación (correlación), secuencia (series de tiempo) y sectorización (división). La arquitectura de BI puede verse en 3.1, cuyo comportamiento se ha descrito durante la presente sección.

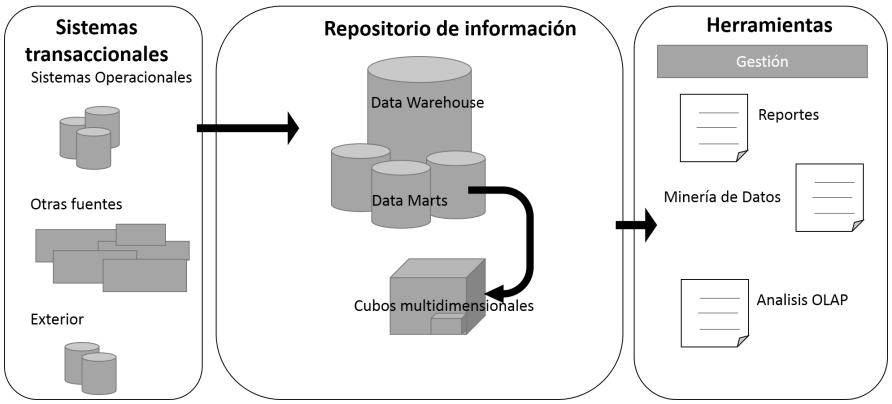


Figura 3.1: Arquitectura del proceso de inteligencia de negocios.

3.2. ¿ Por qué utilizar BI?

La cantidad de datos que alberga el sitio del ESRL es enorme, cada año crece en todos sus sitios activos, es necesario cambiar el enfoque de trabajo centrándonos en primera instancia en la elección del software en el que se hará el desarrollo, es importante recalcar que existen otros programas de software que pueden llevar a cabo técnicas de minería de datos, el ejemplo es $Matlab$, en ella se puede hacer uso de librerías de máquinas de soporte vectorial, algoritmos de clustering, redes neuronales e incluso árboles de decisión. La diferencia entre utilizar $Matlab$ y $SQL - Server$ salta a primera vista, es decir, las ventajas que se obtienen al utilizar Matlab son:

- Fácil carga de datos.
- Fácil interpretación y ejecución de los algoritmos disponibles en la Minería de Datos.
- Los datos pueden estar almacenados en una estructura sencilla y funcional.
- Se usa un solo lenguaje.

Sin embargo las desventajas de utilizar $SQL - SERVER$.

- Basa su funcionamiento en la creación de estructuras dinámicas.

- Es necesario tener almacenada la información con la que se trabajara de una manera bien estructurada, la mencionada estructura puede construirse con MDX a partir de la existencia de una Base de Datos bien estructurada que modele la operación de la organización, incluyendo información que no es de interés y que por lo tanto no cobra relevancia como alguna variable en el modelo a construir.

- En la creación de un modelo de minería de datos es necesesario el uso de MDX.

- Las consultas de predicción se realizan utilizando DMX.

Hasta este punto es previsible que el uso de Matlab supera por la facilidad como una herramienta de trabajo, sin embargo es necesario considerar que en el proceso de grandes cantidades de datos, tal como el trabajo de investigación, es requerido un rápido manejo de datos que sólo se puede conseguir con SQL-Server, cuya principal ventaja es el manejo de grandes almacenes de datos, además de permitir la construcción nuevas estructuras de datos denominadas cubos, que son vistas de la base de datos, que hace más fácil su carga y procesamiento con el objetivo de realizar minería de datos.
Para llevar a cabo BI es necesario recopilar datos de las fuentes que sean necesarias para organizarlas y después almacenarlas en repositorios que permiten generar estructuras dinámicas que a su vez permiten el proceso analítico en línea para generar conocimiento que funciona como insumo para las herramientas de uso táctico y estratégico.
Desde hace muchos años las aplicaciones de BI están trabajando con éxito en la gestión de seguros, comercio minorista, telecomunicaciones, aviación, relaciones con clientes, gestión de desastres, la banca [26], así como en otros dominios. Es necesario entender como "Conocimiento"dentro de BI al proceso de identificar patrones significativos de los datos que sean válidos, novedosos, potencialmente útiles y comprensibles para el personal involucrado en la toma de decisiones y que impacta en la organización[29], los datos de bajo nivel son transformados de forma iterativa e incremental.

El proceso de extracción de Conocimiento se muestra en la figura 3.2, los pasos son los siguientes:

- Comprensión del dominio de la aplicación.

- Selección de datos.
- Preprocesar los datos.
- Transformación de datos.
- Minería de datos.
- Interpretación de datos.
- Interpretación y evaluación de la información.

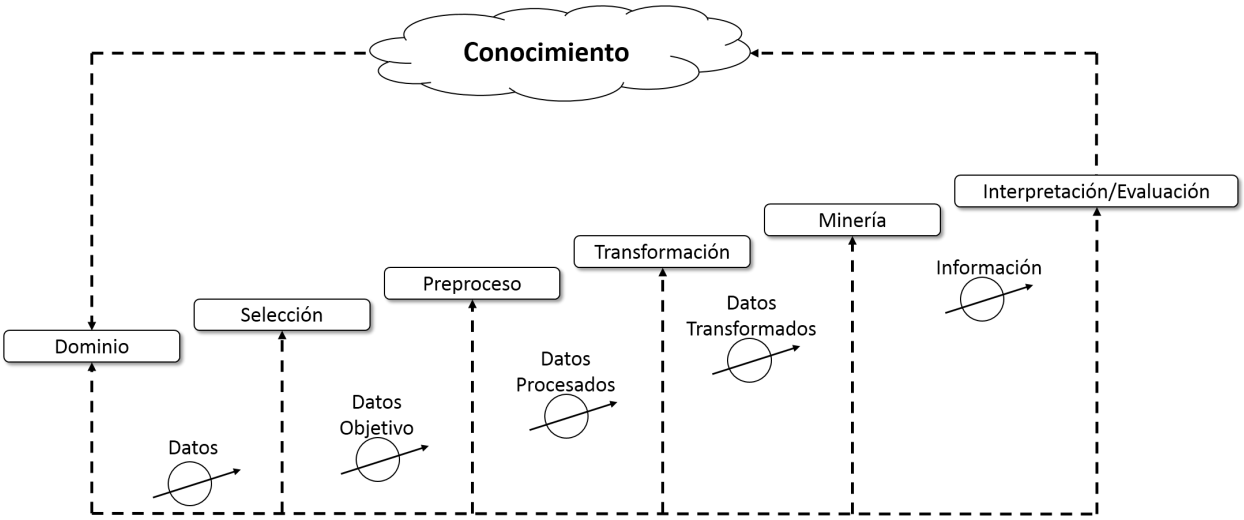


Figura 3.2: Proceso de generación del conocimiento.

Es necesario establecer el dominio de la aplicación, seleccionar los datos para después aplicar un preproceso, posteriormente se transforman para aplicar minería de datos, para así encontrar patrones y finalmente interpretar los datos, es observable que el proceso de generación del conocimiento es un proceso iterativo e incremental, por lo que es posible conocer más acerca del caso de estudio.

3.3. Dominio de la aplicación

En capítulos anteriores se ha hecho énfasis en el caso de estudio, se han mencionado las redes de monitoreo del ESRL pertenecientes al NOAA (National Oceanic and Atmospheric Administration) , además posee una gran cantidad de datos que son gratis y no son exclusivos para la comunidad científica, se espera que los datos proporcionados por los diferentes centros de monitoreo tengan un impacto de tipo en la misión del NOAA, se hace especial énfasis en el objetivo parcial del objetivo de investigación, que consiste en modelar una nueva estructura de datos que permita no solamente tener un nuevo almacén de datos, sino que sirva como base en la obtención de modelos dinámicos que faciliten el uso de los mismos para ser usados y así poder construir modelos de predicción para los componentes monitoreados por el ESRL.

3.4. Selección de datos

De acuerdo al análisis y la exploración de los datos proporcionados por el ESRL y disponibles en su sitio Web[1] (ver figura3.3) es importante tener las siguientes consideraciones:

- Algunas de las concentraciones de componentes monitoreados se reportan en unidades de partes por millón (ppm) en su mayoría los componentes que tienen este tipo de unidad de medida corresponden a dióxido de carbono, monóxido de carbono, ozono, etano, i-butano, entre otros gases, cabe destacar que la unidad de medida cambia según el parámetro medido, por ejemplo la radiación esta reportada en Watts/m^2.

- Algunos componentes del ambiente reportado por el ESRL por su naturaleza reportan más de una variable importante, es el caso del vapor de agua que reporta la humedad relativa basada en una unidad propuesta por el NOAA, la humedad relativa basada en medidas de humedad de radiosonda, entre otras. Otro caso similar proporciona 6 diferentes magnitudes: radiación solar directa, irradiancia

[1] http://www.esrl.noaa.gov/gmd/dv/data/

solar difusa, hundimiento total de la radiación solar, irradiancia de onda larga, irradiancia solar al alza, irradiancia de onda larga.

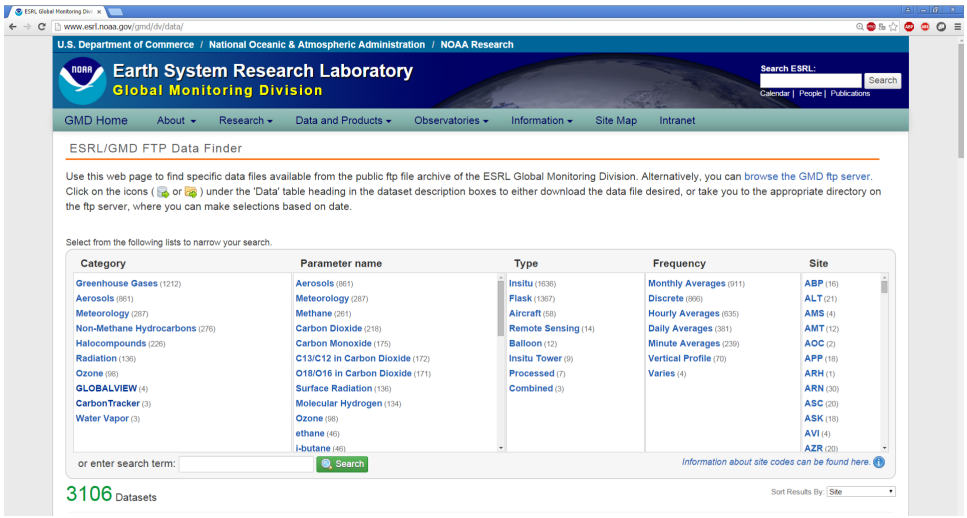


Figura 3.3: Sitio Web del ESRL.

- Los componentes están clasificados por categorías de manera tal que no son disjuntos, es decir, un componente puede estar en más de una categoría, las clases son:
  - Gases de efecto invernadero.
  - Aerosoles.
  - Meteorológicos.
  - Radiación.
  - Ozono.
  - Vapor de agua.
  - Vista global.
  - Compuestos alogenados.
  - Hidrocarburos No-Metano.
  - Carbon Tracker.

- Hay datos disponibles para los siguientes 25 componentes:

Aerosoles	Clorofluorocarbono-12
Datos meteorológicos	Metilcloroformo
Metano	Tetracloruro de Carbono
Dióxido de carbono	Óxido Nitroso
Monóxido de carbono	Clorofluorocarbonos-113
C13/C12 in Dióxido de carbono	Halón -1211
O18/O16 in Dióxido de carbono	Hexafluoruro de azufre
Radiación	Hidrofluorocarburos -142b
Hidrógeno molecular	Hidrofluorocarburos -22
Ozono	Cloruro de metilo
Etano	Radiocarbono
i-Butano	Vapor de agua
i-Pentano	Sulfuro de carbonilo
n-Butano	Hidrofluorocarburos-141b
n-Pentano	Hidrofluorocarburos-141b
Propano	Bromuro de metilo
C13/C12 in Metano	Bromo, Cloro
Clorofluorocarbono-11	

- Los sitios no reportan la concentración de los mismos componentes, por ejemplo, el sitio que esta en México solo reporta dióxido de carbono, monóxido carbono, metano, etano, i-butano, i-pentano, n-butano, n-pentano, propano, mientras que el sitio ubicado en Arizona solo reporta las concentraciones de metano.

- Cada componente puede medirse con diferentes instrumentos (insitu, frascos, sensado remoto, aeronave, insitu tower,globo, combinado, procesado), en los reportes de las concentraciones el formato es diferente(Diferentes columnas) dependiendo del tipo de instrumento empleado.

- Cada componente puede medirse en diferentes frecuencias, el reporte de las con-

centraciones varía para cada tipo de frecuencia, a continuación se enlistan las frecuencias disponibles:

  - Promedio mensual.
  - Promedio por hora.
  - Promedio diario.
  - Promedio por minuto.
  - Discreto.
  - Varios.
  - Perfil vertical.

- El ESRL contiene un total de 215 sitios que se encargan del monitoreo, de los cuales 69 se encuentran descontinuados, dentro de cada sitio es posible que este asignado uno o más proyectos de investigación que igualmente pueden estar habilitados o descontinuados. Cada sitio tiene un nombre, un código que lo identifica, una ciudad, así como su latitud, longitud y elevación, su tiempo en GTM y el proyecto asociado a él.

Las observaciones antes mencionadas son necesarias ya que de ello depende la realización del nuevo modelo para almacenar los datos.

