Universidad del Azuay

Facultad de Ciencias de la Administración

Escuela de Ingeniería de Sistemas y Telemática

APLICACIÓN DE MINERÍA DE DATOS Y VISUALIZACIÓN DE GRANDES VOLÚMENES DE DATOS EN EL ANÁLISIS DE CONTAMINANTES ATMOSFÉRICOS Y VARIABLES METEOROLÓGICAS RECOLECTADOS POR EL IERSE

Trabajo de titulación previo a la obtención del grado de Ingeniero de Sistemas y Telemática

Autor:
Juan Diego Peñaloza Bonilla

Director:
Ing. Marcos Patricio Orellana Cordero

Cuenca – Ecuador
2021

DEDICATORIA

Este trabajo de graduación, está dedicado a Dios, por ser el apoyo espiritual inspirador, que me ha dado fortaleza y constancia en este largo proceso para obtener una meta indispensable anhelada en mi vida.

De manera muy especial dedico este trabajo de graduación a mi madre **Martha**, quien ha sido el pilar y la principal guía, la persona que me ha enseñado a transcender por el camino correcto, por medio de sus consejos, palabras y reflexiones que siempre me ha incentivado, fortalecido, y acompañado en todo momento dentro de mi desarrollo personal, deportivo, académico y profesional, con su inmensurable amor, sacrificio, por esa particular manera de nunca rendirse y todo su apoyo incondicional a lo largo de estos años. Este logro es tuyo y para ti madrecita. Así también, para mi hermano **Marco**, quien nunca dudo de mí, por su ayuda en cada situación, por su comprensión y porque ha sido siempre mi gran ejemplo a seguir, de triunfo, de lucha, de gran trabajo, dedicación y perseverancia para obtener las metas trazadas.

Para mi amada abuelita Irene, "Mamita", quien desde el cielo me cuida, me protege y me acompaña en toda instancia, quien fue un gran soporte en mi vida, como una madre; y de quien aprendí los más sinceros valores, físicamente no estás aquí a mi lado, pero siento que estás junto a mí siempre, en cada momento y aunque faltaron muchas cosas por vivir juntos, sé que este momento hubiera sido tan especial para ti como lo es para mí.

A mi padre Manuel, por las experiencias compartidas, por la paciencia, por inculcar en mí el esfuerzo, valentía y humildad, y por los sacrificios realizados para brindarme siempre lo mejor. A mi hermano Paúl, por su incentivo moral, por su disposición, carácter, y ser otro ejemplo influyente en mi desarrollo.

Finalmente, a mi tío Hugo, por siempre estar pendiente de mi bienestar, encaminándome siempre a ser una buena persona, un buen hombre; a mi tío Hernán, por su bondad, a mi tía Bertha por sus detalles, realmente con todo mi cariño a toda mi familia cercana: abuelos, primas, primos, cuñadas, sobrinas, sobrino, etc. Quienes de diferentes maneras son parte de esta gran meta y siempre han sabido animarme para llegar a cumplir este gran objetivo.

AGRADECIMIENTO

Quiero primero agradecer a Dios, por todo lo que me ha dado en el trayecto de mi vida, por la salud, perseverancia, constancia y sabiduría para culminar los objetivos propuestos en esta etapa académica.

En realidad, me faltan palabras para agradecer a todas las personas que han sido parte fundamental para el desempeño de este trabajo de titulación, sin embargo, la mención y reconocimiento del esfuerzo y sacrificio es necesaria. De todo corazón agradezco principalmente a mi madre Martha, la cual siempre estuvo en los momentos indicados, escuchándome, guiándome e incentivándome cada instante, con el propósito de que culmine mi carrera universitaria, dándome el suficiente apoyo y el impulso necesario para no decaer ni fracasar cuando todo parecía verdaderamente difícil, complicado, e imposible; a mi padre Manuel, a mi hermano Paúl, y en especial a mi hermano Marco, por la confianza depositada en mí, que a pesar de cualquier error o desacierto cometido, siempre ha sabido brindarme sus sabias palabras, recomendaciones y el gran ejemplo a seguir, transmitiéndome sus experiencias, para lograr formar la persona que soy actualmente.

Agradezco mucho a mi tío Hugo, que supo aconsejarme de la mejor manera, por compartir su sabiduría, y por siempre estar al tanto de mí, en los procesos de formación que he atravesado. A mi abuelita Irene, que fue pionera en plantearme este reto, y que me enseño el sentido del respeto, la veracidad y la unidad familiar. A mi abuelo Rigoberto Bonilla, por estar compartiendo junto a mí este objetivo. A mis tíos Bertha y Hernán, por su gran nobleza y bondad.

Agradezco infinitamente a toda mi familia, por ser un importante pilar en mi vida, por su incondicional y valioso apoyo, por la paciencia, por ayudarme a perseverar, por el ánimo transmitido, por estar a mi lado en los buenos y malos momentos de manera incondicional, siendo parte de este arduo proceso.

Expreso un agradecimiento muy cordial para mi director de tesis, Ing. Marcos Orellana Cordero, por brindarme su tiempo, orientarme y compartirme sus conocimientos para el desarrollo de este trabajo de titulación correctamente.

ÍNDICE

ÍNDICE DE CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

RESUMEN:

El trabajo presenta la metodología para la visualización de grandes volúmenes de información utilizando técnicas de preprocesamiento y visualización aplicadas a gráficas de Mapas de Coordenadas para información recolectada de contaminantes atmosféricos y variables meteorológicas de la ciudad de Cuenca. Empleando la metodología CRISP-DM y aplicando algoritmos de clusterización se realizó el análisis y la interpretación para identificar patrones de comportamiento en periodos estacionarios del año 2018. Estos patrones serán útiles para la toma de decisiones del gestor ambiental. Se analizó grandes volúmenes de datos mediante procesos de minería de datos, obteniendo correlaciones que se visualizan en una aplicación web dinámica. El gráfico multidimensional utilizó la herramienta D3.js, que interactúa con determinaciones colorimétricas, selecciones, agrupamientos, detalle de datos, pincelados donde es posible seleccionar datos desde el propio gráfico e incluir o descartar dimensiones. Se encontraron varios patrones, entre los que cuentan el ozono (O3) que tiene correlación directa con la temperatura (TEMPAIRE) y los rayos ultravioletas (UVA).

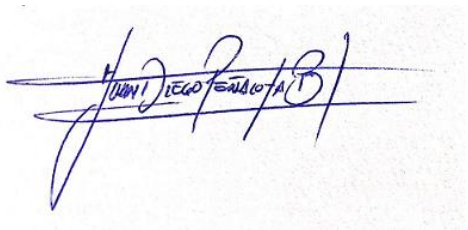
Palabras clave: IERSE, LIDI, CRISP-DM, Clustering, X-Means, K-Means, D3.js, visualización VIZ.

ABSTRACT:

The work presents the methodology for the visualization of large volumes of information using pre-processing and visualization techniques applied to coordinate map charts for information collected on atmospheric pollutants and meteorological variables of the city of Cuenca. Using the CRISP-DM methodology and applying clustering algorithms, the analysis and interpretation was carried out to identify behavior patterns in stationary periods of the year 2018. These patterns will be useful for the decision making of the environmental manager. Large volumes of data were analyzed through data mining processes, obtaining correlations that are visualized in a dynamic web application. The multidimensional graph used the tool D3.js, which interacts with colorimetric determinations, selections, groupings, data detail, brushstrokes where it is possible to select data from the graph itself and include or discard dimensions. Several patterns were found, including ozone (O3) which has direct correlation with temperature (TEMPAIRE) and ultraviolet rays (UVA).

Keywords: IERSE, LIDI, CRISP-DM, Clustering, X-Means, K-Means, D3.js, VIZ visualization.

Translated by



1. INTRODUCCIÓN

El área informática ha evolucionado de forma muy rápida en los últimos años y se ha concentrado en la construcción de sistemas de información y tratamiento de datos, es decir, se aprovecha de mejor manera el conocimiento a las bases de datos (BD). La vanguardia tecnológica ha determinado a la minería de datos (DM), los extensos contenidos de información (Big Data) y técnicas de visualización de datos (VIZ) como ciencias pioneras en la creación de conocimiento, que facilitan aplicaciones en diferentes áreas, y son de gran aporte para el usuario o sus gestores (Furht & Villanustre, 2016).

Entonces, se puede utilizar herramientas de software de gestión, existiendo en la actualidad, herramientas buenas y eficaces. Las alternativas, conforme al progreso de los sistemas informáticos y las grandes cantidades de información que se almacenan, han determinado excelentes resultados hacia DM. Se conceptualiza a la minería de datos como una herramienta generadora de conocimiento, siendo el medio para las decisiones en los diferentes campos de estudio y aplicación. (Kingsy et al., 2016).

Una de las causas del deceso de miles de personas en el mundo es atribuido a la contaminación del aire provocado por los gases de invernadero, esta situación ha provocado un gran deterioro en la capa de ozono, afectando gravemente al planeta y generando un descontrolado calentamiento global que ha modificado las condiciones de vida en la naturaleza de varias especies. (Kingsy, Manimegalai, Geetha, Rajathi, Usha, Raabiathul 2016; Lanjewar & Shah, 2012).

El alto índice de contaminación es uno de los mayores causantes de víctimas mortales con diferentes afecciones en sus sistemas respiratorios, e incluso se ha evidenciado agentes cancerígenos en los pulmones que se atribuyen a este problema. Así también, puede afectar con enfermedades oftalmológicas, o causar sensibilidades en la piel que se convierten en rigurosos problemas dermatológicos, esto según las estadísticas valoradas por la Organización Mundial de la Salud (OMS, 2017). La contaminación del aire es una grave preocupación para la sociedad y autoridades, debido a que estos agentes son muy peligrosos por naturaleza (Gore & Deshpande, 2017).

Desde un punto de vista socio político, la contaminación ambiental, ha sido un tema de análisis recurrente para los organismos internacionales, entes gubernamentales, instituciones locales y municipalidades, los mismos que han implementado normativas, reglamentos y distintas campañas para concientizar a los generadores de esta problemática; en este contexto, la información que se pueda recopilar y que contribuya para la ejecución de acciones que lleven a su reducción resulta imprescindible.