3.5. Preproceso de datos

Realizado el análisis de datos y la estructura en la que se encuentran almacenados surgen las observaciones siguientes:

- No existen campos sin valores.
- No existen espacios vacíos, sin embargo existen algunos valores de concentraciones y/o incertidumbre cuyo registro contiene el valor de -0.999, según la descripción de los metadatos y por convención en todos los sitios de monitoreo corresponde a una ausencia de valor.

Aunque los datos que proporciona el ESRL son muy variados es importante mencionar que para el presente caso de estudio se acordó analizar los datos correspondientes al dióxido de carbono, sin embargo el modelado de la estructura de datos afectara el almacenamiento de todos los componentes que son monitoreados por el ESRL ya que todas se almacenan en una misma estructura, la concentración del dióxido de carbono se mide el partes por millón, puede además ser medido por varios instrumentos y pueden ser reportados en diferentes lapsos de tiempo en diferentes centros de Monitoreo, cada reporte puede descargarse en el sitio del ESRL que esta en texto plano. El preproceso es una tarea exhaustiva debido a la cantidad de datos, sin embargo debe señalarse que el sitio del ESRL provee de los datos necesarios así como los metadatos para cada tipo de reporte proporcionado por los sitios de monitoreo.
Se enlistan las siguientes cualidades que tomaran las concentraciones de dióxido de Carbono de acuerdo al enfoque del trabajo de investigación:

- Se trabajara con las concentraciones del periodo 2009-2013, ya que es el único periodo reportado por México en la concentración de CO_2.

- Los datos se obtendrán de 91 de los sitios disponibles, ya que son todos los sitios que cumplen con el punto anterior.

- La frecuencia a analizar corresponde al promedio mensual de CO_2, debido a que es el único reportado por México.

- El instrumento a considerar serán Frascos(Flask) puesto que es el único instrumento utilizado en el sitio Mexicano.

Las cualidades de los datos fueron seleccionadas para que los resultados del trabajo con el monitoreo sean equiparables.

3.6. Transformación

De acuerdo a los datos proporcionados por el ESRL que corresponden a documentos en texto plano(ver un ejemplo en la figura3.4) así como su sitio Web, fue necesario construir primero una estructura eficiente que los almacene y que así mismo fuera una base para realizar BI. De acuerdo al análisis señalado la estructura propuesta puede ser vista en la figura3.5.

```
comment:
comment:  ************ DATA DOCUMENTATION ***************
comment:
comment: Please refer to the species-specific README file in the
comment: appropriate directory folder at ftp://aftp.cmdl.noaa.gov/data/trace_gases/.
comment:
data_fields: sample_site_code sample_year sample_month sample_day sample_hour sample_minute sample_seconds sample_id sample_method parameter_formula analysis_group_abbr analysis_value analysis_uncertainty analysis_flag
nalysis_instrument analysis_year analysis_month analysis_day analysis_hour analysis_minute analysis_seconds sample_latitude sample_longitude sample_altitude event_number
X 2009 01 09 18 21 00  1307-99 P       co     CCGG   100.990     0.856 ...  R5 2009 01 30 11 22 00  18.9841  -97.3110  4469.00   268330
X 2009 01 09 18 21 00  1308-99 P       co     CCGG   100.730     0.855 ...  R5 2009 01 30 11 36 00  18.9841  -97.3110  4469.00   268331
X 2009 01 19 17 50 00  3543-99 P       co     CCGG    96.280  -999.990 P..  R5 2009 02 05 16 48 00  18.9841  -97.3110  4469.00   268778
X 2009 01 19 17 50 00  3544-99 P       co     CCGG   103.120  -999.990 P..  R5 2009 02 05 17 03 00  18.9841  -97.3110  4469.00   268779
X 2009 01 23 16 32 00   661-99 P       co     CCGG   123.150     0.908 ...  R5 2009 02 05 16 20 00  18.9841  -97.3110  4469.00   268776
X 2009 01 23 16 32 00   662-99 P       co     CCGG   123.150     0.908 ...  R5 2009 02 05 16 34 00  18.9841  -97.3110  4469.00   268777
X 2009 01 29 18 00 00   479-99 P       co     CCGG   182.720     1.406 ...  V2 2009 03 13 17 22 00  18.9841  -97.3110  4469.00   271298
X 2009 01 29 18 00 00   480-99 P       co     CCGG   184.450     1.411 ...  V2 2009 03 13 17 37 00  18.9841  -97.3110  4469.00   271299
X 2009 02 23 17 11 00  1165-99 P       co     CCGG   144.670     0.977 ...  V2 2009 03 13 16 20 00  18.9841  -97.3110  4469.00   271296
X 2009 02 23 17 11 00  1166-99 P       co     CCGG   144.580     0.977 ...  V2 2009 03 13 16 36 00  18.9841  -97.3110  4469.00   271297
X 2009 03 02 16 04 00  6607-66 P       co     CCGG   130.120     1.748 ...  R5 2009 05 18 14 28 00  18.9841  -97.3110  4469.00   275396
X 2009 03 02 16 04 00  6608-66 P       co     CCGG   133.080     1.753 ...  R5 2009 05 18 14 42 00  18.9841  -97.3110  4469.00   275397
X 2009 03 03 15 00 00  1727-99 P       co     CCGG    76.480     0.831 ...  V2 2009 03 13 16 51 00  18.9841  -97.3110  4469.00   271294
X 2009 03 03 15 00 00  1728-99 P       co     CCGG    76.740     0.832 ...  V2 2009 03 13 17 06 00  18.9841  -97.3110  4469.00   271295
X 2009 03 19 15 58 00  1651-99 P       co     CCGG   235.910  -999.990 P..  V2 2009 04 25 13 52 00  18.9841  -97.3110  4469.00   274312
X 2009 03 19 15 58 00  1652-99 P       co     CCGG   117.140  -999.990 P..  V2 2009 04 25 14 07 00  18.9841  -97.3110  4469.00   274313
X 2009 03 26 15 38 00  3087-99 P       co     CCGG   101.670     1.552 ...  V2 2009 04 25 13 21 00  18.9841  -97.3110  4469.00   274314
X 2009 03 26 15 38 00  3088-99 P       co     CCGG    99.070     1.577 ...  V2 2009 04 25 13 37 00  18.9841  -97.3110  4469.00   274315
X 2009 04 16 15 20 00  3081-99 P       co     CCGG   139.330     1.153 ...  R5 2009 05 18 13 59 00  18.9841  -97.3110  4469.00   275398
X 2009 04 16 15 20 00  3082-99 P       co     CCGG   140.610     1.156 ...  R5 2009 05 18 14 14 00  18.9841  -97.3110  4469.00   275399
X 2009 04 23 18 30 00   451-99 P       co     CCGG   159.560     1.092 ...  R5 2009 05 28 14 40 00  18.9841  -97.3110  4469.00   276019
X 2009 04 23 18 30 00   452-99 P       co     CCGG   158.810     1.090 ...  R5 2009 05 28 14 54 00  18.9841  -97.3110  4469.00   276020
X 2009 04 25 16 27 00  1773-99 P       co     CCGG   109.520  -999.990 P..  V2 2009 06 01 11 37 00  18.9841  -97.3110  4469.00   276017
X 2009 04 25 16 27 00  1774-99 P       co     CCGG   113.870  -999.990 P..  V2 2009 06 01 11 53 00  18.9841  -97.3110  4469.00   276018
X 2009 05 13 17 15 00   763-99 P       co     CCGG   196.440     1.176 ...  V2 2009 06 18 12 48 00  18.9841  -97.3110  4469.00   276548
X 2009 05 13 17 15 00   764-99 P       co     CCGG   196.040     1.174 ...  V2 2009 06 18 13 03 00  18.9841  -97.3110  4469.00   276549
X 2009 05 21 21 12 00  2183-99 P       co     CCGG   142.690     1.002 ...  V2 2009 06 18 13 19 00  18.9841  -97.3110  4469.00   276550
```

Figura 3.4: Ejemplo de un documento suministrado por el ESRL del monitoreo de CO_2.

La transformación de los datos se hace partiendo del análisis de los datos para poder pasar de la base de datos construida en la figura 3.5 para generar otra estructura semejante a la mostrada en la figura 3.6, correpondiente a un modelo general, llamada estructura multidimensional, aplicando estos conceptos obtenemos 3.7 para el caso de estudio.

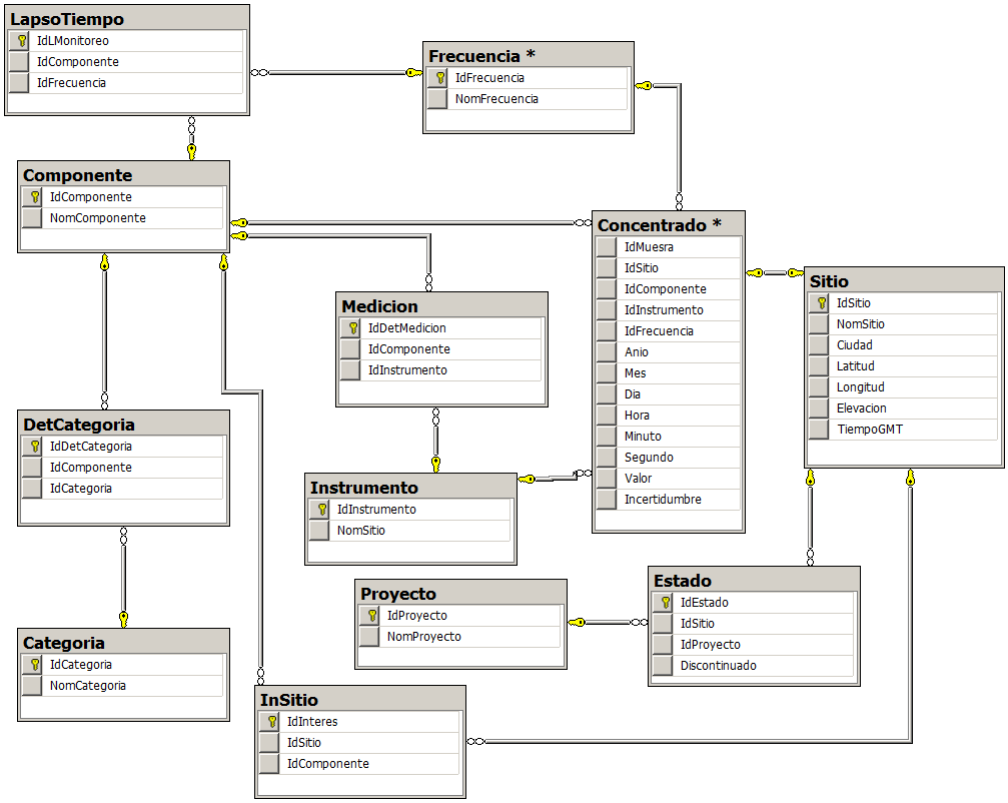


Figura 3.5: Estructura que almacena los datos proporcionados por el ESRL.

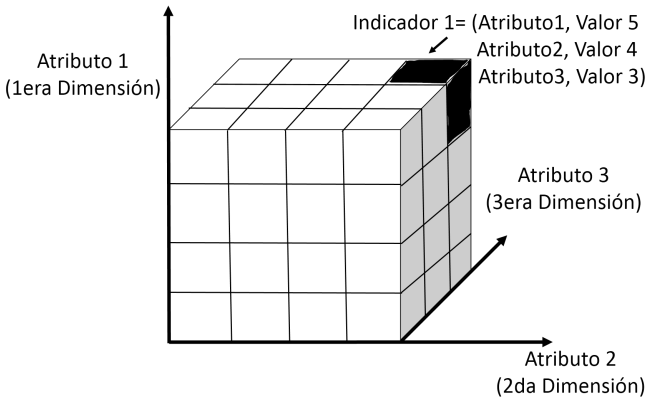


Figura 3.6: Estructura general multidimensional resultado de la transformación de datos.

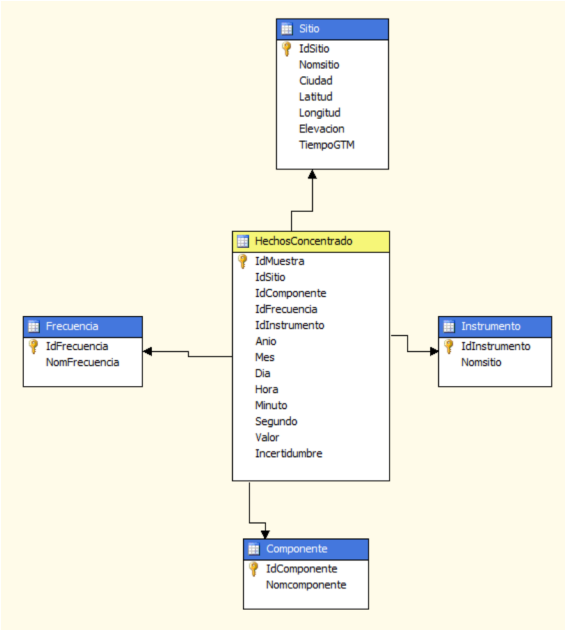


Figura 3.7: Estructura multidimensional resultado de la transformación de datos aplicado al caso de estudio.