Un gran causante de la contaminación se atribuye a la industrialización de las sociedades, otro de los factores de importante consideración es la generalización de uso de vehículos, los cuales se han incrementado de forma exagerada en las urbes (Gore & Deshpande, 2017). Hoy en día, se ha vuelto común encontrar algún tipo de contaminación atmosférica en los entornos que se frecuenta diariamente, por tal razón, se debe tomar medidas preventivas.

Es crucial analizar, gestionar e intentar controlar la contaminación del aire; para ello, se debe recopilar, tratar y analizar información donde sea posible realizar una medición, de tal forma que los resultados permitan construir y elegir correctos planes de acción, todo esto con el propósito de reducir o eliminar esta afección (Gore & Deshpande, 2017).

Distintas localidades a través de sus gobiernos municipales, han implementado sofisticados aparatos de medición de la contaminación, cuyo propósito es realizar política pública a partir de los resultados, es necesario entonces, aplicar técnicas y conocimientos para la recolección, extracción y tratamiento de datos para analizar la distribución de la calidad del aire en las diferentes ciudades y sectores (Christina & Komathy, 2013).

Con los grandes avances tecnológicos, existe la manera de recolectar información de forma automatizada, mediante el uso de sensores electrónicos ambientales, de tal manera que se pueda obtener información para el análisis y gestión de los agentes que influyen de forma negativa a nuestro ambiente. Las partes interesadas en la gestión de la calidad del aire y la contaminación ambiental, en distintos países y ciudades del mundo han invertido, tiempo, dinero y conocimiento; para crear y distribuir estaciones de monitoreo en puntos estratégicos, que permitan la recolección de datos, con los que se puedan a partir de los resultados tomar desiciones.

Estas redes de monitoreo, recopilan registros en bases de datos y servidores de las diferentes variables: atmosféricas y meteorológicas; y con las técnicas de análisis adecuadas, permiten revelar resultados efectivos de fácil comprensión de los niveles de contaminación en base a factores de relación (Andrade Durazno, 2018; EMOV EP, 2017).

Se puede aplicar la minería de datos (DM), para el tratamiento de los datos que son recolectados por las estaciones de monitoreo, en la cual se implementan metodologías, por ejemplo (CRISP-DM), que permite la ejecución de los pasos guías, donde se puede adaptar de forma flexible un problema específico, procesos y fases detalladas, con un orden de flujo de aplicación, obteniendo una fácil interpretación de los resultados procesados. Existen varios tipos de software especializado que permite la ejecución de una metodología de minería de datos, por ejemplo (RapidMiner), y con el uso de estas herramientas, se puede cumplir con los objetivos del análisis de los datos. Dentro de la metodología, se determina la zona de estudio y la representación de resultados de manera comprensiva, e inclusive técnicas de visualización avanzada (Baloian, 2016).

Al hablar de técnicas avanzadas de visualización e interpretación de esta gran cantidad de información, se puede mencionar a los Mapas de Coordenadas Paralelas, que es una técnica poderosa, ya que transmite la información de manera efectiva de las bases de datos en un solo repaso visual. Estas interpretaciones visuales nos permitirán identificar causas raíces de lo que se trata, facilita además las suposiciones adaptables, e incluso podrían aportar para aplicar ciertos métodos de predicción, según sean los datos estadísticos (Peña, 2018).

Permiten, presentar tendencias y comportamientos que aporte a los organismos gestores para tener un conocimiento, una comprensión simple y general del problema, consecuentemente, plantear la debida intervención para control y erradicación de riesgos de ciertos sectores; se puede concientizar de mejor manera el uso de estrategias, tanto de manera preventiva, como también de forma correctiva. Además, que no todos los contaminantes y variables meteorológicas, tienen un mismo comportamiento; es necesario manejarlos de forma diferenciada (Peña, 2018).

Este trabajo de titulación permitirá, confirmar hipótesis planteadas sobre correlaciones relevantes, por ejemplo, la relación directa entre el ozono (O3) y la temperatura (TEMPAIRE) que encontró como resultado Andrade (2018). También, aporta la fácil visualización las asociaciones entre variables, haciendo que el análisis sea mucho más eficiente e intuitivo para el gestor ambiental.

1.1 Antecedentes

Es importante señalar que uno de los principales problemas que existen, en casi todas las ciudades, incluyendo la nuestra (Cuenca-Ecuador), es la contaminación ambiental, debido a muchos diversos motivos. Por tal razón, es de vital importancia realizar estudios que ayuden al análisis del comportamiento de los factores que intervienen en este problema (Matute Rivera, 2018).

La Empresa Municipal de Movilidad Tránsito y Trasporte de la ciudad de Cuenca (EMOV EP), actualmente tiene una estación de monitoreo continuo, la cual consta de sensores ambientales para la medición de agentes contaminantes atmosféricos, tomando registros de monitoreo de la calidad del aire desde el año 2008, y generando información fiable (EMOV EP, 2017; Sellers, 2013).

De tal manera, y considerando, que la contaminación ambiental es un conflicto real de suma importancia en nuestra ciudad, el Departamento de Investigación del Instituto de Estudios de Régimen Seccional del Ecuador (IERSE) de la Universidad del Azuay (UDA), inició un proyecto específico de seguimiento y evolución, para monitorear la calidad del aire en la ciudad de Cuenca (Sellers, 2013).

En una primera fase de este proyecto, se realizó un proceso de recolección de datos relacionados con los niveles de contaminación del aire, considerando los factores atmosféricos; determinando la relación que existe entre los factores de contaminación atmosférica (Cabrera Lituma, 2017; Ortega Guamán, 2018).

En una segunda fase, los resultados que obtuvo Cabrera Lituma (2017), se presentan por medio de dos sitios web para el monitoreo y publicación, con normativas, Environmental Protection Agency (EPA) y Texto Unificado de Legislación Ambiental Secundaria (TULSMA) de los contaminantes atmosféricos y variables meteorológicas, donde se observa gráficas y tablas que muestran la relación entre dichos contamintantes. El propósito principal fue establecer un medio de comunicación para la recolección automatizada de la información colectada por la estación de monitoreo de la EMOV-EP.

También, Andrade (2018) analizó el aspecto relacional entre las dimensiones atmosféricas y las variables meteorológicas, específicamente aplicó una metodología flexible y técnicas de minería de datos. Los datos utilizados correspondieron al periodo de tiempo de un mes, además las muestras extraídas, fueron mediciones realizadas cada diez minutos, y presentó sus resultados a través de gráficas estadísticas.

El IERSE, ha considerado analizar, tratar y gestionar un margen de variables de contaminantes atmosféricos y variables meteorológicas dentro de sus proyectos de investigación, comprobando la disposición de datos recolectados a lo largo de algunos años por la unidad de monitoreo de la EMOV-EP. Tomando en cuenta que los mismos varían constantemente, pero que son de mayor importancia para el gestor ambiental.

Las variables que intervienen en el análisis de contaminantes ambientales son: Ozono (O3), Monóxido de Carbono (CO), Dióxido de nitrógeno (NO2), Dióxido de Azufre (SO2), Material Particulado (PM2.5). Se generan grandes volúmenes de datos que se almacenan en un determinado tiempo (EMOV EP, 2017; Sellers, 2013).

De igual manera, las variables meteorológicas que intervienen son: la Temperatura del Aire (TEMPAIRE), la Presión Atmosférica (PATM), la Radiación Global (RADGLOBAL), la Radiación UVA, la Radiación UVE, la Precipitación (PRECIP_SUM), la Humedad Relativa (HR), el Punto del rocío (DP), la Velocidad del Viento (WINDSPEED), y la Dirección del Viento (WINDDIR) (EMOV EP, 2017).

1.2 Objetivos

1.2.1 Objetivo General

Aplicar técnicas de minería de datos basadas en la metodología CRISP-DM sobre los contaminantes atmosféricos y las variables meteorológicas, de un año de recolección de datos, y analizar e implementar el método más adecuado en la visualización que permitan la toma de decisiones de los gestores ambientales.

1.2.2 Objetivos Específicos

- Identificar las variables de estudio, de acuerdo al objetivo general del presente trabajo de titulación, considerando el comportamiento de las variables en la ciudad.
- Elaborar un marco teórico referente a las técnicas de minería de datos y técnicas visuales interactivas, aplicables al problema, basados en la metodología CRISP-DM.
- Identificar los patrones relevantes que determinen la relación entre las variables de contaminación atmosférica y variables meteorológicas.
- Visualizar grandes volúmenes de resultados de la aplicación de la minería de datos, a través de una técnica de visualización multidimensional dinámica e interactiva, generando conocimiento que permita la toma de decisiones del gestor ambiental.

1.3 Preguntas de Investigación

¿Cómo aplicar técnicas de minerías de datos y visualización de grandes volúmenes de datos relacionados a variables atmosféricas y meteorológicas para la toma de decisiones de gestión ambiental?

1.4 Justificación

Los trabajos e investigaciones que el departamento de investigación IERSE ha venido desarrollando, han presentado evoluciones dentro de sus análisis, metodologías de tratamiento y presentación de resultados, lo que ha sido de gran utilidad para la gestión del "Índice de Calidad del Aire" (ICA).

Si bien se ha recolectado la información, han existido limitaciones relacionadas con la publicación de la información capturada. Esta información, obtenida por los sensores de la unidad de monitoreo de la EMOV-EP, anteriormente se la exponía únicamente medios impresos, lo que limitaba su interpretación en tiempo real y un análisis temporal. Por lo que se desarrolló, un mecanismo para explotar la información, a través del tratamiento de datos y difusión de la información en la web, donde se presentó resultados en medios de fácil acceso y comprensión (Sellers, 2013).

Con el pasar del tiempo se ha condensado una gran cantidad de información, se ha planteado desarrollos a través del uso de metodologías y técnicas de minería de datos, y se ha extraído patrones de comportamiento de contaminantes (Ortega Guamán, 2018).

El avance tecnológico y la gran acogida de los medios digitales, han previsto proporcionar información útil en distintos ámbitos, como los negocios, la política, la medicina, el ambiente, etc., de tal forma se sugiere, un proceso eficiente de tratamiento y visualización de los datos, para la toma de decisiones. La información registrada del IERSE es almacenada en un repositorio de datos tipo texto (CSV y XLSX), por el momento, no existe un procedimiento altamente visual, con gráficas dinámicas, en donde se pueda interactuar con los datos capturados y los análisis que pretenda realizar el gestor, adicionalmente que se encuentren disponible en la web.

Debido a esto, dentro del presente trabajo se realizó procesos de análisis a través software especializado para minería y visualización de Big Data, aplicando metodologías y técnicas visuales, que facilitaron la extracción de patrones de interés, de conjuntos de datos, por medio de gráficos interactivos que permitan analizar, tratar e interactuar con estas grandes cantidades de información (Castro, Luján, &Martig, 2006).

1.5 Alcance

En la tercera fase propuesta, con la información recopilada por el IERSE de los contaminantes atmosféricos y variables meteorológicas dentro de la zona de estudio de Cuenca, se aplican procesos de minería de datos, empleando la metodología CRISP-DM, que permite el tratamiento del conjunto de datos recolectado, aplicando los principales algoritmos de clustering[1] (X-Means, K-Means, DBSCAN) para su evaluación y comparativa de agrupamiento, dando como resultado patrones de relación y asociación que puedan ser visualizados por medio de herramientas dinámicas, incluyendo grandes volúmenes de datos, específicamente la colección de un año referente al 2018; permitiendo la toma de decisiones para el gestor ambiental.

1.6 Resultados Esperados

Los resultados que se pueden extraer de los análisis de grandes volúmenes de información, implican la toma de decisiones. En general, muchas empresas ven el futuro de los análisis de Big Data a través de la visualización (VIZ), y están desarrollando nuevas plataformas interactivas que apoyan la investigación en este campo, para facilitar la interpretación del usuario (Furht & Villanustre, 2016). Se espera obtener una gráfica de dinámica desarrollada en una aplicación web, que indican las asociaciones de los contaminantes atmosféricos y las variables meteorológicas, que finalmente puede aportar a la toma de decisiones.

Las técnicas de visualización avanzadas y la aplicación de herramientas dinámicas a la vanguardia(D3.js), han sido aplicadas a muchos problemas de la representación visual ambiental, se puede obtener un alto grado de éxito en el procesamiento e interpretación de los resultados con respecto a los criterios visuales (Gupta, Singh, & Singh, 2017).

Culminado este trabajo de titulación, la información resultante se sintetizará de forma generalizada, publicándose en una aplicación web, que contendrá un gráfico multidimensional, gestionado por la herramienta D3.js, que permite un análisis altamente visual, en donde se puede interactuar con los datos de manera sencilla y dinámica, con determinaciones: colorimétricas, selecciones, agrupamientos, detalle de registro de dato, pincelados, siendo posible minar datos desde el propio gráfico, y descartar dimensiones directamente.

[1] Clustering: agrupamiento de datos por medio de un algoritmo, técnica o método.

2. REVISIÓN DE LITERATURA

A continuación, se estudiarán los conceptos fundamentales que se relacionan al desarrollo de proyectos de minería de datos en cuanto a su organización, la que se esquematiza dentro de un marco teórico para la comprensión del trabajo de titulación desarrollado.

2.1 Marco Teórico

Dentro de este marco teórico, se detallan definiciones que han sido necesarias mencionar, para un adecuado desarrollo del estudio. En consecuencia, se obtiene una visión general, que permite al usuario gestor familiarizarse con los temas y los elementos involucrados en el desarrollo de la aplicación experimental del presente trabajo de titulación.

2.1.1 Conceptos Básicos Generales

Al desarrollar análisis de conjuntos de datos con un amplio contenido de registros de información, el usuario se relaciona con el campo de Big Data, el cual hoy en día se encuentra en auge dentro de las grandes negocios y organizaciones, permitiendo obtener grandes beneficios con los múltiples enfoques de tratamiento de datos. En este punto, la minería de datos, se involucra directamente a Big Data, para realizar procesos de tratamiento, mediante técnicas y metodologías que permitan encontrar patrones ocultos en los volúmenes de datos, resultados que permiten la toma de decisiones y control de procesos con eficacia.

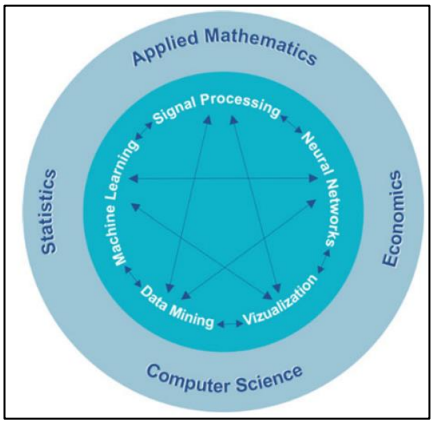
2.1.1.1 Big Data

Big Data, es una tecnología que se caracteriza por aspectos de importancia que influyen al manejo de los datos, a través de un modelo específico, por ejemplo: gran velocidad, diversidad de preparación de datos, resultados valiosos y veraces, etc. (Demchenko, De Laat, & Membrey, 2014).

Dentro de Big Data, se necesita de métodos y técnicas de distintas ramas de conocimiento y estudio, lo que permite que sea fácil crear estructuras y relaciones entre las dichas variables. Esto determina en consecuencia, que los procesos dentro de Big Data están en constante evolución.

Figura 1

Métodos de procesamiento de Big Data



Fuente: (Furht & Villanustre, 2016)

En la figura 1, los métodos de procesamiento son: la minería de datos, el aprendizaje automático, el procesamiento de señales, las redes neuronales, y la visualización, los cuales tienen una fuerte relación, es decir, interactúan y se usan simultáneamente durante el procesamiento de datos. Esto aumenta la eficacia, por su estructura estrella, en donde la combinación de ciertos métodos potencializa resultados precisos y acertados. Las ciencias en la parte exterior, (matemáticas aplicadas, las estadísticas, la economía y la informática) son la base estructural de los métodos de procesamiento de los datos, por lo que se permite una orientación definida (Furht & Villanustre, 2016).

Según (Gandomi & Haider, 2015) aclaró que, las empresas, así como las organizaciones públicas y privadas, requieren metodologías eficientes, con el objetivo de transformar los grandes volúmenes de datos en aspectos significativos, que generen un beneficio, enfocó la necesidad de métodos analíticos y nuevas herramientas predictivas, que permitan explotar los grandes almacenes de datos heterogéneos no estructurados, abriendo un campo muy extenso para diversas aplicaciones innovadoras de gestión. A través de la minería de datos se obtiene patrones significativos de los datos ocultos que permiten obtener beneficios significativos para la toma de decisiones.

2.1.1.2 Minería de Datos (Data Mining-DM)

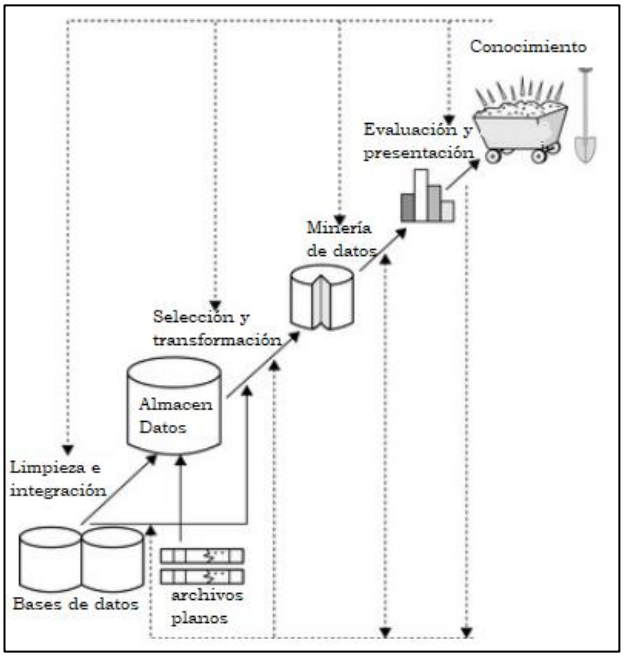
La evolución natural de la tecnología de la información, encamina a la minería de los datos como una ciencia que facilita descubrir conocimiento, las grandes cantidades de datos que se almacenan a cada minuto alrededor del mundo, permiten la necesidad de análisis de los mismos, de donde, se puede extraer datos valiosos para obtener hallazgos por medio de herramientas informáticas (Jiawei, Micheline, & Jian, 2012).

La minería de datos, desempeña un rol diferencial, ya que aplica algoritmos específicos para extraer patrones de los datos, un subproceso particular del proceso general Knowledge Discovery in Databases (KDD), transforma conjuntos de datos en conocimiento, lo que permite resolver problemáticas que se presentan varios aspectos de gran interés, es decir, es la búsqueda de conocimiento, a través de procesos y métodos inteligentes, para extraer patrones de comportamiento en la información, identificando, los más relevantes, y publicando el conocimiento tratado con técnicas de visualización (Jiawei et al., 2012).

Los algoritmos de proceso de la minería de datos se fundamentan en la inteligencia artificial y el análisis estadístico, que solucione problemáticas por medio de la construcción de modelos que permitan realizar predicciones, segmentaciones, etc. (Beltran, Poveda, Bolívar, Corrales, & Sarmiento, 2010). Existen diferentes técnicas como: sistemas de agrupamiento, reglas de asociación, correlaciones, clasificación, regresión. El enfoque para seleccionar la mejor técnica, debe ser claro, pues existe eficacia cuando se conoce que tipo de información se está procesando y extrayendo, que cuando se conoce que buscar en concreto dentro de un grupo de datos, así también el tipo de análisis de que realice:

Figura 2

Procesos de minería de datos para descubrir conocimiento.



Fuente: (Jiawei et al., 2012)

En la figura 2, se observan los procesos básicos de la minería de datos, en donde se realiza un preprocesamiento de la información, que permite integrar los datos, los mismos que pueden ser de diferentes fuentes, este resultado parcial, se aloja en un nuevo almacén de datos, para la posterior selección de las variables relevantes para el análisis según la necesidad. Luego, se transforman en tipos de datos de interés que permite su procesamiento. A continuación, se aplica un método inteligente para la extracción de patrones de estos datos, después, se evalúa los esquemas de comportamiento y se extrae el conocimiento de utilidad, para que se presente dentro de un material visual de fácil entendimiento.