3.7. Minería de datos

Los datos almacenados son un tesoro para las organizaciones, es donde se guardan las interacciones pasadas con los clientes, la contabilidad de sus procesos internos, representan la memoria de la organización. Pero con tener memoria no es suficiente, hay que pasar a la acción inteligente sobre los datos para extraer la información que almacenan[5]. Este es el objetivo de la minería de datos. Cualquier problema para el que existan datos históricos almacenados es un problema suceptible de ser tratado mediante técnicas de minería de datos, a continuación se presenta una lista con algunos problemas que pueden ser abordados:

- Búsqueda de lo inesperado por descripción de la realidad multivariante. Un principio clásico de la Estadística, el principio de la parsimonia, ya no es ahora válido (si bien siempre serán preferibles los modelos simples). Para describir un fenómeno cuantas más variables tengamos mejor, más ricas, más globales y más coherentes

serán las descripciones y más fácil será detectar lo inesperado, esto es, aquello que no habíamos previsto y que resulta valioso para entender mejor el comportamiento de algún grupo de individuos, lo cual se ve favorecido por el hecho de trabajar con muestras grandes. Las muestras aleatorias son suficientes para describir la regularidad estadística global, pero no para detectar comportamientos particulares de subgrupos.

- Búsqueda de asociaciones. Un cierto suceso, ¿está asociado a otro suceso?, ¿podemos inferir que determinados sucesos ocurren simultaneamente más de lo que seria esperable si fuesen independientes?, ¿es posible sugerir un producto, sabiendo que otro ha sido adquirido?

- Definición de tipologías. Los consumidores son, a efectos prácticos, infinitos, pero los tipos de consumidores distintos son un número mucho más pequeño. Detectar estos tipos distintos, su perfil de compra y proyectarlos sobre toda la población, es una operación imprescindible a la hora de programar una política de marketing. Por otro lado, las tipologías no tienen que ser necesariamente de consumo, pueden ser de opiniones, valores, condiciones de vida, etc.

- Detección de ciclos temporales. Todo consumidor sigue un ciclo de necesidades que ocasionan actos de compra distintos a lo largo de su vida. Detectar los diferentes ciclos y la fase donde se sitúa cada consumidor ayudará a crear complicidades y adecuar la oferta de productos a las necesidades y crear fidelización.

- Predicción. A menudo deberemos efectuar predicciones: ¿cuál es la probabilidad de baja de un cliente?, ¿cuál es el precio de una vivienda concreta?, ¿lloverá mañana? Estas y muchas más son preguntas que deberemos responder, para ello construiremos un modelo a partir de los datos históricos. Si la variable de respuesta es continua (p. e. la rentabilidad de un cliente) diremos que se trata de un problema de regresión, mientras que si la variable de respuesta es categórica (p. e. la compra o no de un producto) diremos que se trata de un problema de clasificación.

3.7.1. Técnicas de minería de datos

En general, cualquiera que sea el problema a resolver, no existe una única técnica para solucionarlo, sino que puede ser abordado siguiendo aproximaciones distintas. El número de técnicas es muy grande y puede crecer en el futuro, la siguiente es una lista de técnicas con una breve reseña:

- Análisis Factoriales Descriptivos. Permiten hacer visualizaciones de realidades multivariantes complejas y, por ende, manifestar las regularidades estadísticas, así como eventuales discrepancias respecto de aquella y sugerir hipótesis de explicación.

- Market Basket Analysis o análisis de la cesta de la compra. Permite detectar qué productos se adquieren conjuntamente, permite incorporar variables técnicas que ayudan en la interpretación, como el día de la semana, localización, forma de pago. También puede aplicarse en contextos diferentes del de las grandes superficies, en particular el e-comercio, e incorporar el factor temporal.

- Técnicas de clustering. Son técnicas que parten de una medida de proximidad entre individuos y a partir de ahí, buscar los grupos de individuos más parecidos entre sí, según una serie de variables mesuradas.

- Series temporales. A partir de la serie de comportamiento histórica, permite modelizar las componentes básicas de la serie, tendencia, ciclo y estacionalidad y así poder hacer predicciones para el futuro, tales como cifra de ventas, previsión de consumo de un producto o servicio, etc.

- Redes bayesianas. Consiste en representar todos los posibles sucesos en que estamos interesados mediante un grafo de probabilidades condicionales de transición entre sucesos. Puede codificarse a partir del conocimiento de un experto o puede ser inferido a partir de los datos. Permite establecer relaciones causales y efectuar predicciones.

- Modelos lineales generalizados. Son modelos que permiten tratar diferentes tipos de variables de respuesta, por ejemplo la preferencia entre productos concurrentes

en el mercado. Al mismo tiempo, los modelos estadísticos se enriquecen cada vez más y se hacen más flexibles y adaptativos, permitiendo abordar problemas cada vez más complejos: (GAM, Projection Pursuit, PLS, MARS,...).

- Previsión local. La idea de base es que individuos parecidos tendrán comportamientos similares respecto de una cierta variable de respuesta. La técnica consiste en situar los individuos en un espacio euclídeo y hacer predicciones de su comportamiento a partir del comportamiento observado en sus vecinos.

- Redes neuronales. Inspiradas en el modelo biológico, son generalizaciones de modelos estadísticos clásicos. Su novedad radica en el aprendizaje secuencial, el hecho de utilizar transformaciones de las variables originales para la predicción y la no linealidad del modelo. Permite aprender en contextos difíciles, sin precisar la fórmulación de un modelo concreto. Su principal inconveniente es que para el usuario son una caja negra.

- Árboles de decisión. Permiten obtener de forma visual las reglas de decisión bajo las cuales operan los consumidores, a partir de datos históricos almacenados. Su principal ventaja es la facilidad de interpretación.

Se aborda la minería de datos mediante la construcción de un modelo de predicción basado en árboles de decisión. Recordemos que la utilización del algoritmo de minería de datos debe acoplarse a nuestras necesidades, no podemos usar algún algoritmo de clustering, ya que por su definición gastaría demasiados recursos de computo por la complejidad de sus operaciones, la naturaleza de la investigación requiere un algoritmo capaz de manejar grandes cantidades de datos y para ese hecho un algoritmo basado en árboles de decisión representa una opción viable para el problema, su definición se basa más en el cálculo de probabilidades condicionales, entro otros cálculos estadísticos, cuyo procesamiento de cómputo es en comparación de otros más costoso.

3.7.2. Árboles de decisión

Los árboles de decisión son uno de los métodos más importantes que existen en el área del aprendizaje de máquina y en el reconocimiento de patrones[13]. Un árbol de deci-

sión es un algoritmo que conjuga habilidades de análisis y clasificación, cuyos principios básicos son la combinación de la teoría de la probabilidad y herramientas de análisis de en forma de análisis árboles, deriva una jerarquía de reglas de partición con respecto a un atributo objetivo[32].

Existen diferentes maneras de clasificar datos usando redes neuronales, análisis B., árboles de decisión etcétera, sin embargo los árboles de decisión han sido eficientes analizando grandes bases de datos. Actualmente los árboles de decisión son usados exitosamente en la Minería de Datos, ya que es una herramienta de apoyo que usa un árbol o un modelo de decisiones y sus posibles consecuencias.

Los árboles de decisión trabajan dividiendo los conjuntos de datos en subconjuntos más pequeños y más homogéneos. Por lo tanto, se trabaja en un enfoque de divide y vencerás. Esta división recursiva de los conjuntos de datos significa que los subconjuntos en los niveles más bajos tienen muy pocos casos[31].

Para los atributos discretos, el algoritmo hace predicciones basándose en las relaciones entre las columnas de entrada de un conjunto de datos. Utiliza los valores, conocidos como estados, de estas columnas para predecir los estados de una columna que se designa como elemento de predicción.

Específicamente, el algoritmo identifica las columnas de entrada que se correlacionan con la columna de predicción. Por ejemplo, en un escenario para predecir qué clientes van a adquirir probablemente una bicicleta, si nueve de diez clientes jóvenes compran una bicicleta, pero solo lo hacen dos de diez clientes de edad mayor, el algoritmo infiere que la edad es un buen elemento de predicción en la compra de bicicletas. El árbol de decisión realiza predicciones basándose en la tendencia hacia un resultado concreto.

Los datos proporcionados por el NOAA son de tipo continuo, para este caso la Regresión lineal es el método más usado en estadística para predecir valores de variables continuas debido a su fácil interpretación. La naturaleza de los datos a trabajar es de tipo continuo, cada nodo construido en el árbol contiene una fórmula de regresión. Se produce una división en un punto de no linealidad de la fórmula de regresión.

Cuando el algoritmo crea el conjunto de posibles valores de entrada, realiza una selección de características para identificar los atributos y los valores que ofrecen la mayor cantidad de información, el árbol es generado mediante la determinación de las corre-

laciones entre una entrada y el resultado deseado. Una vez correlacionados todos los atributos, se identifica el atributo único que separa más claramente los resultados. Este punto de la mejor separación se mide usando una ecuación que calcula la obtención de información. El atributo que tiene la mejor puntuación para la obtención de información se usa para dividir los casos en subconjuntos, que posteriormente son analizados de forma recursiva por el mismo proceso hasta que no sea posible dividir más el árbol. El uso de regresión lineal es importante durante el trabajo de datos continuos. El análisis de regresión es una técnica estadística para investigar la relación funcional entre dos o más variables, ajustando algún modelo matemático. La regresión lineal simple utiliza una sola variable de regresión y el caso más sencillo es el modelo de línea recta. Supóngase que se tiene un conjunto de n pares de observaciones (x_i, y_i), se busca encontrar una recta que describa de la mejor manera cada uno de esos pares observados

3.8. Árboles autorregresivos

Sea una secuencia temporal de variables $Y = (Y_1, Y_2, \ldots, Y_T)$, las series de tiempo es una secuencia de valores para esas variables denotadas por $y = (y_1, y_2, \ldots, y_T)$, una serie de tiempo es una secuencia de valores para esas variables denotadas por $y = (y_1, y_2, \ldots, y_T)$, se trabaja con modelos del tipo:

$$P(y_t|y_1, \ldots, y_{t-1}, \theta) = f(y_t|t_{t-p}, \ldots, y_{t-1}, \theta), p < t < T \tag{3.1}$$

Donde $f(\cdot|\cdot, \theta)$ es una familia de distribuciones de probabilidades condicionales que representan la forma funcional del modelo y θ es el modelo de parámetros.
El modelo más común usado en el análisis de series de tiempo es el modelo lineal autorregresivo. Un modelo lineal autorregresivo de tamaño p como el de la ecuación 3.1, denotado por AR(p) esta descrito por la siguiente ecuación, en la cual $f(y_t|y_{t-p}, \ldots, y_{t-1}, \theta)$ es una regresión lineal:

$$f(y_t|y_{t-p}, \ldots, y_{t-1}, \theta) = N(m + \sum_{j=1}^{p} b_{iyt-j}, \sigma^2) \tag{3.2}$$

Un Árbol autorregresivo(ART) es un modelo autorregresivo lineal por tramos en los

que los límites son definidos por un árbol de decisiones, y las hojas del árbol de decisión contienen modelos autorregresivos lineales. La Figura 3.8 es un ejemplo de un modelo ART en el que hay tres regiones definidos utilizando la variable Y_{t-1} y cada hoja contiene un modelo AR (1) descrito por la ecuación en la hoja.

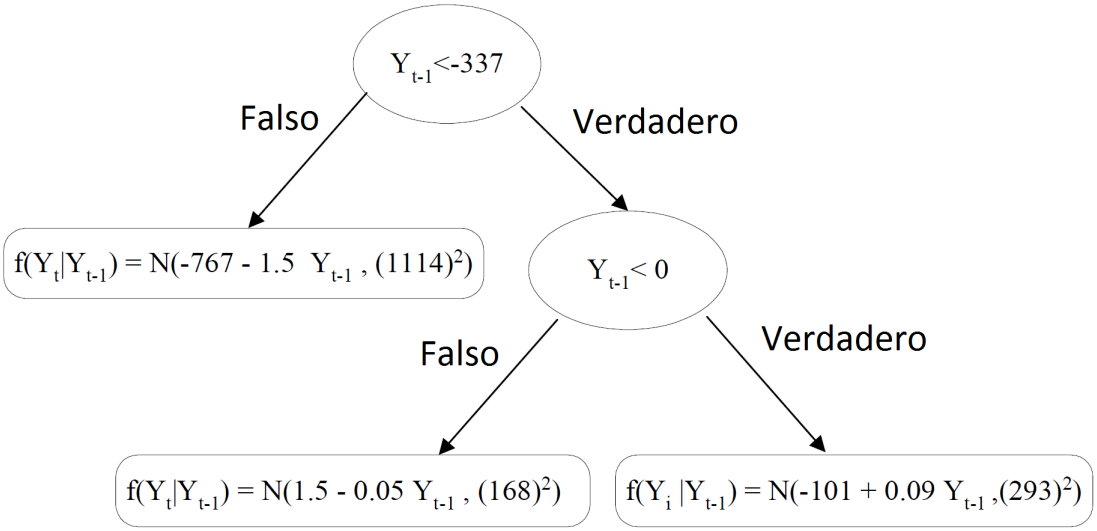


Figura 3.8: Ejemplo de un árbol tipo AR.

3.9. Marco de aprendizaje

Un modelo ART que llamado modelo de árboles autorregresivos de longitud p, denotado por ART(p). Un modelo de ART(p) es un modelo de ART en el que cada hoja del árbol de decisión contiene un modelo AR (p), y las variables de división para el árbol de decisión se eligen de entre las p variables en la serie de tiempo.

En un modelo de tipo ART(p), cada nodo hoja en un árbol de decisión está asociado con una fórmula booleana que es una función de p variables $Y_{t-p}, \ldots, Y_{t-1}$. Por ejemplo el nodo raíz del modelo de tipo ART en la figura 3.8 comprueba si $Y_{t-1} < -337$, se asocia con cada vértice en el árbol de la fórmula(la negación de la fórmula) para este nodo padre su etiqueta en el vértice es verdadero(falso). Se asocia cada hoja l_i un función indicadora, ϕ, que retorna 1 cuando la conjunción de todas las fórmulas asociadas con

los ejes a lo largo de la ruta desde el nodo raíz a la hoja l_i es verdadero, y 0 de otra manera. Por ejemplo, la función indicadora asociada con la mitad de la hoja en la figura 3.8 retorna 1 cuando $(Y_{i-1} < -337) \wedge (Y_{i-1} >= 0)$, y 0 de otra manera. Entones un modelo de tipo ART(p) definido por la ecuación 3.1 donde:

$$F(y_t|y_{t-p},\ldots,y_{t-1},\theta) = \prod_{i=1}^{L} f_i(y_t|y_{t-p},\ldots,y_{t-1},\theta_i)^{\sigma_i} = \prod_{i=1}^{L} N(m_i + \sum_{j=1}^{p} b_{ij}y_{t-j}, \sigma_i^2) \quad (3.3)$$

L es el número de hojas, $\theta = (\theta_1,\ldots,\theta_L)$, y $\theta_i = (m_i, b_{i1},\ldots,b_{ip},\sigma_i^2)$, son los parámetros del modelo para la regresión lineal en la hoja l_i, i=1, ..., L.

Consideremos una definición general del aprendizaje, entonces se tiene una colección de estructuras de modelos alternativos $(s_1, s_2,\ldots,s_S)$ teniendo un modelo desconocido de parámetros $\theta_{s_1},\ldots,\theta_{s_S}$ respectivamente, y nosotros expresamos nuestra incertidumbre acerca de la estructura y los parámetros mediante el colocado de las distribuciones de probabilidad sobre ellos: $p(s)$ y $p(\theta_s|s)$, entonces se usa la regla de Bayes en conjunción con los datos d para inferir distribuciones posteriores sobre estas cantidades: $p(s|d)$ y $p(\theta|d,s)$. En el caso más general se puede hacer predicciones mediante el promedio de estas distribuciones, consecuentemente se usa un modelo bayesiano de selección en el cual se escoge una estructura s que tiene la más alta probabilidad a posteriori del modelo de estructura $p(s|d)$. Mediante la regla de Bayes esta probabilidad a posteriori esta dada por $p(s|d)$ =p(s) $p(d|s)/$ $p(d)$ ya que $p(d)$ es una constante que podemos usar en el producto $p(s)p(d|s)$ para elegir el mejor modelo, nos referimos a este producto como la puntuación Bayesiana para el modelo. El primer término en este modelo es simple, corresponde la estructura a priori, el segundo término $p(d|s)$ es llamada probabilidad marginal y es igual a $\int p(d|\theta,s)p(\theta,s)d\theta$, la probabilidad de los datos $p(d|\theta)$ promediando sobre la incertidumbre en θ. Es interesante notar que cuando se usa un modelo de selección la probabilidad marginal balancea la aptitud de la estructura del modelo de datos con la complejidad del modelo. Una manera de entender este hecho es notar que cuando el número de casos N es grande, la probabilidad marginal puede ser aproximadamente:

$$p(d|\hat{\theta}_s, s) - \frac{|\theta|}{2} logN \quad (3.4)$$

Donde $\hat{\theta}_s$ es el máximo estimador de probabilidad marginal de los datos[**?**] para el modelo s. La primera cantidad en esta expresión representa el grado en que el modelo

se ajusta a los datos, lo que aumenta como el modelo de complejidad aumenta. La segunda cantidad, en contraste, penaliza la complejidad del modelo. Ahora, volvamos a la aplicación del enfoque bayesiano para el aprendizaje de un modelo de series de tiempo fijo, de orden Markov. De acuerdo con la ecuación 3.1, la probabilidad de los datos es:

$$p(y_{p+1}, \ldots, y_T | y_1, \ldots, y_p, \theta, s) = \prod_{t=p+1}^{T} f(y_t | y_{t-p}, \ldots, y_{t-1}, \theta, s) \tag{3.5}$$

Existe un *enfoque de ventanas* en la literatura que reduce el problema de aprendizaje que reduce el problema del aprendizaje en un ordinario modelo de regresión lineal, el corazón del enfoque de ventanas es una transformación de una secuencia $y = (y_1, \ldots, y_t)$ a un conjunto de casos $x^1, \ldots, x^{T-p}$, la transformación está dada por $x^i = (x_1^i, \ldots, x_{p+1}^i)$, para $1 < i < T - p$, donde $x_j^i = y_{i+j-1}$, esta transformación del conjunto de datos "transformación de tamaño p de una serie de tiempo". Como un ejemplo consideremos la secuencia $y = (1, 2, 3, 4)$, entonces la transformación de tamaño 2 es :

$$x^1 = (1, 3), x^2 = (3, 2), x^3 = (2, 4) \tag{3.6}$$

Y la transformación de tamaño 3 es:

$$x^1 = (1, 2, 3), x^2 = (2, 3, 4). \tag{3.7}$$

Dada la transformación podemos reescribir la probabilidad del modelo en la siguiente ecuación:

$$p(p_{y+1}, \ldots, y_T | y_1, \ldots, y_p, \theta, s) = \prod_{t=p+1}^{T} f(x_{p+1}^t | x_1^t, \ldots, x_p^t, \theta, s) \tag{3.8}$$

3.10. La puntuación bayesiana de un modelo ART

Para facilitar el cálculo eficiente, es deseable tener puntuaciones de modelo que puede ser computado en forma cerrada y el factor de acuerdo a la estructura del árbol de decisión. Este es el deseo que nos lleva a hacer las siguientes dos supuestos. Uno, la probabilidad a priori de una estructura de modelo s está dado por:

$$p(s) = k^{|\theta|} \tag{3.9}$$

Donde $0 < k <= 1$ y $|\theta|$ es el número de parámetros del modelo. Los parámetros $\theta_1, \ldots, \theta_L$ son los parámetros asociados con las hojas del árbol de Decisión y son mutuamente independientes, estas suposiciones implican:

$$Puntuacion(s) = \prod_{i=1}^{L} Puntuacionnodohoja(l_i) \tag{3.10}$$

Donde:

$$PuntuacionNodoHoja(l_i) = k^{p+2} \int \prod_{x^t enl_i} f_i(x^t_{p+1}|x^t_1, \ldots, x^t_{p+1}, \ldots, x^t_p, \theta_i, s) p(\theta_i|s) d\theta_i \tag{3.11}$$

f_i es una distribucion Normal correspondente a la regresión lineal de la hoja l_i, la puntuación del nodo hoja es el producto de 3.1 la probabilidad a priori de el componente hoja de la estructura(hay $p+2$ parámetros en cada hoja) y 3.2 la probabilidad marginal de los datos que caen en la hoja.

Hasta ahora se ha seguido una metodología para la generación de conocimiento, y en gran medida la creación de una nueva estructura para el almacenamiento de los datos permitirá la construcción de modelos de predicción de algún gas, además se analizó de manera teórica como es construido un árbol de decisión y se fundamentó el por qué de la elección de esa técnica de Minería de Datos. En el capítulo siguiente se proponen algunos modelos de predicción de CO_2, realizados mediante la construcción de árboles de decisión, para después evaluarlos y así determinar cuál es el más adecuado, sobre las cuales un experto en materia ambiental sería capaz de usar esa herramienta para vaticinar los efectos en el ambiente para poder proponer estrategias de mitigación y/o adaptación en contra del cambio climático.

4

Propuesta de los modelos de predicción

Después del análisis del funcionamiento del algoritmo de árboles de decisión es importante realizar la construcción de varios de ellos basándose en las mismas condiciones para que sus resultados puedan ser equiparables y así poder hacer comparaciones acerca de su efectividad. Durante el desarrollo del capítulo se hace una explicación acerca de la metodología usada así como la forma en la que se calcula la puntuación de cada árbol construido para conocer su efectividad.

4.1. Metodología

Se realizaron modelos para realizar la minería de datos utilizando árboles de decisión, en las siguientes secciones se mostraran algunos resultados obtenidos de la realización de dichos modelos. Las características que adoptan los datos a considerar son las siguientes:

- Se tiene disponible una muestra de concentraciones de CO_2 del periodo 2009-2013 reportadas mediante un valor que equivale al promedio mensual de las concentraciones.
- El número total de sitios que pueden proporcionar esta misma información son 91(el sitio de monitoreo de México se encuentra contenido en el conjunto).
- Cada algoritmo puede calibrarse para lograr un mayor índice de confiabilidad con el fin de descartar los modelos que sean menos eficientes.

Las tareas involucradas en la construcción de modelos que predicen la concentración de CO_2 requieren definir:

- El modelo de minería de datos toma como datos de entrada: Sitio, Año, Mes.
- El modelo de minería de datos debe de arrojar como valor de predicción el valor de la concentración mensual dado un Sitio, un es y un Año.
- Se realizan los modelos con los datos del ciclo 2009-2012 y los parámetros por defecto para ambos casos.
- Se evalúa el modelo construido utilizando como entrada los Sitios, Años y Meses correspondientes al año 2013, con el fin de compararse con los valores reales proporcionados por el ESRL.
- Calibrar ambos modelos hasta encontrar alguno que ofrezca un mayor índice de confiabilidad.
- Analizar los resultados y realizar conclusiones.

4.2. Diagnóstico de un modelo de regresión

Recordemos que el criterio que utiliza un árbol de decisión para hacer las ramas es el de la regresión lineal. Un aspecto que se olvida frecuentemente es que los modelos de regresión se basan en hacer determinadas suposiciones sobre los datos y que éstas no siempre se cumplen, por lo que es preciso comprobar si las hipótesis básicas del modelo se dan en nuestros datos, es lo que se conoce como diagnóstico del modelo.
En el caso de los modelos de regresión lineal se utiliza el concepto de residuo: diferencia entre el valor observado y el valor estimado por la ecuación de regresión, es decir lo que la ecuación de regresión no explica para cada unidad de observación.
En un modelo de regresión lineal que sea adecuado los residuos deben seguir una distribución normal con media 0 y varianza constante. La representación de los residuos frente a cada una de las variables independientes X nos permite detectar la falta de linealidad o la heterocedasticidad (se dice que existe heterocedasticidad cuando la dispersión o varianza de la variable no es constante y varía con el valor de ésta).

4.3. Construyendo un modelo de árboles de decisión

Previamente se ha analizado la manera en la que un árbol de decisión funciona para predecir datos de tipo continuo, se abordará la construcción de modelos de predicción utilizando ese algoritmo, cuya concepción está sujeta a varias condiciones y eso implica que pueda generarse más de un modelo de predicción cada vez que se establezcan diferentes parámetros. La situación que se pesenta es calibrar al algoritmo de árbol de decisión para producir un buen modelo de predicción. Para este caso se debe considerar los parámetros que afectan directamente el árbol:

- Método de puntuación.
- Establecimiento de los regresores.
- Umbral para el crecimiento del árbol.

En la secciones siguientes se describe de manera más detallada cada uno de los aspectos mencionados así como su importancia dentro de la construcción del árbol.

4.3.1. Método de puntuación

Parte importante dentro de la minería de datos es la utilización previa de algoritmos de selección de características para mejorar el análisis y así reducir la carga del procesamiento, a continuación se hace mención de los algoritmos utilizados para realizar esta tarea.

Entropía de Shannon

La entropía de Shannon mide la incertidumbre de una variable aleatoria para un determinado resultado.Por ejemplo, la entropía de lanzar una moneda al aire para decidir algo a cara o cruz se puede representar como una función de la probabilidad de que salga cara. Se utiliza la fórmula siguiente para calcular la entropía de Shannon:

$$H(X) = -\sum P(xi)log(P(xi)) \tag{4.1}$$

Bayesiano con prioridad K2

Una red bayesiana es un gráfico dirigido o acíclico de estados y de transiciones entre ellos; esto significa que algunos estados siempre son anteriores al estado actual y otros son posteriores, y que el gráfico no se repite ni realiza bucles. Las redes bayesianas permiten el uso del conocimiento previo. Sin embargo, la pregunta sobre que estados anteriores se deben utilizar para calcular las probabilidades de los estados posteriores es importante para la precisión, el rendimiento y el diseño del algoritmo. El algoritmo K2 está basado en la optimización de una medida. Esa medida se usa para explorar, mediante un algoritmo de ascensión de colinas, el espacio de búsqueda formado por todas las redes que contienen las variables de la base de datos. Se parte de una red inicial y ésta se va modificando (añadiendo arcos, borrándolos o cambiándolos de dirección) obteniendo una nueva red con mejor medida. En concreto, la medida K2 (Cooper and Herskovits 1992) para una red G y una base de datos D es la siguiente:

$$f(G:D) = logP(G) + \sum_{i=1}^{n}[\sum_{k=1}^{S_i}[log\frac{\Gamma(\eta_{ik})}{\Gamma(N_{ik}+\eta_{ik})} + \sum_{j=1}^{r_i} log\frac{\Gamma(N_{ijk}+\eta_{ijk})}{\Gamma(\eta_{ijk})}]] \qquad (4.2)$$

donde N_{ijk} es la frecuencia de las configuraciones encontradas en la base de datos D de las variables x_i, donde n es el número de variables, tomando su j-ésimo valor y sus padres en G tomando su k-ésima configuración, donde si es el número de configuraciones posibles del conjunto de padres y r_i es el número de valores que puede tomar la variable x_i. Además, $N_{ik} = \sum_{r_i}^{j=1} N_{ijk}$ y Γ es la función Gamma.