Cabe recalcar que, para la aplicación de la minería de datos y el desarrollo de proyectos, existen varias metodologías que cumplen con estos procesos, y se aplican con diferentes herramientas informáticas disponibles en el medio actualmente.

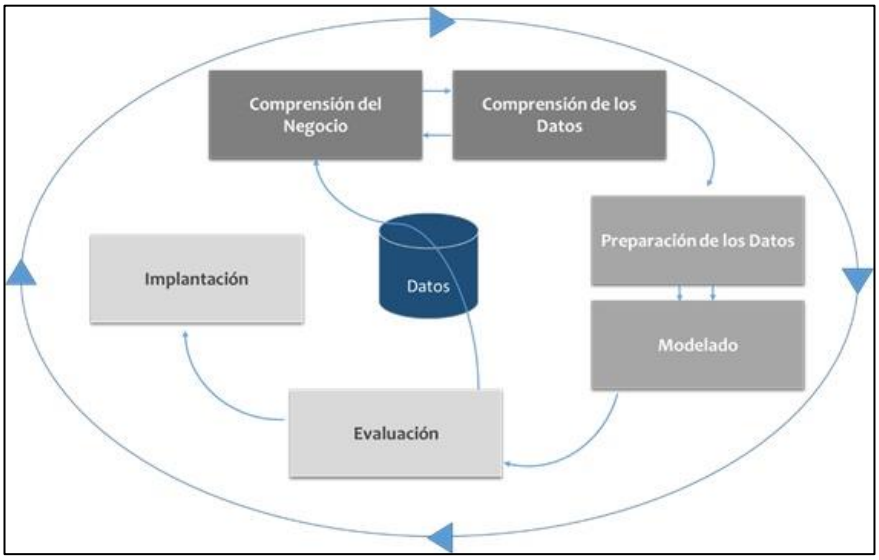
Se ha elegido la metodología CRISP-DM, ya que dicha metodología ha sido usada en proyectos de investigación antecesores, con el propósito de implementar una comparativa y un enfoque visual. Además, esta metodología, facilita el desarrollo adaptándose de una forma eficiente a los objetivos propuestos dentro del desarrollo del presente trabajo de titulación, en resumen, esta metodología responde de manera satisfactoria los propósitos del estudio experimental que se realiza conforme al enfoque de investigación.

2.1.1.3 Metodología CRISP-DM

La metodología del Proceso estándar de la industria para la minería de datos, Cross Industry Standard Process for Data Mining (CRISP-DM), es generalmente muy utilizada para todo tipo de desarrollo de proyectos de minería de datos. Es una metodología estructurada que asegura el rendimiento del proyecto, flexible en sus etapas, lo que permite su personalización fácilmente, es decir, esta metodología puede adaptarse a cada proyecto de estudio, realizar informes en cualquier momento, crear un modelo de minería de datos con base en necesidades concretas (IBM, 2012).

Figura 3

Modelo de proceso de la Metodología CRISP–DM



Fuente: (IBM, 2012)

Orientación de modelo de proceso, determina síntesis del ciclo de vida de minería.

Según kdnuggets.com (2007), como se citó en (Agila-Palacios, Ramirez-Montoya, García-Valcárcel, & Samaniego-Franco, 2017) CRISP-DM se convirtió en la metodología más utilizada en proyectos de minería de datos, justificando su aplicación por su gran eficacia.

CRISP-DM está distribuida en cuatro niveles de abstracción jerárquica, y cuando se enfoca como metodología, describe etapas, tareas necesarias y la ilustración de las relaciones entre éstas. Las etapas que se observan en la figura 2 son: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación (Gallardo Arancibia, 2013).

2.1.1.3.1 Comprensión del Negocio

La etapa de comprensión del negocio, es de gran importancia dentro de la metodología CRISP-DM, principalmente la comprensión de los objetivos y requisitos del negocio, permite entender completamente el problema que se desea resolver, colectar datos correctos e interpretar los resultados de forma correcta (Gallardo Arancibia, 2013).

Dentro de las principales tareas que se realizan en esta etapa, tenemos: determinar los objetivos del negocio, valoración actual (pre proceso), objetivos DM, y desarrollar plan de proyecto.

2.1.1.3.2 Comprensión de los Datos

Esta etapa, se refiere a conocer más a detalle las estructuras disponibles, acceder a los datos y explorarlos, con la ayuda de tablas, diccionarios y esquemas, de manera que la calidad de los datos para el proceso de minería sea efectiva. El desarrollo de esta comprensión es crucial para evitar problemas en la siguiente etapa y representar los resultados correctamente (IBM, 2012).

Según esta etapa, se tiene las siguientes tareas programadas: recolección de datos iniciales, descripción, exploración de estructuras y verificación de la calidad de los datos.

2.1.1.3.3 Preparación de los Datos

La etapa más importante y que más tiempo exige en la minería de datos. Preparar y empaquetar los datos, conlleva hasta un 70% de tiempo y esfuerzo de todo el proyecto. Esta se encuentra directamente relacionada con la etapa de modelado, y que dependiendo de la técnica de modelado, los datos se procesan de distintas maneras (IBM, 2012). En la figura 4, se detalla las tareas que intervienen dentro de esta etapa.

Figura 4

Tareas de la etapa preparación de los datos

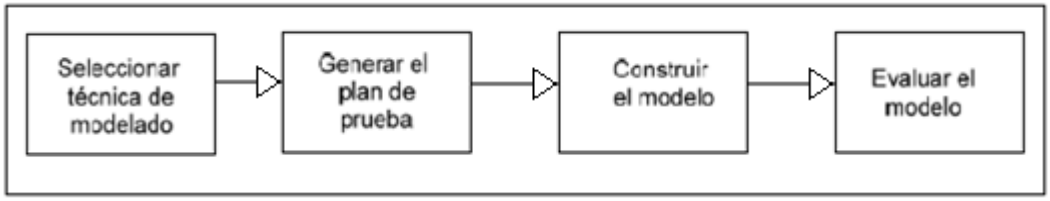
2.1.1.3.3.1 Normalización de los Datos

La normalización de los datos, hace referencia a los ajustes necesarios de las escalas de los datos, con respecto a una escala común, debido a que la dispersión de los rangos, por las diferentes unidades de medida de los datos. El propósito es obtener la comparación de los valores en conjuntos de datos de manera que los datos se distribuyan de manera uniforme para determinar comportamientos y no se dispersen de una manera generalizada.

2.1.1.3.4 Modelado

En esta etapa de CRISP-DM, se seleccionan la técnica de modelado considerando criterios para la correcta aplicación de la minería de datos. Es importante definir y construir el modelo tomando en consideración los tipos de datos disponibles, los objetivos de la minería, y los requisitos específicos del modelado. Las tareas que intervienen en esta etapa son: seleccionar técnica de modelado, generar plan de prueba, construcción y evaluación del modelo (Gallardo Arancibia, 2013).

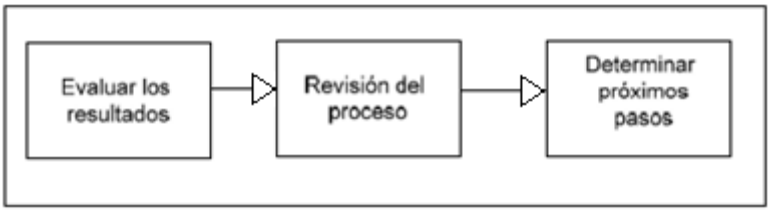
Figura 5
Tareas de la etapa de modelado



2.1.1.3.5 Evaluación

Esta etapa, permite revisar el modelado, que se desarrolló en la etapa anterior, de tal forma que se acate a los criterios satisfactorios de rendimiento del problema del negocio, esto se realiza de manera concreta sobre los datos específicos que se aplicó el análisis, aquí se puede encontrar errores, pudiendo corregir estos, considerando un tiempo para retroalimentar los aciertos y errores. Cuando ya el modelo construido es válido, y cumple con los objetivos de éxito, se explota el modelo y se emplea herramientas de interpretación de resultados (Gallardo Arancibia, 2013).

Figura 6
Tareas de la etapa evaluación



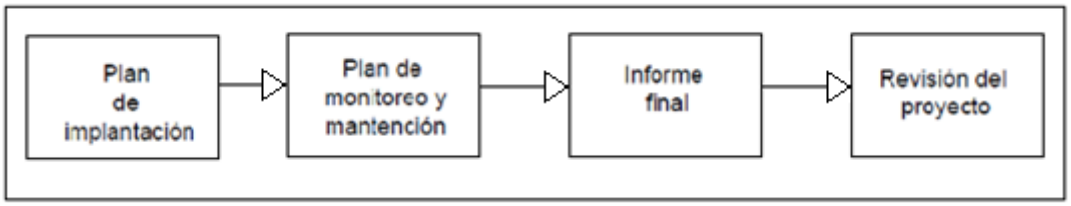
En la figura 6, se describe las tareas involucradas en esta etapa, donde se realiza una valoración de resultados, una calificación y mejoras del proceso de DM y plantear un retroceso de fases, si los resultados no son acordes a los propósitos del análisis propuesto.

2.1.1.3.6 Implantación

En esta etapa, el conocimiento que se obtuvo de la creación y validación del modelo, mismo que se convierte en operaciones de acción, dentro de la gestión del objetivo del negocio, basadas en las observaciones del modelo o aplicando el modelo a diferentes conjuntos de datos para incrementar conocimiento de la data, generando un informe final que garantice su eficacia y se permita realizar mejoras continuas de la experimentación a lo largo de la metodología (IBM, 2012).

Figura 7

Tareas de la etapa implantación



2.1.2 Herramientas Software

Las herramientas que se van a utilizar para el desarrollo del presente trabajo, se estructuran con base en aspectos como:

- Modelamiento computacional
- Visualización

El modelamiento computacional implica el uso de una herramienta que realice la limpieza de los datos, conocida comúnmente como técnicas de pre-procesamiento. Un aspecto fundamental de las herramientas de modelamiento es la aplicación de técnicas de minería de datos, para lo cual se utilizará la herramienta Rapidminer.

2.1.2.1 Rapidminer

Es una herramienta de software, desarrollado en Java[2], multiplataforma, extensible y con un gran liderazgo reconocido a nivel mundial, es característicamente de código abierto, que brinda utilidad para la minería de datos. Realiza sus análisis de datos, a través del encadenamiento de operadores hacia una pantalla gráfica, lo que facilita la interacción, y cuenta con varios ejecutores para entrada, tratamiento, procesamiento y visualización (Beltran et al., 2010).

[2] Java: Lenguaje de programación de propósito general orientado a objetos (Java, 2019).

Es una herramienta de fácil uso, con un entorno intuitivo y amigable con el usuario, carece de costo, para un determinado número de registros o para uso en roles investigativos y de aprendizaje.

El software Rapidminer se ejecuta en las plataformas operativas principales, e implementa algoritmos de preparación de datos y de aprendizaje automático entre otros, permitiendo maximizar su eficacia, según Kdnuggets (2015), se posicionó entre las dos primeras herramientas más utilizadas para la minería de datos. Además, permite incluir algoritmos de evaluación Waikato Environment for Knowledge Analysis (WEKA)[3], para dar soporte a procesos con grandes volúmenes de datos (Burget, Karásek, Smékal, Uher, & Dostál, 2014; Rapidminer, 2019).

La visualización de datos se fundamenta en el uso de lenguajes que presenten al usuario un resultado de fácil interpretación, para este propósito, se realiza la combinación de varios lenguajes y herramientas que permiten el desarrollo y la creación del resumen de resultados; incluyendo las herramientas tecnológicas que permiten la interacción con el conjunto de datos, cumpliendo el objetivo propuesto. El presente trabajo de titulación utiliza HTML5, Sublime Text y D3.js.

2.1.2.2 Lenguaje de Programación HTML 5

Es un lenguaje de marcado, Hyper Text Markup Language (HTML), se utiliza para el desarrollo de páginas de Internet. La importante acogida del uso de internet, permite que se obtengan medios para el desarrollo de un sitio web. Las posibilidades que ofrece la herramienta estándar HTML5, son favorables para crear páginas con etiquetas, semánticas, multimedia, funciones y eventos que facilitan la interactividad en los diseños web, World Wide Web (W3C)[4] trabaja conjuntamente en el desarrollo de estándares web que sirvan de guía a desarrolladores (Diez, Dominguez, Martinez, & Sáenz, 2012).

[3] WEKA: lenguaje de análisis del conocimiento de la Universidad de Waikato (University Waikato, 2019).
[4] W3C: Consorcio internacional que genera estándares (W3S).

2.1.2.3 Sublime Text 3

Sublime Text 3, es un editor de texto de código fuente con descripción visual con base en colorimetría, orientado para el desarrollo de código: HTML, JavaScript [5], PHP[6], etc. Es de gran rendimiento, y facilita la detección de errores, basados en los colores con una interacción partidaria para el desarrollador; se puede obtener de forma gratuita, pero de ser necesario el uso continuo, es obligatorio adquirir una licencia, esto según Sublime Text, (2017), consultado por (Ontaneda, 2017).

2.1.2.4 Documentos Basados en Datos (Data Driven Documents D3.js)

Herramienta conocida como D3.js, es una librería JavaScript para manipular o interactuar de forma eficiente documentos basados en datos, cumple enfáticamente los estándares web combinando componentes de visualización potentes, éstos, pueden adaptarse a los diferentes navegadores. Permite exaltar la visualización de los datos usando HTML (contenido de página), CSS[7] (estética), SVG (gráficos basados en vectores escalables) y basado en datos para la manipulación de la estructura jerárquica, Document Object Model (DOM), además, se puede reutilizar código, a través de diversos módulos oficiales de libre acceso. D3.js ofrece una flexibilidad extraordinaria, es considerablemente rápido y admite grandes conjuntos de datos y comportamientos dinámicos para la interacción y la animación (Bostock, Ogievetsky, & Heer, 2011).

La herramienta permite: establecer atributos, establecer estilos; registrar eventos; agregar, ordenar y eliminar; controlar el contenido de texto (“d3js.org”, 2019). La escalabilidad se refiere al dominio de la estructura, lo que optimiza recursos, cumpliendo los objetivos de: compatibilidad, depuración y actuación (Bostock et al., 2011).

SVG: es un formato de gráficos vectoriales bidimensionales, estáticos y animados, en formato XML[8], estándar W3C, proporciona un control exacto sobre el procedimiento de coordenadas donde se localizan los objetos gráficos, las transformaciones que se aplicarán durante el renderizado[9] (Ferraiolo, Jun, & Jackson, 2003).

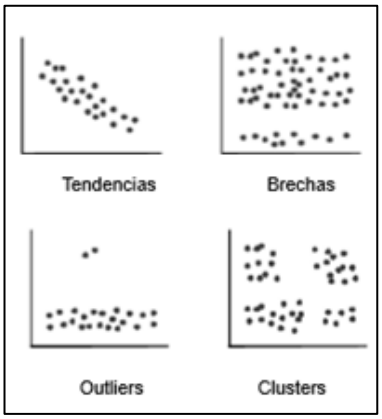
[5] JavaScript: Lenguaje de programación para crear interacciones en páginas web (JS).
[6] PHP: Lenguaje de programación independiente de plataforma, que realiza accesos a bases de datos (PHP).
[7] CSS: “Cascading Style Sheets”. Hojas de estilo en cascada.
[8] XML: Lenguaje de marcado extensible,
[9] Renderizado: proceso de generar una imagen visible a partir de información digital.

2.1.3 Técnicas de Visualización

Visualizar, se refiere a transmitir la información por medio de representaciones gráficas de los datos, formar una imagen de un concepto de manera que se pueda hacer posible por medio de una herramienta informática. Al usar determinaciones visuales de datos abstractos, compatibles con una computadora, se puede amplificar el conocimiento facilitando a los usuarios realizar tareas de manera más eficiente, es así, como la visualización hace su aporte significativo (Castro et al., 2006). Según (Zhao, 2013), “La visualización transforma la comprensión de los análisis, y permiten que sean mucho más eficientes".

Al existir grandes cantidades de datos, el conocimiento es limitado, inapreciable y se torna confuso; en consecuencia, las técnicas de visualización se plantean como principal objetivo, poder transmitir de manera simple, clara y efectiva los resultados por medio de métodos de representación visual. La visualización de datos, permite construir gráficos interactivos o esquemas dinámicos, que resultan muy interesantes para la manipulación del usuario, permitiendo gestionar los análisis de manera intuitiva y divertida (Jiawei et al., 2012).

Figura 8
Patrones a través del Análisis



Fuente: (Castro & Luján, 2017)

Al hacer énfasis de un gráfico dinámico, se describe en una aplicación interactiva, donde el gestor administrativo puede navegar en las cifras en tiempo real, con una presentación de los resultados de los datos amigable y por medio de una herramienta informática, en donde se puede especificar condicionantes directamente en las gráficas, haciendo de la interpretación de los datos una experiencia que muestra las especificidades e integra funcionalidades.

Otros propósitos de la visualización son; almacenamiento, que permite registrar los hallazgos obtenidos; comunicación, se presenta oportunamente los datos, permitiendo posteriormente; el análisis y exploración, lo que permite ser el soporte para el razonamiento o procesamiento y también, la revelación de patrones por medio de la percepción que permita destacar intereses; y la confirmación, que se obtiene la verificación de hipótesis planteadas (Castro & Luján, 2017).

2.1.3.1 Percepción de Visualización

El cerebro humano solo puede captar cierta proporción de información si se muestra dispersa. La única manera de ver los patrones es mirando con el sentido visual (Zhao, 2013). Al hablar de percepción, se refiere al procesamiento perceptual en general, existen tres etapas para el procesamiento de la información visual consecuente de un estímulo: el sistema visual humano, procesamiento preatentivo y percepción de patrones. Según (Castro & Luján-Ganuza, 2017):

El sistema visual humano: está comprendido por el ojo y una parte del cerebro que procesa las señales, de tal forma que transforman un aspecto óptico, en una percepción visual de una escena.

Procesamiento preatentivo: se refiere, que no existe necesidad de una atención focalizada, importante para el diseño y construcción de visualizaciones, determinando que puede percibirse inmediatamente, propiedades discriminatorias, y que puede prestarse como erróneas interpretaciones.

Percepción de patrones: procesos activos que diferencian estructuras y dividen el escenario visual en regiones, considerando aspectos: color, textura y patrones de movimiento.

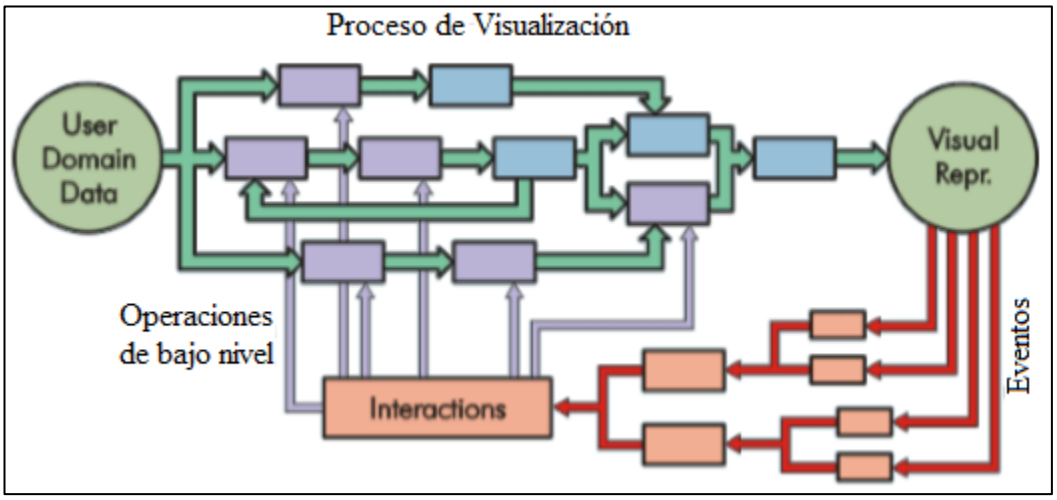
Dentro de un procesamiento de alto nivel, hace referencia a una percepción del espacio no simultáneo. El espacio 3D se enfoca comúnmente, a criterios para la percepción de la profundidad, por ejemplo se puede listar algunos importantes: perspectiva lineal, gradiente de textura, gradiente de tamaño, oclusión, profundidad de foco, sombras, convergencia, etc.(Castro & Luján-Ganuza, 2017f).