Equivalente Dirichlet bayesiano con prioridad uniforme

La puntuación Equivalente Dirichlet bayesiano (BDE) también utiliza el análisis bayesiano para evaluar una red dado un conjunto de datos.El método de puntuación BDE fue desarrollado por Heckerman y está basado en la métrica BD desarrollada por Cooper y Herskovits.La distribución Dirichlet es una distribución multinomial que describe la probabilidad condicional de cada variable de la red y dispone de muchas propiedades que son útiles para el aprendizaje. El método Equivalente Dirichlet bayesiano con prioridad uniforme (BDEU) considera un caso especial de la distribución Dirichlet, en el que se utiliza una constante matemática para crear una distribución fija o uniforme de esta-

dos anteriores.La puntuación BDE también considera la equivalencia de probabilidad; esto significa que no es de esperar que los datos diferencien estructuras equivalentes.En otras palabras, si la puntuación para *If A Then B* es la misma que la puntuación para *If B Then A*, las estructuras no se pueden diferenciar basándose en los datos, y no se puede deducir la causalidad.

De acuerdo a la metodología señalada a continuación se enlistaran una serie de variaciones realizadas en los parámetros para la realización de un modelo basado en árboles de decisión que sea el más apto en la predicción de la concentración de CO_2, para después ser diagnosticado y determinar el mejor modelo que predicción.

4.3.2. Establecer los regresores

La elección de n atributos a utilizar como regresores dentro del algoritmo fuerza al algoritmo independientemente de su importancia en los cálculos del algoritmo, mejor conocidas dentro de la regresión lineal como variables explicativas, independientes o regresores.

4.3.3. Umbral de crecimiento del árbol de decisión

La selección de características es aplicada a las salidas del árbol para reducir el número de resultados posibles y generar más rápido el modelo, es importante cambiar el umbral para la realización de la selección de características incrementando o disminuyendo el número de valores posibles, un umbral bajo aumenta el número de divisiones y un valor alto los reduce, los valores recomendados son 0.5 (de 1 a 9 atributos), 0.9 (de 10 a 99 atributos) y 0.99(para 100 o más atributos).

4.4. Modelos de predicción basado en árboles de decisión

Durante esta sección se presentan los modelos desarrollados con las variantes mencionadas en la sección anterior. En principio se trabaja con la creación de un modelo de

predicción x basado en árboles de decisión sujeto a las especificaciones mostradas en el cuadro 4.1; es posible generar el árbol que puede verse en la Figura 4.1.

Modelo x	
Regresores: Año	
U. Crecimiento: 0.9	
Puntuación: Entropía	**Niveles Generados: 6**

Cuadro 4.1: Variaciones y resultados de las variaciones en el algoritmo de árboles de decisión para obtener el modelo x.

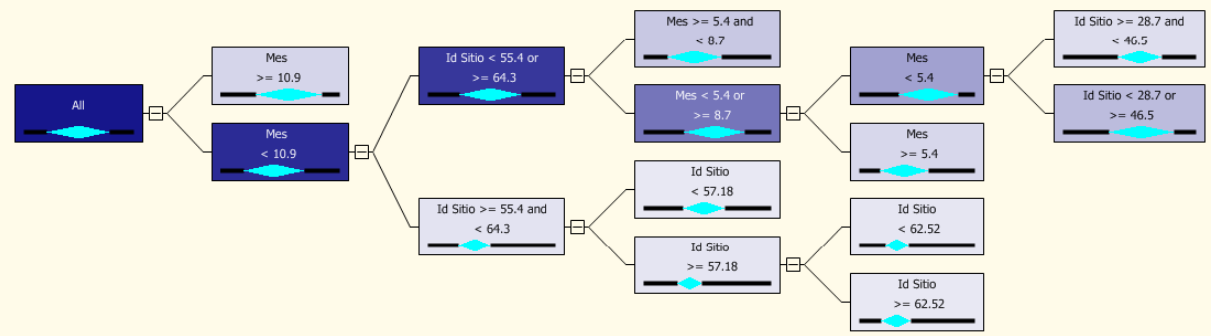


Figura 4.1: Árbol de decisión correspondiente al modelo x.

Una manera de poder observar y analizar el comportamiento de los valores en el modelo x es mediante un diagrama de correlación, en la Figura 4.2 , se puede observar inmediatamente que la correlación existente entre la concentración de los valores predichos y los valores reales es de tipo positiva o directa, ya que el coeficiente de correlación es de 1.41.

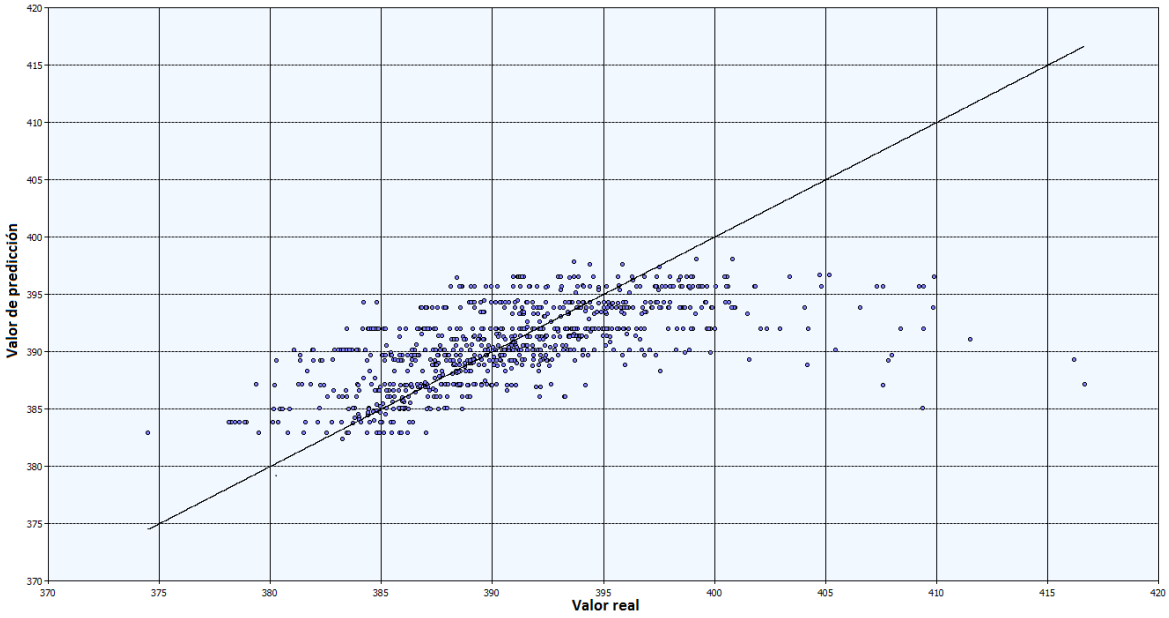


Figura 4.2: Gráfica de correlación correspondiente al modelo x.

Sin embargo el modelo x no corresponde a un buen modelo para la predicción de CO_2, puede observarse de manera clara que la correlación existente para el modelo x es bastante débil, en el intervalo de [400, 420] los puntos que se predicen están demasiado alejados de la recta de correlación, lo ideal en la búsqueda de un modelo de predicción basándonos en árboles de decisión es que los puntos se encuentren lo más cercano posible a la recta.

Se realizaron variaciones en el algoritmo para generar distintos árboles de decisión(ver Cuadro 4.2), algunos diagramas de correlación pueden observarse en la Figura 4.3. A partir de los modelos construidos es posible hacer las siguientes observaciones:

- Definir como regresor el identificador de *Sitio* causa que los modelos obtenidos no brinden un coeficiente de correlación alto, debido a que no influye en el cálculo la variación de otros parámetros dentro de la construcción del árbol.

- El uso de Año y Mes como regresores brindan un coeficiente de relación más alto. Sin embargo esto no quiere decir que la predicción de los datos tenga un buen diagnóstico.

- El método de puntuación basado en entropía no brinda un apoyo considerable en la construcción de un árbol eficiente.

- El umbral para el crecimiento del árbol que brinda mejores predicciones es de 0.99, esto se debe a la gran cantidad de datos manejados.

- Utilizar el Equivalente Bayesiano de Dirichlet como puntuación arroja buenos resultados en la predicción.

Sin embargo la interpretación gráfica puede ser engañosa, es necesaria una medida para describir que tan preciso es el pronóstico de Y con base en X, o a la inversa, es decir, que tan exacta es la estimación. Esta medida se denomina error estándar de estimación, es el mismo concepto que la desviación estándar. Mientras que la desviación estándar mide la dispersión respecto a la media, el error estándar de estimación mide la dispersión respecto de la recta de correlasión.

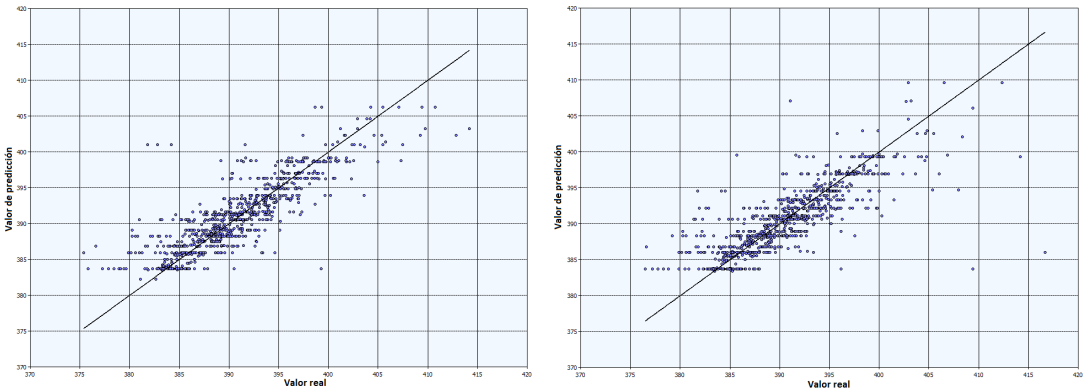
$$\textbf{\textit{Error estándar de estimación}} = \sqrt{\frac{(Y-\hat{Y})^2}{n-k}}$$

Donde n=Tamaño de la muestra, Y=Valor de la variable de predicción, $\hat{Y} = Valorestimado$, k=Número de variables explicativas.

En la tabla 4.3 podemos observar el error estándar de estimación calculado para cada modelo construido, es visible que el Modelo e es quien predice los datos de una mejor manera, recordemos que mientras más pequeño es el error estándar de estimación los datos están menos dispersos sobre la línea de correlación, en la Figura 4.4 observamos que los puntos predichos se acercan más a la recta de correlación en comparación a los modelos contenidos en la Figura 4.3.

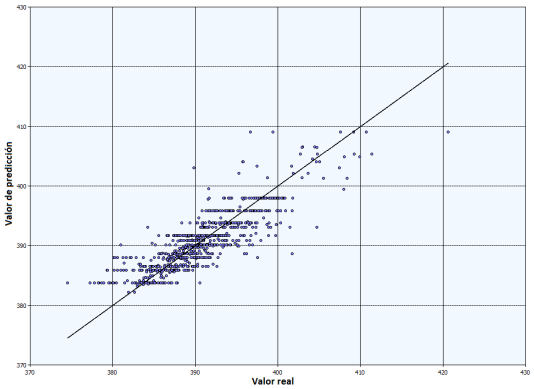
Mod	Regresor	Crecimiento	Puntuación	Niveles	Coef. Correlación
a	Año	0.5	Entropía	15	0.45
b	Año	0.9	Entropía	21	0.32
c	Año	0.99	Entropía	33	0.9
d	Año	0.5	Eq. B.de Dirichlet	6	0.7
e	Año	0.99	Eq. B. de Dirichlet	35	2.66
g	Año	0.99	Eq. B. de Dirichlet	32	1.45
h	Año	0.5	B. con prioridad k2	20	1.76
i	Año	0.9	B. con prioridad k2	27	0.63
j	Año	0.99	B. con prioridad k2	35	0.65
k	Mes	0.5	Entropía	23	0.89
l	Mes	0.9	Entropía	14	1.7
m	Mes	0.99	Entropía	26	2.32
n	Mes	0.5	Eq. B. de Dirichlet	7	1.52
o	Mes	0.9	Eq. B. de Dirichlet	4	1.02
p	Mes	0.99	Eq. B. de Dirichlet	5	1.9
q	Mes	0.5	B. con prioridad k2	12	1.26
r	Mes	0.9	B. con prioridad k2	14	1.34
s	Mes	0.99	B. con prioridad k2	35	1.1
t	Sitio	0.5	Entropía	12	0.5
u	Sitio	0.9	Entropía	6	0.54
v	Sitio	0.99	Entropía	38	0.8
w	Sitio	0.5	Eq. B. de Dirichlet	16	0.89
x	Sitio	0.9	Eq. B. de Dirichlet	9	0.56
y	Sitio	0.99	Eq. B. de Dirichlet	34	0.63
z	Sitio	0.5	B. con prioridad k2	14	0.67
$a1$	Sitio	0.9	B. con prioridad k2	8	0.56
$b1$	Sitio	0.99	B. con prioridad k2	34	0.67

Cuadro 4.2: Variaciones y resultados de las variaciones en el algoritmo de árboles de decisión en la generación de modelos.