2.1.3.2 Proceso y Diseño de Visualización

Para que el proceso de la visualización sea correcto, se debe considerar algunos aspectos: en primer lugar y como lo más importante, se determinará el alcance visual y hacia quien se dirige la interpretación visual, segundo, se definen las tareas del proceso y finalmente se define el set de datos que interviene, todo basado en cumplir el objetivo que se planteó.

Dentro del proceso de desarrollo, se establece las tareas que intervendrán en la codificación, con base en una esquematización, interacción y soporte. Esto permite una evaluación de satisfacción visual, verificando si el resultado no cumple con las expectativas (Castro & Luján-Ganuza, 2017d).

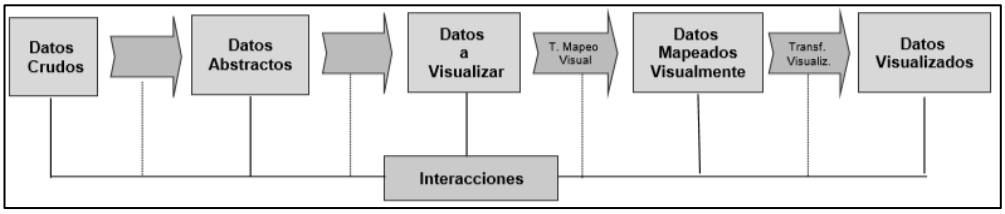
Figura 9
Proceso de Visualización



Fuente: (Castro & Luján-Ganuza, 2017d)

En la figura 9, se esquematiza los estados, transformaciones, acciones y operaciones. Los eventos son tareas de interacción visual, que varían de cierta manera en el proceso.

Figura 10
Proceso de Visualización - Diagrama Bloques



Fuente: (Castro & Luján-Ganuza, 2017d)

En la figura 10, por medio de interacción, los datos mapeados[10], se pueden mostrar mediante una técnica de visualización, la misma que es altamente expresiva. Esta muestra atributos del conjunto de datos; con efectividad, siendo los atributos más relevantes, los de mayor importancia, los que son más notables (Castro & Luján-Ganuza, 2017d).

La estructura visual se refiere al uso del espacio, señala la importancia de la ubicación es dominante ante la percepción. Dentro de esta estructura (Castro & Luján-Ganuza, 2017d) señala:

Sustrato Espacial: elementales tipos de ejes: no estructurados que son los que no disponen de ejes; nominales, una zona es clasificada en subzonas; ordinales, clasifica subzonas con importancia a un orden; y cuantitativos, ejes que disponen de una métrica. Los ejes pueden ser de forma: ortogonal, paralela, radial o libre.

Marcas: elementos visuales gráficos, visibles en el espacio, por ejemplo: puntos, líneas, áreas, volúmenes, y elementos n-dimensionales.

Canales Visuales: atributos de los elementos visuales, estos controlan la apariencia de las marcas, por ejemplo: posición, color, tamaño, orientación, forma y textura.

2.1.3.3 Clasificación de Visualización

La clasificación se la realiza considerando algunos aspectos, por ejemplo: objetivo de la visualización, los tipos de las variables, los mapeos de las variables, etc.

VyGLab (2015), determinó una clasificación, en consideración, a los datos que presentan de forma multivariada[11]: Técnicas basadas en puntos; Múltiples pantallas; Escalamiento Multidimensional (MDS), Técnicas basadas en la fuerza; y Técnicas basadas en líneas.

Otras clasificaciones se describen, considerando el estilo de sus gráficas y dimensiones, por lo que, dependerá estrictamente de los datos y los objetivos del problema, para la elección de la gráfica óptima de la interpretación, considerando incluso, datos multivariados.

[10] Mapeados: conjunto de elementos de un mismo tipo que tienen una distribución espacial determinada.
[11] Multivariados (as): comprende más de tres dimensiones (Multidimensional).

Cabe mencionar, que existen muchas variantes de tipos de gráficos, que están contenidos en las diferentes orientaciones. Según Jiawei et al. (2012), clasificó estas técnicas de visualización como:

2.1.3.3.1 Técnicas orientadas a píxeles

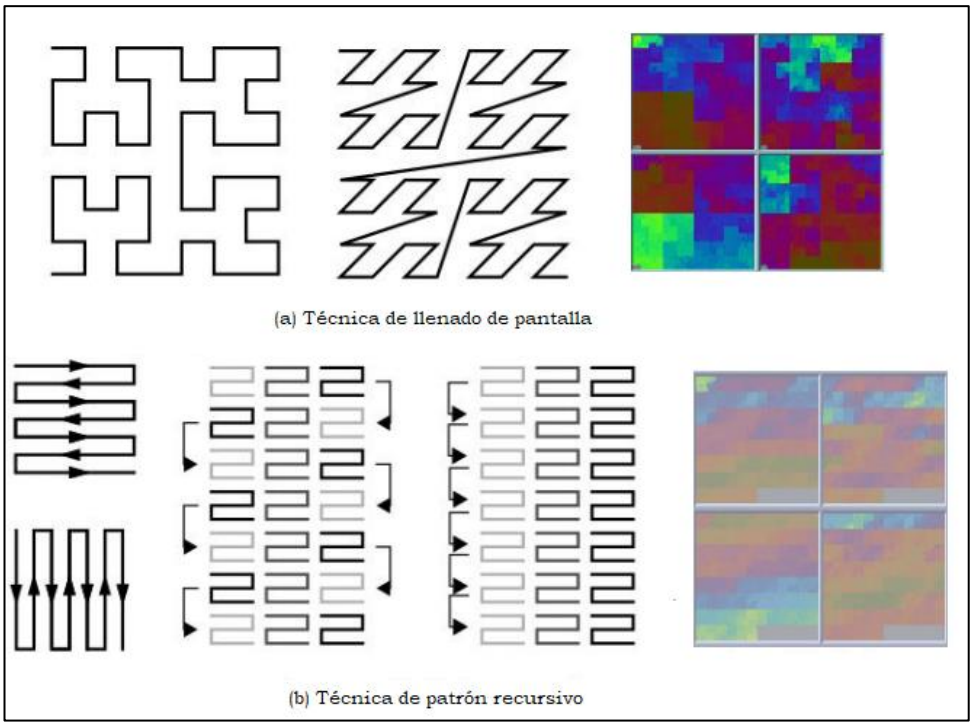
En estas técnicas intervienen datos de "k" dimensiones, que se representan, como "k" píxeles[12] coloreados en "k" ventanas en la pantalla. El color del píxel, es la representación de su valor, esto en consideración de las características de los datos y las tareas de visualización (D. Y. Liu, 2012).

Dentro de esta orientación de las técnicas de visualización, se dividen en dos grupos que consideran el porcentaje de datos:

Técnicas independientes de consulta: visualiza todo el conjunto de datos. Dentro de estas técnicas existen diversos gráficos. Por ejemplo:

- Técnica de llenado de pantalla.
- Técnica de patrón recursivo.

Figura 11
Técnicas orientadas a píxeles independientes de consulta



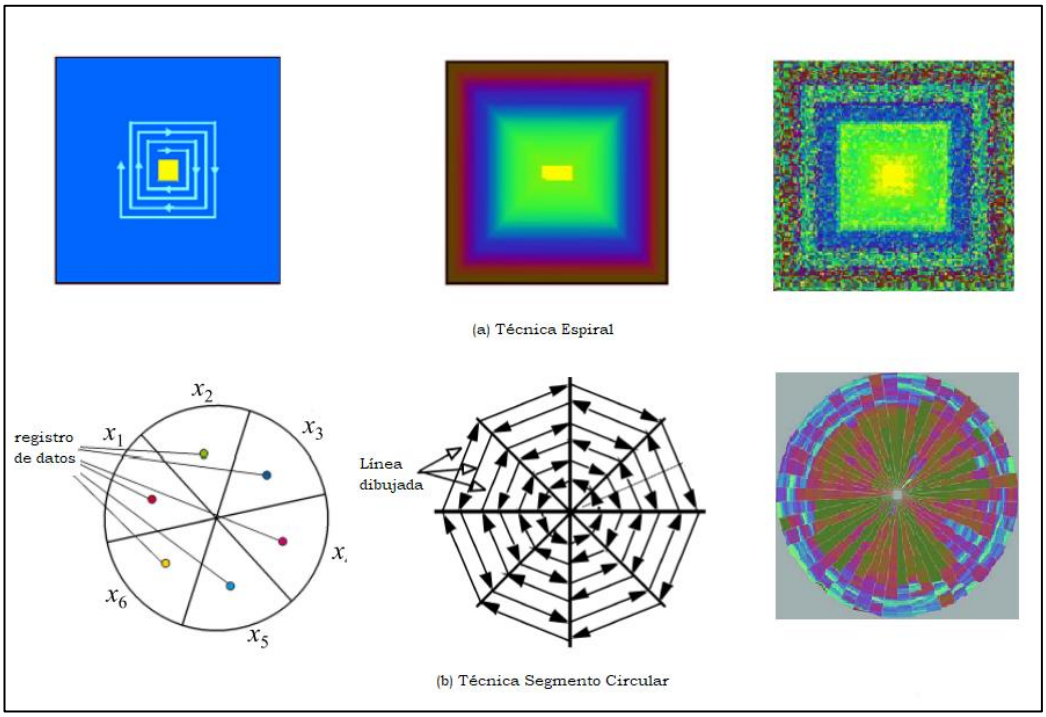
Fuente: (VyGLab, 2015)

[12] Píxel: Unidad básica de imagen digitalizada en pantalla a base de puntos de color (DefinicionABC).

Técnicas dependientes de consulta: visualiza una porción de datos, con mayor importancia para el consultante respecto del contexto del problema, es decir una consulta específica. De estas técnicas existen diversos gráficos. Por ejemplo: Técnica espiral; Segmento circular; Pantallas densas de píxeles; y Gráficos de barras de píxeles.