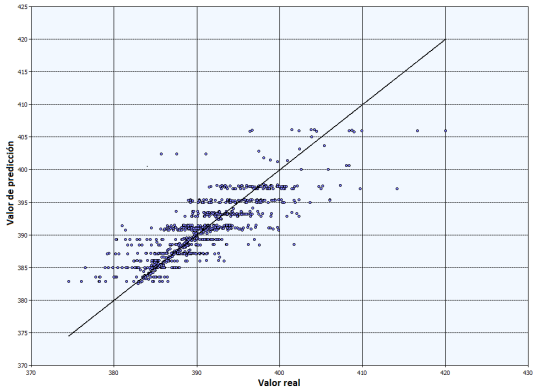


(a) Modelo a

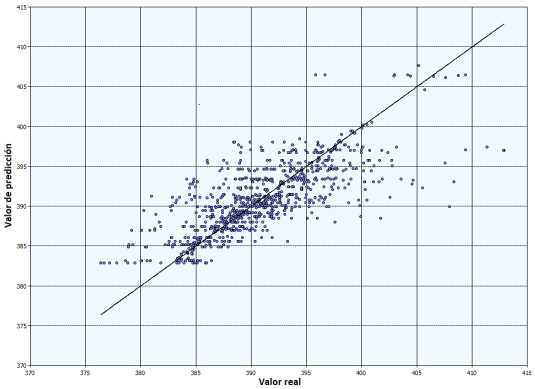
(b) Modelo b



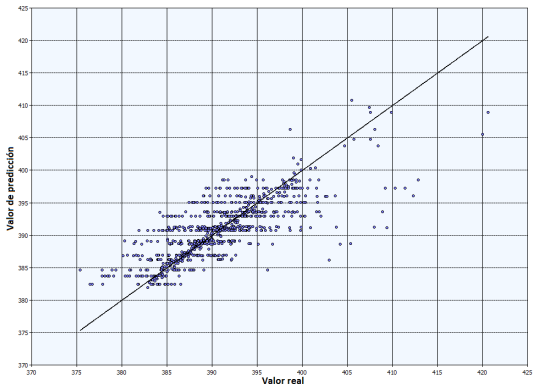
(c) Modelo c



(d) Modelo f



(e) Modelo g



(f) Modelo i

Figura 4.3: Diagramas de correlación pertenecientes a algunos modelos de árboles de decisión construidos

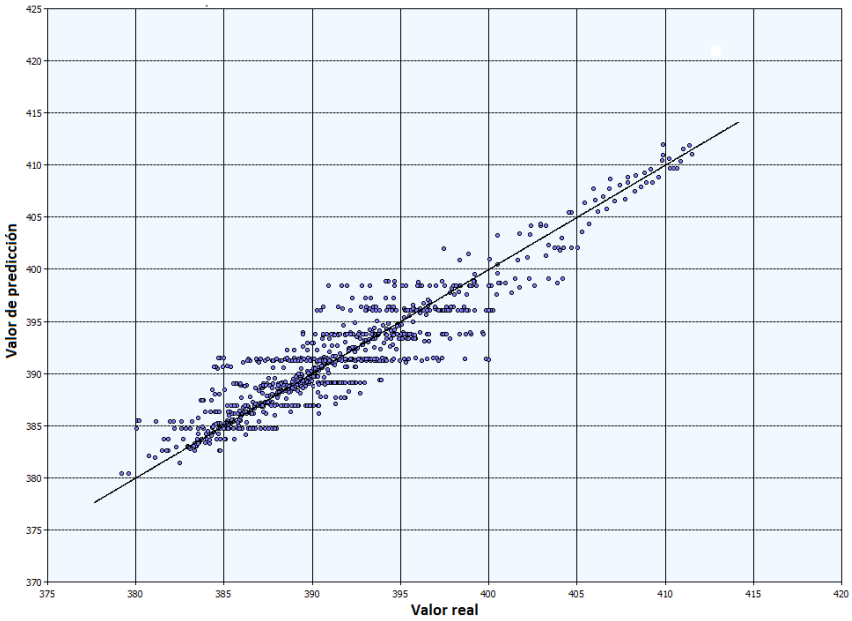


Figura 4.4: Gráfica de dispersión correspondiente al modelo e.

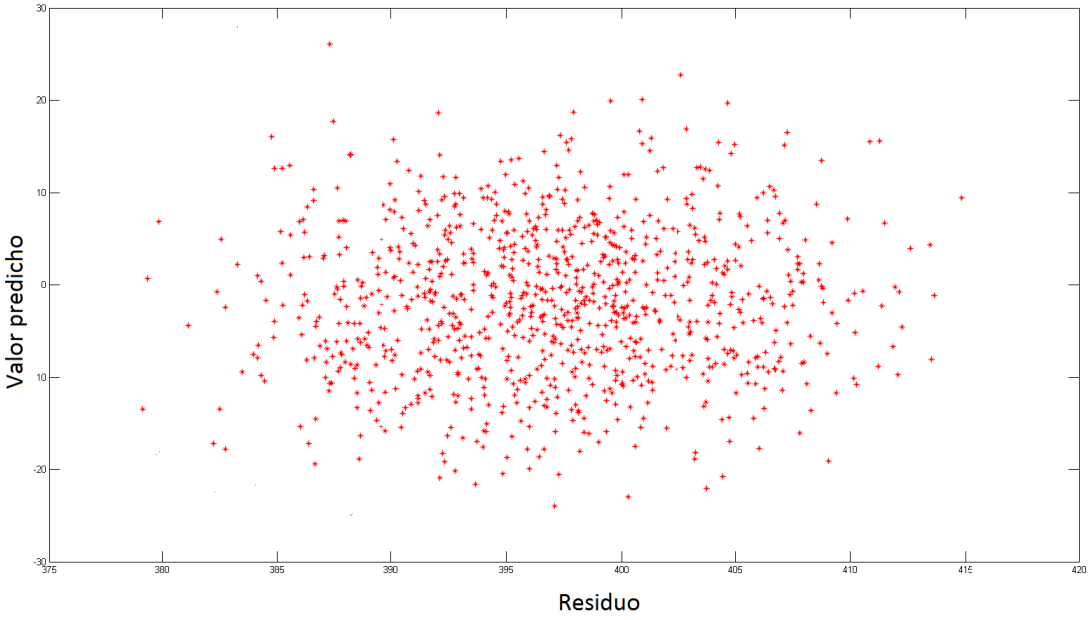


Figura 4.5: Gráfica de relación entre el residuo y el valor predicho.

Mod	Error estándar de estimación	Mod	Error estándar de estimación
a	14.749300233	b	9.783444290
c	6.783499281	d	8.349848934
e	2.520756499	g	9.673499221
h	19.673434010	i	28.672399220
j	21.459377299	k	10.784377001
l	13.783300007	m	18.459037289
n	24.673490566	o	28.746637711
p	16.563944781	q	16.569364777
r	20.784366722	s	19.674593675
t	17.673948220	u	19.673499288
v	13.674904893	w	21.673452008
x	18.674555390	y	12.897845983
z	9.7834598555	a1	10.784599239

Cuadro 4.3: Error estándar de estimación para los modelos de predición.

Se hace total énfasis en el modelo e que hasta ahora ha presentado ser un buen modelo de predicción. Podemos ver en la Figura 4.5 una gráfica de los datos realizada como se ha mencionado en la sección de diagnóstico para un modelo de regresión, cuyo comportamiento parece estar distribuido de manera normal, al menos gráficamente hablando, puede verse claramenteque tiene una media de cero y la varianza parece ser constante, sin embargo es necesario hacer otra prueba para para definir la normalidad de los datos de manera más efectiva, se determinó que el conjunto de datos predichos por el *Modelo e* siguen una distribución normal, se utilizó la prueba de Lilliefors que está incluida en Matlab para determinar si corresponde a una distribución normal, la gráfica que comprueba la naturaleza de la distribución puede verse en 4.6, podemos asegurar que está normalmente distribuído y por lo tanto las hipótesis básicas del modelo propuesto se cumplen.

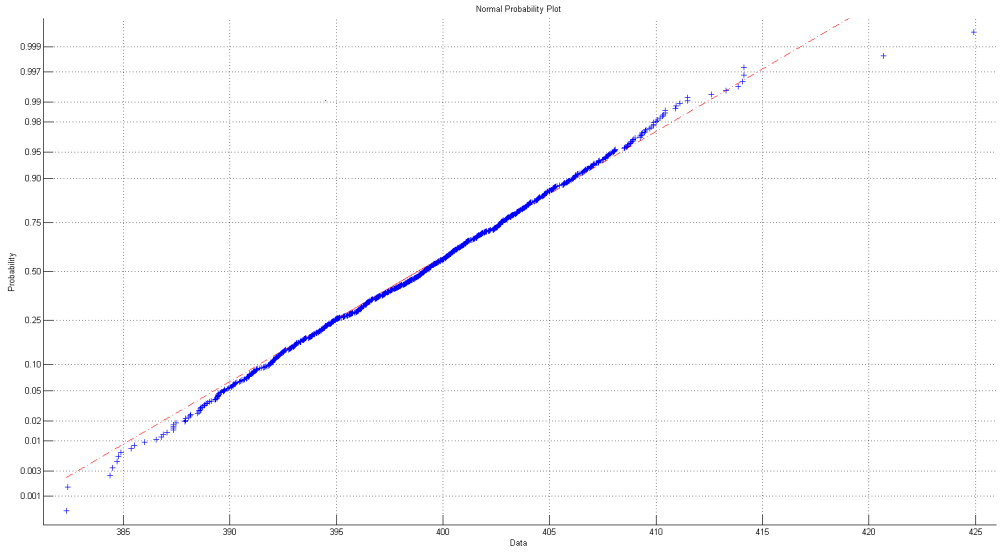


Figura 4.6: Comportamiento normal de los datos predichos por el Modelo e.

De acuerdo a las observaciones mencionadas es fácil inferir que el modelo que mejores resultados obtuvo fue el modelo e por sus características, cuya gráfica de correlación puede observarse en la figura 4.4. Los datos generados corresponden igual que el modelo x a una correlación positiva, sin embargo de acuerdo a las variaciones realizadas es mucho más fuerte. Una parte del árbol construido del modelo e puede observarse en la Figura 4.7, se muestran 4 de 35 niveles.

Cada nodo contenido en los árboles de decisión contiene una fórmula, para el caso específico del nodo con la leyenda $Idsitio < 85 or >= 86$; el valor a predecir puede ser calculado mediante:

$$\boldsymbol{Concentración}\ CO_2\ = 390{,}579 + 1{,}976 * (\boldsymbol{Mes} - 10{,}495 + 2{,}326 * (\boldsymbol{Año} - 2, 10{,}444))$$

En materia de predicción no existe un modelo que sea considerado como correcto, sin embargo puede ofrecerse uno que brinde una "buena"predicción mediante la verificación de los parámetros mencionados anteriormente.

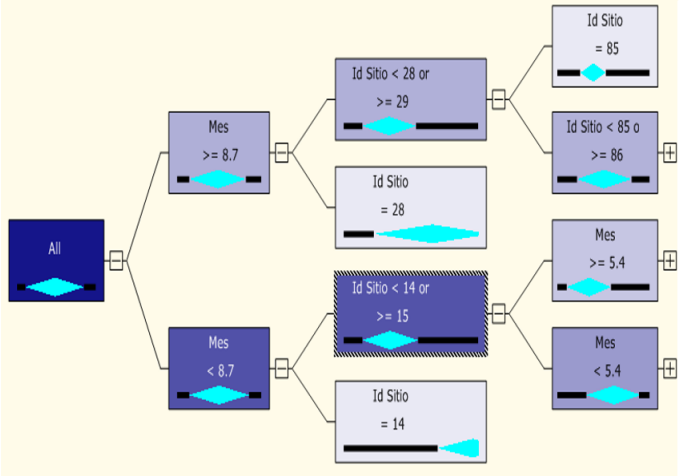


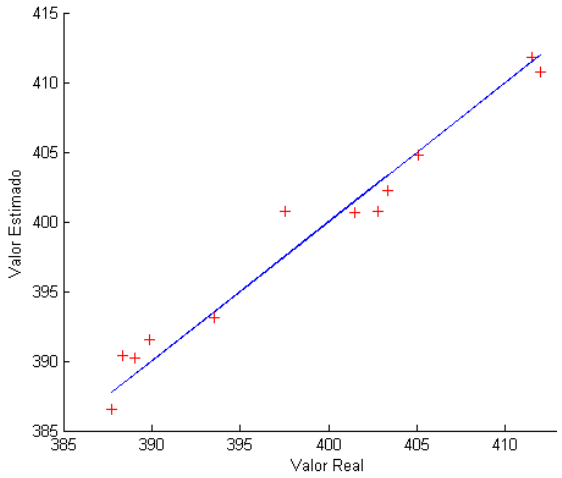
Figura 4.7: Árbol correspondiente al Modelo e.

4.4.1. Prediciendo con valores conocidos

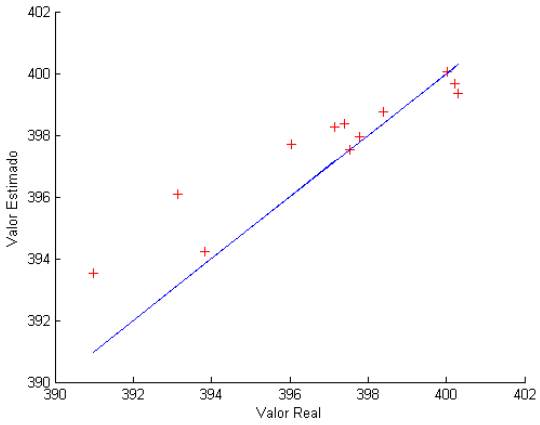
La ventaja más grande de trabajar con una visión de BI es que no se hacen conjeturas acerca de los datos de interés, sino que, es posible abarcar con todos los datos puesto que se trabaja con el histórico de las concentraciones de CO_2, lo que permite llevar a cabo la comparación y medir la calidad de las predicciones.