Figura 12

Técnicas orientadas a píxeles dependientes de consulta

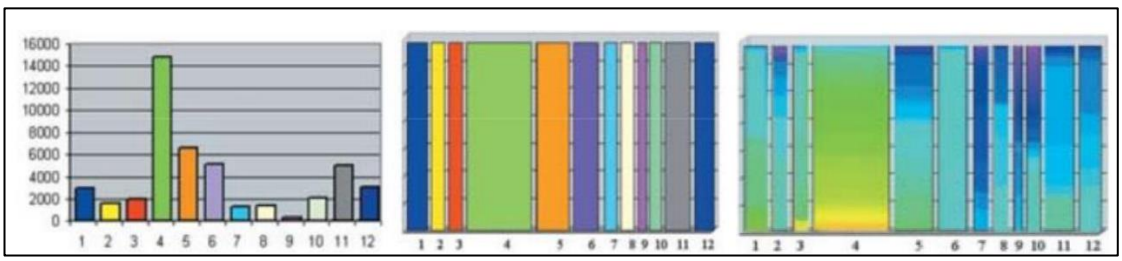


Fuente: (Y. Liu, 2010)

Como mencionó Y. Liu (2010), esta técnica maneja de forma considerable, conjuntos de datos de mediana y alta dimensión, cada registro de datos se lo asigna a un píxel, evitando superposición y desorden visual, pero esta orientación es inconsistente para valores y relaciones cuantitativas.

Figura 13

Técnicas orientadas a píxeles - Gráfica de barras en píxeles



Fuente: (VyGLab, 2015)

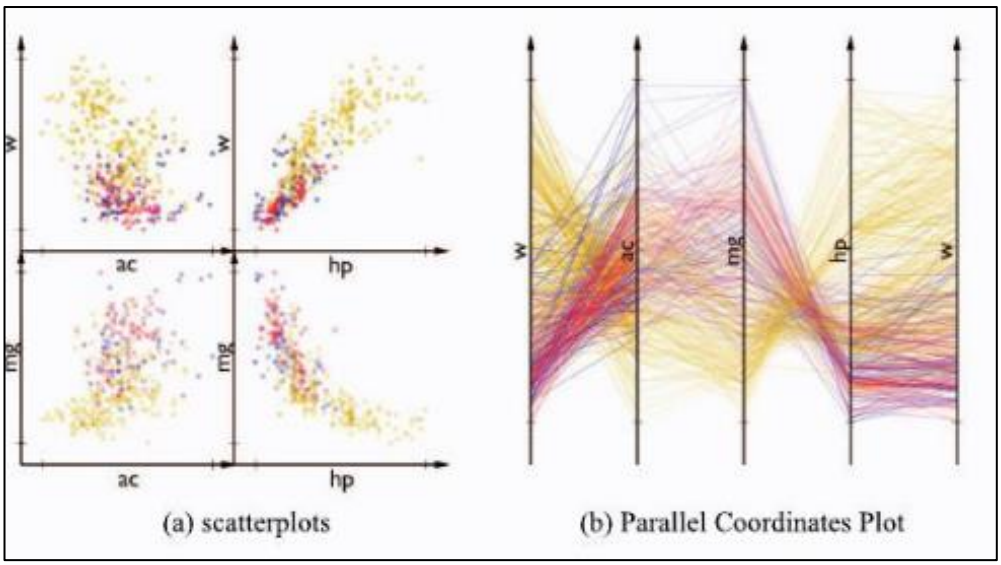
2.1.3.3.2 Técnicas de proyección geométrica

Orientada a transformaciones geométricas y proyecciones de los datos, para los registros de datos se asignan a puntos multidimensionales, de manera que permite visualizar un espacio de alta dimensión en una pantalla 2D. En esta sección, se puede listar los gráficos de este enfoque: Gráfico de dispersión (Scatterplot); Matriz de gráficos de dispersión (Scatterplot Matrix); RadViz; Pantalla enrejada (Trellis Display); y Coordenadas Paralelas (Parallel Coordinates).

De manera general, esta técnica altamente efectiva, muy eficiente para detectar valores outliers[13] y la correlación entre determinadas variables, además, permite trabajar conjuntos de datos grandes, y con la interacción apropiada, soporta grandes volúmenes.

Sin embargo, el desorden visual conjuntamente con la superposición, aumenta de manera crítica al procesar conjuntos de datos grandes (Y. Liu, 2010), una desorganización de los ejes, puede afectar el resultado su percepción e interpretación.

Figura 14
Técnicas orientadas proyección geométrica



Fuente: (Claessen & Van Wijk, 2011)

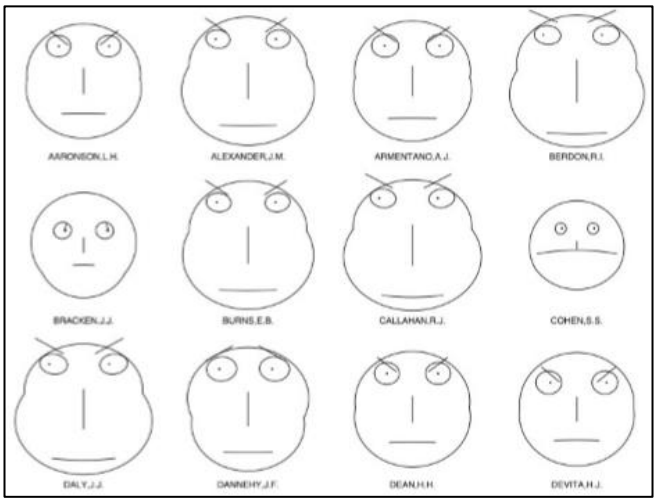
[13] Outliers: Un valor más extremo, sobre rangos normales, no habitual.

2.1.3.3.3 Técnicas basadas en iconos

Representa valores multidimensionales, por medio de iconos, con el objeto de manifestar descripciones los datos, es decir realiza una visualización de los valores como características de iconos. En esta sección, se puede listar los gráficos de este enfoque: Caras de Chernoff; Stick figures; Star plots; Color icons, etc. (Jiawei et al., 2012).

Figura 15

Técnicas basadas en iconos - Caras de Chernoff



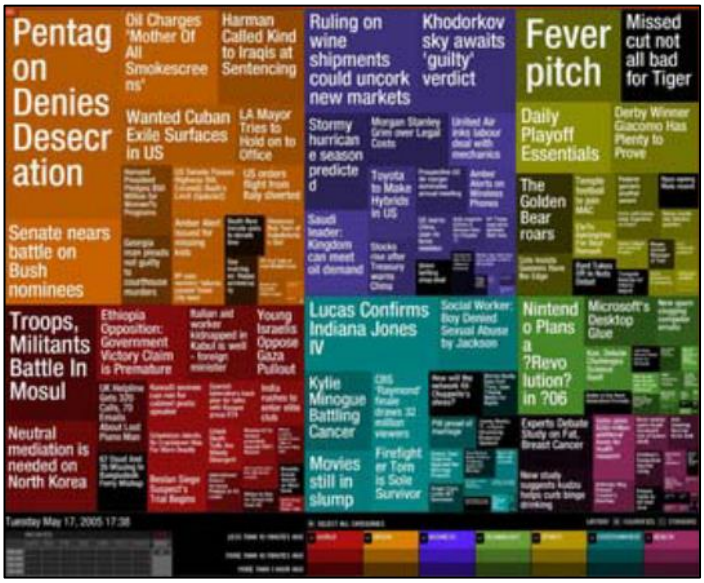
Fuente: (Castro & Luján-Ganuza, 2017b)

Manejan conjuntos de datos pequeños a medianos de alta dimensionalidad, debido a que los íconos ocupan espacio en pantalla de diversos píxeles, algunas características visuales de los íconos pueden atraer más atención, las variables de datos se asignan a las características de los iconos lo que determina la expresividad de la visualización resultante (Y. Liu, 2010).

2.1.3.3.4 Técnicas jerárquicas y gráficas

Permite visualizar múltiples dimensiones simultáneamente, soporta y gestiona conjuntos de datos de pequeña a mediana proporción, trato diferencial de asignaciones que producen diferentes vistas de datos, es el óptimo para manejo de dimensionalidad baja a media. La interpretación resultante tiene complejidad. Dentro de esta sección, se puede listar los gráficos de este enfoque: Apilamiento dimensional (Dimensional stacking); Mosaic Plot; Mapas de árboles (Treemap), etc. (D. Y. Liu, 2012).

Figura 16
Técnicas jerárquicas y gráficas - Mapa de Árbol



Fuente: (Castro & Luján-Ganuza, 2017b)

NewsMap permite visualizar noticias por extensión geográfica, de forma jerárquica a las noticias más populares.

2.1.3.3.5 Visualización de datos complejos y relaciones

Técnica novedosa de visualización de datos de texto, generada por la evolución y crecimiento de datos de tipo no-numéricos, se usa para representar etiquetas o metadatos[14] en sitios web, existe un gran interés, por analizar y gestionar estos datos, expresando la importancia de estos, con el tamaño o el color de los resultados del tratamiento. Dentro de esta técnica existen gráficos como: Tag clouds (Jiawei et al., 2012).

Figura 17
Visualización de datos complejos - Tag clouds

Fuente: (Hassan-Montero & Herrero-Solana, 2006)

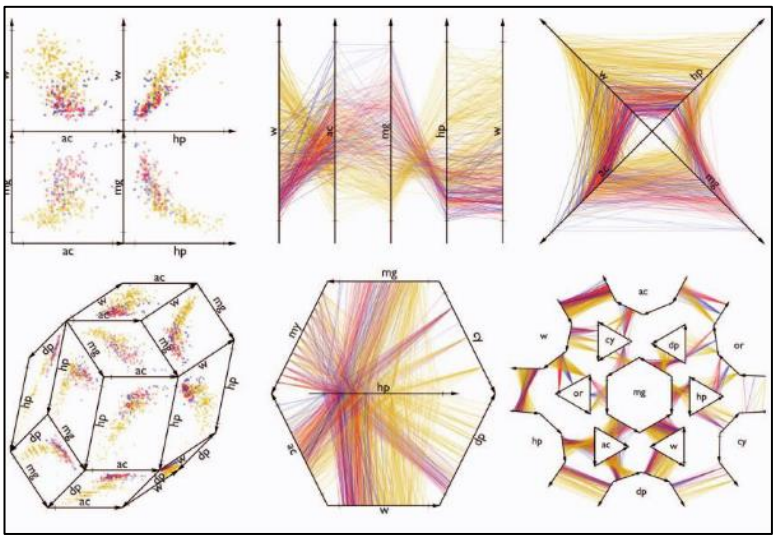
Etiquetas seleccionadas y visualmente ponderadas según su frecuencia de uso.