Para analizar la efectividad del modelo en el comportamiento de valores predichos de una manera más específica se realiza una comparación con los valores reales registrados, se presentan las gráficas correspondientes al nivel de la concentración de CO_2 en México, Alemania, Canadá, Estados unidos, China y Reino Unido en el año 2013, los sitios mencionados son importantes ya que podemos comparar las concentraciones de CO_2 de los países que se encuentran en el grupo de los 7 exceptuando México y China, la importancia de las últimas 2 radica en que China se ha hecho bastante popular en la emisión de contaminantes perjudiciales en la atmósfera, y México es el país en donde se desarrolla la investigación.

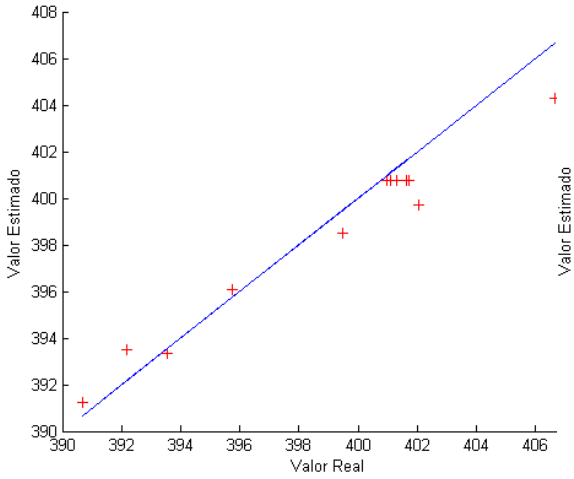
Hasta ahora se han hecho comparaciones con los datos prueba que se encuentran dentro de los datos correspondientes al periodo 2009-2012, se ha identificado un modelo de predicción que ha sido evaluado correctamente de acuerdo a un modelo de regresión lineal, es momento de analizar el comportamiento de las predicciones después de introducir



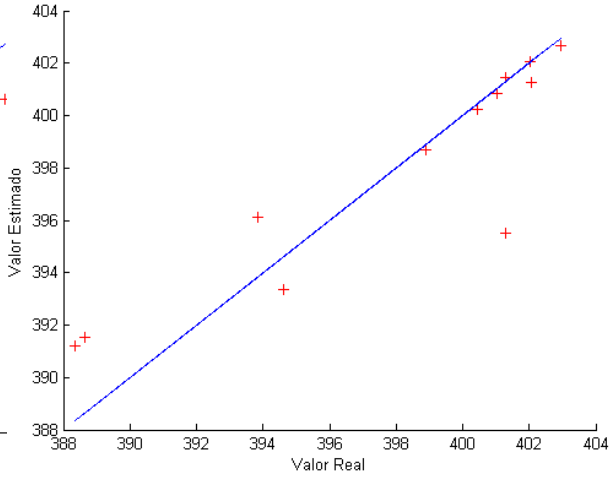
(a) Alemania



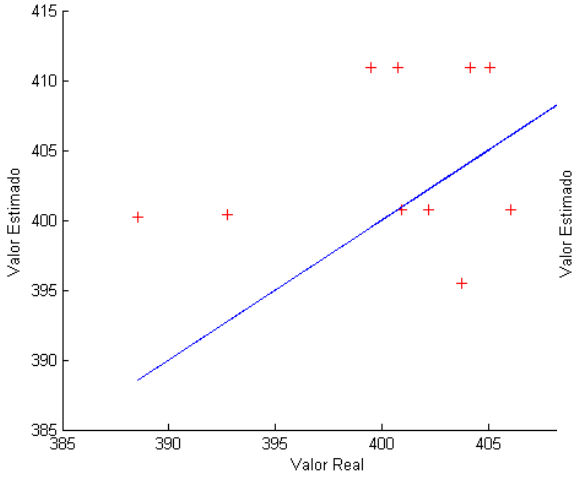
(b) Reino Unido



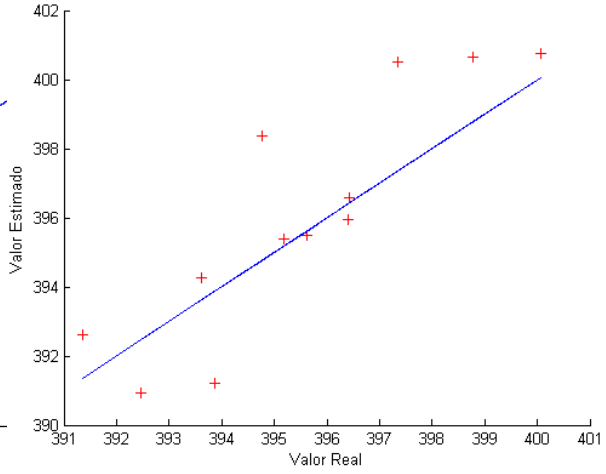
(c) Estados Unidos



(d) Canadá



(e) China



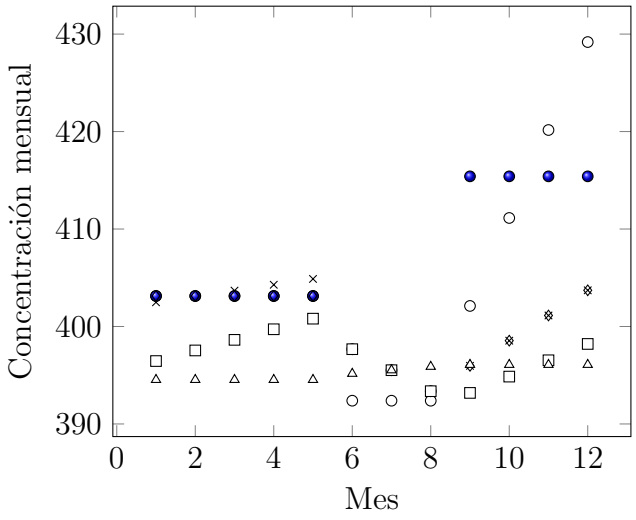
(f) México

Figura 4.8: Predicciones para la concentración de CO_2 obtenidas mediante el Modelo e para el año 2013

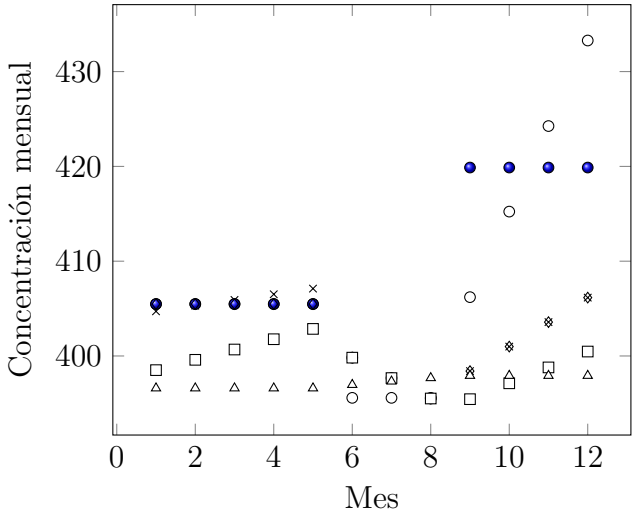
los parámetros correspondientes al año 2013, una vez devuelto el valor predicho de la concentración CO_2 mediante el modelo basado en árboles de decisión se procederá a la comparación con el valor real registrado por los centros de monitoreo. Las predicciones efectuadas en los sitios mencionados con el modelo desarrollado pueden observarse en las figuras contenidas en 4.8, es posible verificar que las predicciones arrojadas por el modelo estan ubicadas alrededor de la recta de correlación con poca dispersión a comparación de las que se han analizado en secciones anteriores.
El modelo desarrollado hace posible obtener las concentraciones de CO_2 correspondientes al año 2014(véase Figura 4.9a) que para algunos sitios no ha sido registrado, México es uno de esos casos, y predecir las concentraciones del 2015 y 2016 en adelante(véase Figura 4.9 y véase Figura 4.10), las observaciones de tales predicciones pueden ser una herramienta importante en la toma de decisiones, ya que de primera vista vemos que el aumento en las emisiones de CO_2 es innegable.

4.4.2. Interpretación del modelo

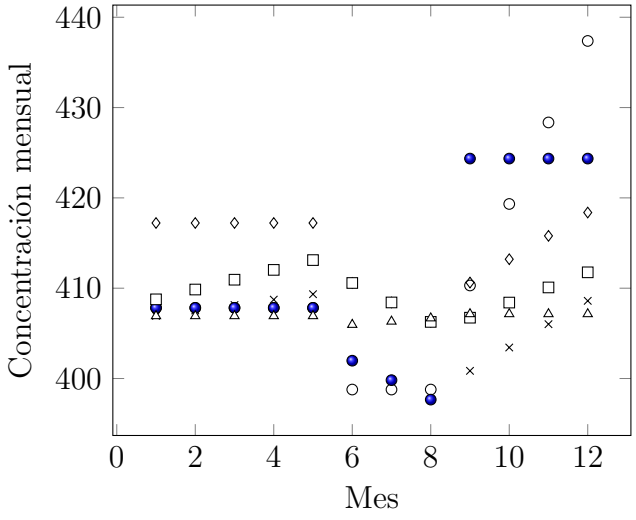
En el desarrollo de modelos de solución de inteligencia de negocios existe un ámbito en el que funcionan como apoyo en la toma de decisiones, a continuación se hacen algunas observaciones acerca de las repercusiones que trae el creciente aumento de CO_2 que pueden enriquecerse más con la intervención de un experto en materia ambiental.
Dentro del grupo de los GEI es importante el estudio de aquellos gases cuyo incremento se ha elevado de manera súbita desde la aparición del hombre, dentro de ellos está el CO_2, su importancia dentro del desarrollo del trabajo de tesis ha desembocado en modelo de predicción. De acuerdo a las gráficas obtenidas es claro el aumento en la concentración de dicho gas, las gráficas no son solo cifras, en primera instancia es evidente que se presenta mayor aumento de cncentración CO_2 en los meses donde hace más frio, esto se debe al llamado efecto de inversión térmica que provoca que el aire sea más denso, además el aumento del CO_2 trae consecuencias graves en el ambiente debido a que las moléculas que componen este gas son capaces de absorber y reemitir con energía extra otro fotón infrarrojo, esta habilidad causa que el CO_2 sea un gas de efecto invernadero que atrapa calor de manera efectiva. Los efectos del CO_2 no involucran solamente a los devastadores fenómenos de tipo climático adjudicados también a los



(a) Predicciones de CO_2 para el año 2014.

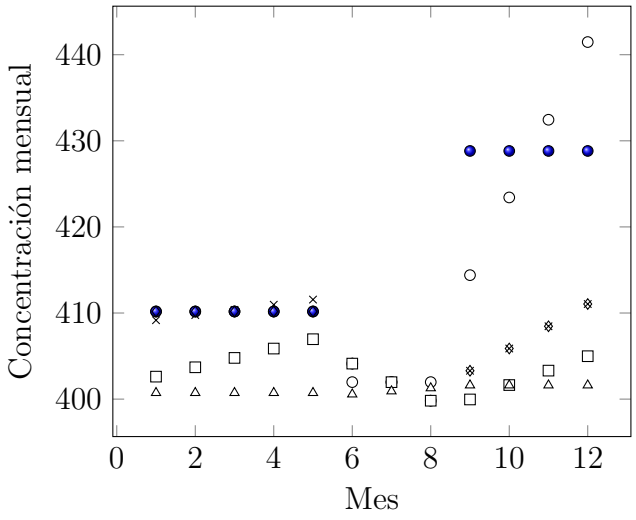


(b) Predicciones de CO_2 para el año 2015.

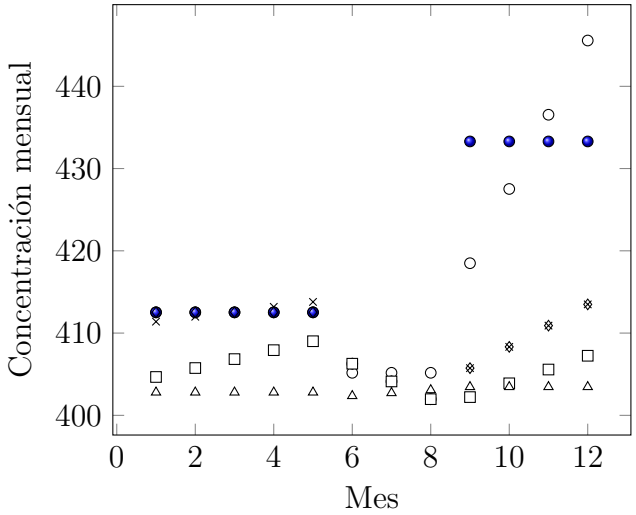


(c) Predicciones de CO_2 para el año 2016.

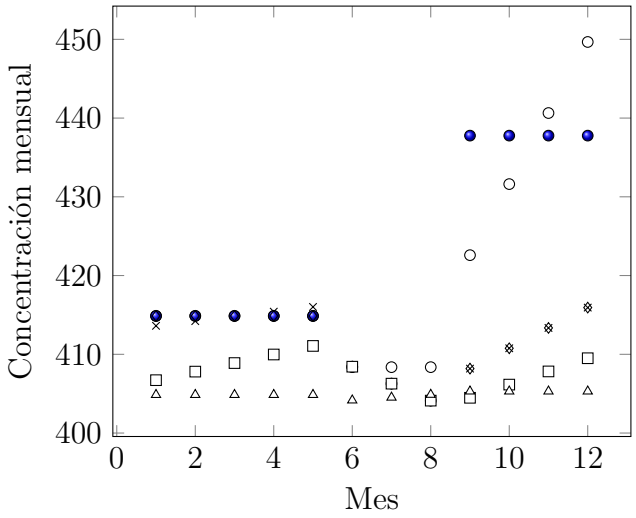
Figura 4.9: Alemania(○), Canadá(x), China(●), E.U(◇), México(□), Reino Unido(△).



(a) Predicciones para el año 2017.



(b) Predicciones para el año 2018.



(c) Predicciones para el año 2019.

Figura 4.10: Alemania(○), Canadá(x), China(●), E. U(◇), México(□), Reino Unido(△).

otros GEI, como inundaciones, ciclones, sequias, terremotos, olas de frio o calor, en específico las afectaciones a nivel vegetación generan el encogimiento de los estomas durante el proceso de fotosíntesis, que causa que se libere menos agua reduciendo la capacidad de refrescamiento de las plantas lo que reduce su tasa de crecimiento, por si fuera poco cada aumento de un grado Celsius en la temperatura superficial global puede traer consigo 12 por ciento más rayos cuyos rayos van más allá de inconvenientes eléctricos en los hogares, si no que más relámpagos significa más chispas para más incendios forestales y, curiosamente, un poco más de la contaminación en el aire. Los inconvenientes causados por las altas concentraciones de CO_2 van más allá de los visibles hablando de la superficie, ya que modifican de manera catastrófica la acidez de los océanos que a su vez repercute de manera dramática en la flora y la fauna marina, afectando la capacidad de algunos organismos para producir y mantener sus conchas, los estragos envuelven también a los arrecifes de coral, ya que puede verse afectada su capacidad de reconstruirse, lo que pone en peligro a aproximadamente un millón de especies que dependen de su hábitat. Los efectos de la alta concentración del CO_2 son realmente brutales, la mayoría forma la base para una serie de eventos en cadena que afectan a todos los organismos del planeta. A sabiendas de los daños en el ambiente que pueden ocurrir a causa del CO_2 son los líderes de organizaciones gubernamentales quienes deciden qué acciones son necesarias para bajar los niveles proponiendo planes para administrar la huella del carbono en empresas y concientizando a las personas a seguir un estilo de vida enfocado al respeto ambiental, en último caso se procedería a promover acciones de tipo adaptativas ante el inminente cambio en la atmósfera.

5

Conclusiones y trabajo a futuro

Durante el desarrollo del trabajo de investigación se han abordado problemáticas actuales que son enfrentadas en el ámbito científico en la búsqueda de la construcción de modelos de predicción de gases que afectan de manera crucial en el fenómeno del cambio climático, en primera instancia exigen la reestructuración de los almacenes de datos actuales para poder ser consumidos por la comunidad científica de una manera más eficiente, dinámica, que hagan posible que cualquier investigador pueda obtener datos sin la necesidad de realizar tareas que corresponden a la minería de datos, mismas que pueden retraer el desarrollo de un proyecto de investigación. De acuerdo a la necesidad multidisciplinaria actual de conocer el comportamiento a largo plazo del cambio climático se desarrolló un modelo matemático de predicción del CO_2 a partir de datos obtenidos del ESRL mediante el uso de árboles de decisión y obteniendo los datos a partir del modelo de almacenamiento propuesto, que hace posible la generación dinámica de una nueva estructura de datos. En el marco del desarrollo de estrategias que mitiguen el fenómeno del cambio climático es necesario conocer, fomentar y construir el uso de nuevas tecnologías que permitan conocer el comportamiento de los fenómenos, para que a partir de ello se puedan realizar nuevas acciones de prevención o de acondicionamiento en la sociedad actual. Para el caso del trabajo de tesis el concepto de modelo de solución de inteligencia de negocios se acopla como una herramienta capaz de soportar los requerimientos en el eficiente manejo de datos, así como la construcción de nuevos modelos de predicción.

En la construcción de modelos realizados para comprender fenómenos es necesario tomar en cuenta el mayor número de datos posibles, en el contexto del cambio climático por su injerencia, es esencial utilizar la mayor cantidad de datos provenientes de todas las partes del mundo, no solo para tener una mayor variedad, si no que la naturaleza del fenómeno a estudiar establece su devastadores efectos de manera global.

Bibliografía

[1] About NOAA(Octubre 2015). Recuperado de http://www.noaa.gov/about-noaa.html

[2] Adopted, I. P. C. C. (2014). CLIMATE CHANGE 2014 SYNTHESIS REPORT

[3] Alpert, J.C., et al.; 2009: "High Availability Applications for NOMADS at the NOAA Web Operations Center Aimed at Providing Reliable Real Time Access to Operational Model Data.;.American Geophysical Union Spring meeting conference proceedings

[4] Auffhammer, M., & Carson, R. T. (2008). Forecasting the path of China's CO 2 emissions using province-level information. Journal of Environmental Economics and Management, 55(3), 229-247

[5] Banet, T. A. (2001). La minería de datos, entre la estadística y la inteligencia artificial. Questiió: Quaderns d'Estadística, Sistemes, Informatica i Investigació Operativa, 25(3), 479-498

[6] Belll, M.L., A. McDermott, S.L. Zeger, J.M. Samet y F. Dominici. 2004. Ozone and Short-term Mortality in 95 US Urban Communities, 1987-2000. Journal of the American Medical Association 292: 2372-2378

[7] Breiman, L. ; Friedman, J.H. ; Olshen, R.A. ; Stone, C.J.: Classification And Regression Trees. Boca Raton : CHAPMAN & HALL/CRC, 1984

[8] C. Patel, K. Supekar, Y. Lee, and E. K. Park, "OntoKhoj: A Semantic Web Portal for Ontology Searching, Ranking and Classification", In Proc. of the Workshop On Web Information And Data Management, ACM, 2003

[9] Caldeira, K., & Wickett, M. E. (2005). Ocean model predictions of chemistry changes from carbon dioxide emissions to the atmosphere and ocean. Journal of Geophysical Research: Oceans (19782012), 110(C9)

[10] Chris A. Mattmann, Duane Waliser, Jinwon Kim, Paul Ramirez, Cameron Goodale, Andrew F. Hart, Paul Loikith, Huikyo Lee, Michael Joyce, Maziyar Boustani, Shakeh Khudikyan, Kim Whitehall, Jesslyn Whittell, Paul Zimdars, Daniel Crichton, Yolanda Gil, Luca Cinquini1Model Evaluation Using the NASA Regional Climate Model Evaluation System (RCMES).2013

[11] Edenhofer, O., Pichs-Madruga, R., Sokona, Y., Farahani, E., Kadner, S., Seyboth, K., ... & Minx, J. C. (2014). Climate change 2014: mitigation of climate change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, 511-597

[12] Evans, J., J. Spengler, J. Levy, J. Hammitt, H. Suh, P. Serra-no-Trespalacios, L. Rojas-Bracho, C. Santos-Burgoa, H. Rojas-Rodriguez, M. Caballero-Ramirez y M. Castillejos. 2000 Contaminación Atmosférica y Salud Humana en la Ciudad de México, MIT-IPURGAP Reporte No. 10, 109 pp.83

[13] F. Esposito; D. Malerba; G. semeraro. A comparative analysis of methods for pruning decision trees, IEEE transactions on pattern analysis and machine intelligence. Vol. 19, NO. 5, MAY, 1997

[14] Euzenat, J., Stuckenschmidt, H.: the family of language approach to semantic interoperability. In Omelayenko, B., Klein, M., eds.: knowledge transformation for the semantic Web, Amsterdam (NL), IOS press (2003) 49-63

[15] García-Castro, R., et al.: Specification of a methodology, general criteria, and benchmark suites for benchmarking ontology tools. Deliverable D2.1.4, KnowledgeWeb EU-IST Network of Excellence IST-2004-507482 (2005)

[16] Geng SUN, Yan-hong FENG, Xian-jiu GUO. The K-means Clustering algorithm[J]. Journal of Changchun Normal University. February 2011, VoI30 No.I): 1-4

[17] GENL (Gobierno del Estado de Nuevo León), Semarnap y SSA. 1997. Programa de la administración de la calidad del aire del área metropolitana de Monterrey 1997-2000. México 172 pp

[18] Gómez, A. A. R., & Bautista, D. W. R. (2010). Inteligencia de negocios: Estado del arte. Scientia et Technica, 1(44), 321-326

[19] Gruber, T. *A translation approach to portable ontology specifications*. Knowledge Acquisition 5:199-220; 1993

[20] Hong Dayong. *A comprehensive judge and sampling analysis of the citizens environmental awareness*. Science and technology review, 1998 (9) :13-16

[21] IEEE: IEEE Std 610-1990 IEEE Standard Computer Dictionary: A compilation of IEEE Standard Computer Glossary (1990) New York, NY

[22] Inmon, W.H., Bonnie, O. & Fryman, L. (2008). Business Metadata Capturing Enterprise Knowledge. Elsevier Inc. ISBN: 978-0-12-373726-7

[23] Intergovernmental Panel on Climate Change, Climate Change 2007: The Physical Science Basis *http* : $//ipccwg1.ucar.edu/wg1/Report/AR4WG1_{P}rint_{S}PM.pdf, 2007$

[24] IPCC, 2007. Climate Change 2007: Synthesis Report. Contribution of Working Groups I, II and III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, Pachauri, R.K and Reisinger, A. (Eds.)]. IPCC, Geneva, Switzerland

[25] Jayanthi Ranjan, "Business Intelligence: Concepts, Components, Techniques and Benefits", Journal of Theoretical and Applied Information Technology, 2009 Inmon, W.H., Bonnie, O. & Fryman, L. (2008). Business Metadata Capturing Enterprise Knowledge. Elsevier Inc. ISBN: 978-0-12-373726-7

[26] Jayanthi Ranjan, "Business Intelligence: Concepts, Components, Techniques and Benefits", Journal of Theoretical and Applied Information Technology, 2009

[27] Jiawei Han, Micheline Kamber. Data mining: Concepts and techniques[M].2nd.[S.I.]: Morgan Kaufmann Publisher, 2006: 256-266

[28] Juan, W., Hai-zhen, Y., & Zhi-bo, L. (2009, October). Prospect of energy-related carbon dioxide emission in China based on scenario analysis. In Energy and Environment Technology, 2009. ICEET 09. International Conference on (Vol. 3, pp. 90-93). IEEE

[29] Michalewicz, Z., Schmidt, M., Michalewicz, & Chiriac, C. 2007. Adaptative Bussines Intelligence. Springer-Verlag Berlin Heidelberg. ISBN-13, 978-3-540-32928-2 Springer Berlin Heidelberg New York.

[30] MITCHELL, T. M. Machine Learning. McGraw-Hill, 1997^a.

[31] MITCHELL, T. M. Machine Learning. McGraw-Hill, 1997^a.

[32] M. Zhao and X. Li, "An application of spatial decision tree for classification of air pollution index", in Geoinformatics, 2011 19th International Conference on, june 2011, pp. 1-6.

[33] NSC, National Safety Council. Carbon Monoxide.2005

[34] Qin Dahe, Chen Zhenlin, Luo Yong. *Updated understanding of climate change sciences*. Advances in climate change sciences research, 2007,3:63-73

[35] Ranjan, J. (2009). Business intelligence: Concepts, components, techniques and benefits. Journal of Theoretical and Applied Information Technology, 9(1), 60-70. ISO 690

[36] Romieu, I., F. Meneses, J.J.L. Sienra-Monge, J. Huerta, S.R. Velasco, M. C. White, R. A. Etzel y M. Hernandez-Avila. 1995. Effects of urban air pollutants on emergency visits for childhood asthma in Mexico City. American Journal of Epidemiology 141: 546-553

[37] Rubén Darío Alvarado. Metodología para el desarrollo de Ontologías, 2010

[38] Rutledge, G., Crichton, D., Alpert J, "Improving Numerical Weather Prediction Models and Data-Access Latencies", Earthzine, March 29, 2014

[39] Shengyi Jiang and Xia Li, A Hybrid Clustering Algorithm, in proc. of 2009 Sixth International Conference on Fuzzy Systems and Knowledeg Discovery, 2009: 366-370

[40] Sheta, A. F., Ghatasheh, N., & Faris, H. (2015, April). Forecasting global carbon dioxide emission using auto-regressive with eXogenous input and evolutionary product unit neural network models. In Information and Communication Systems (ICICS), 2015 6th International Conference on (pp. 182-187). IEEE

[41] Shih, S. H., & Tsokos, C. P. (2008). Prediction models for carbon dioxide emissions and the atmosphere. Neural, Parallel and Scientific Computations, 16(1), 165

[42] Tellez-Rojo M.M., I. Romieu, M. Polo-Pena, S. Ruiz-Velasco, F. Meneses-Gonzalez y M. Hernandez-Avila. 1997. Effect of environmental pollution on medical visits for respiratory infections in children in Mexico City. Salud Pública de México 39: 513-22

[43] Tim Bernes Lee, James Hendler, and Ora Lasilla. *The semantic Web.* Scientific American, 2001

[44] Theodoridis, S., & Koutroumbas, K. (2010). Introduction to Pattern Recognition: A MATLAB Approach. Burlington: Academic Press

[45] UNEP-GRID ARENDAL. "Vital climate graphics", en: Introduction to climate change: planets and atmospheres. UNEP- GRID ARENDAL. www.climateark.org/vital/02.htm

[46] USEPA. Automobiles and Carbon Monoxide. 2003

[47] Yang Xing. *A Study on Framework Convention on Climate Change.* Doctoral dissertation in Wu Han university . 2005