[14] Metadato: información que caracteriza datos, describen el contenido.

2.1.3.3.6 Técnicas híbridas

Integra múltiples técnicas de visualización, pudiendo expresarla en múltiples ventanas, para mejorar el entendimiento general de la técnica visual. Dentro de esta técnica se encuentra gráficos como: matrices de Scatterplots de vinculación y cepillado, SPLOM[15] Coordenadas Paralelas. La vinculación y el cepillado son herramientas muy eficaces para integrar ventanas de visualización (Y. Liu, 2010).

Figura 18
Técnicas Híbridas - SPLOM Coordenadas Paralelas



Fuente: (Claessen & Van Wijk, 2011)

Con base a la revisión de la literatura, según (Kandogan, Road, & Jose, 2000; Kandogan, 2001) como se citó en (Gui-Fen, Guo-Wei, Li M., Li Y., Jian & Hang, 2009), mencionó la afectividad de la aplicación de diferentes técnicas de visualización, indistintamente de su clasificación: matriz de dispersión, coordenadas paralelas, basadas en ícono y las coordenadas en estrella, etc. Se afirma que la matriz de dispersión facilita observar relaciones y comportamientos de atributos bidimensionales.

[15] SPLOM: Matriz de diagramas de dispersión.

2.1.4 Visualización con Coordenadas Paralelas (Parallel Coordinates)

Mecanismo para estudiar la geometría de dimensiones superiores (Inselberg, 1985), es una técnica de proyección geométrica para el análisis de datos multivariados, sustituye la colocación ortogonal, por la colocación paralela, en donde los ejes se colocan paralelos uno al lado del otro, independientemente del número de dimensiones que se presente en el conjunto de datos. Cada ítem de dato o registro, se dibuja por medio de una polilínea, que atraviesa cada eje paralelo en la posición correspondiente a su valor, esta línea poligonal intercepta cada eje Y en un punto; que corresponde al valor para el atributo (VyGLab, 2015).

El propósito general de esta técnica, es visualizar problemas multivariados multidimensionales, sin pérdida de información, donde se transforman relaciones multivariadas, en modelos 2D, de baja complejidad, que permite un tratamiento uniforme para cada variable, y sin un límite de dimensiones, permite revelar resultados con principios matemáticos sólidos. Su gestión es fácil e intuitiva, donde se obtiene información de los atributos del objeto n-dimensional (Castro & Luján-Ganuza, 2017b).

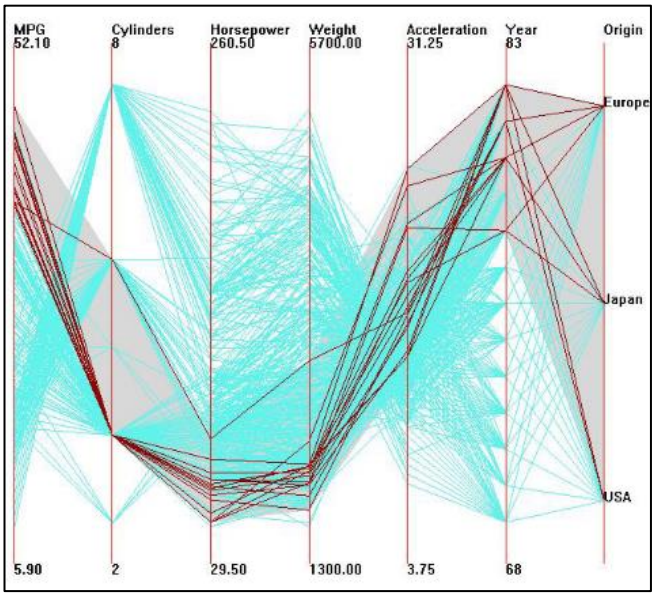
Dentro de los beneficios que se puede obtener de esta técnica, se puede mencionar: representar grandes conjuntos de datos, con valores superiores a tres dimensiones, en un plano 2D, este permite efectividad para el reconocimiento de patrones por el usuario; además, es altamente flexible, debido a su escalabilidad individual para cada eje paralelo permitiendo combinar varias unidades de medias; y también permite la integración de herramientas efectivas (zoom y brush[16]), para una exploración efectiva del conjunto de datos (Castro & Luján, 2017).

Sin embargo, también cuenta con varias limitaciones, a medida que se incrementa la cantidad de dimensiones, los ejes paralelos se acercan entre sí, lo que genera una dificultad para la percepción de comportamientos correlacionales, también puede existir desorden de los ejes paralelos, confundiendo el análisis del usuario gestor, esto, se podría contrarrestar con funcionalidades de reordenamiento y exclusión de dimensiones.

[16] Brush: “cepillado” interacción de selección de un subconjunto de datos en un plano.

Figura 19

Coordenadas Paralelas



Fuente: (VyGLab, 2015), selección interactiva por medio de brush.

Las interacciones son de vital importancia, considerando que el gestor administrativo de un sistema de análisis de datos debe interpretar de manera general una primera instancia, para luego obtener detalles ha pedido, lo que esta técnica permite, es que el gestor pueda modificar parámetros interactivos y obtener retroalimentación inmediata del sistema. El análisis interpreta desde la visualización de la correlación detectada de comportamientos no comunes, hasta la interpretación de conjuntos en múltiples dimensiones, aplicando codificaciones visuales alternativas como envolventes y criterios de la densidad, color o forma de la línea (Heinrich & Weiskopf, 2013).

Al hablar de ciertas interacciones, podemos describir funcionalidades de las coordenadas paralelas:

Brush: se refiere al barrido o cepillado de datos, su función, es envolver y aislar puntos específicos de datos en diagramas, permite al gestor escoger un subconjunto de datos por medio de una herramienta de cepillado o barrido (Heinrich & Weiskopf, 2013).

Interacción con ejes: se refiere a los ejes paralelos, de la técnica, tiene mucha influencia en los comportamientos de los datos, ya que definen un esquema. Algunas de las características de la interacción con los ejes son: la traslación, que permite el intercambio entre variables y la escala, que establece un rango de valores, que pueden ser incluso revertidos (Heinrich & Weiskopf, 2013).

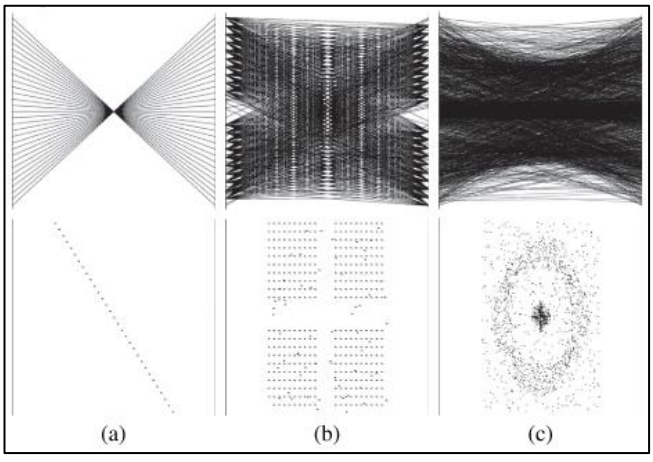
Las series de tiempo, también, pueden ser representadas por medio de esta técnica, pero se refiere a un trato de manera diferente, tienen un orden fijo si se establecen como dimensión conjuntamente con el grupo de datos, son de mucha determinación para descubrir relevantes patrones.

Al existir enfoques con técnicas híbridas, se puede obtener mayor provecho, a los hallazgos de conocimiento que se requiera con base en esta técnica de visualización. Por ejemplo:

2.1.4.1 Scattering Points en Coordenadas Paralelas (SPPC)

Se han integrado visualizaciones de diferentes técnicas en un solo sistema de gestión, para proporcionar una mejor comprensión y exploración del conjunto de datos, la integración de los diagramas de dispersión (escalamiento multidimensional) en coordenadas paralelas, es una técnica visual híbrida, que propone aprovechar las ventajas de las dos técnicas, por ejemplo para evitar desorden de las líneas y problemas de agrupamiento en coordenadas paralelas donde los resultados no son claros, debido a la alta densidad de datos (Xiaoru Yuan, Peihong Guo, He Xiao, Hong Zhou, & Huamin Qu, 2009).

Figura 20
Análisis integración de técnicas visuales



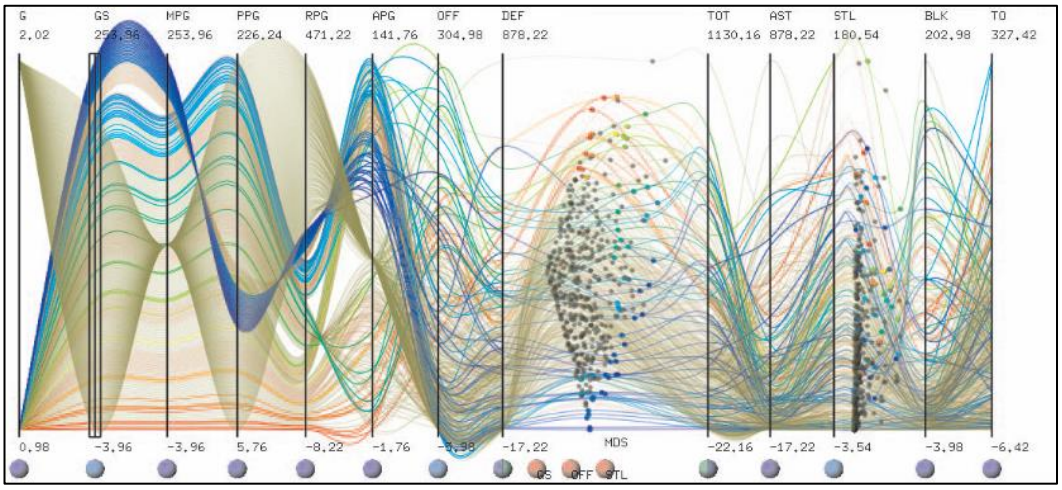
Fuente: (Xiaoru Yuan et al., 2009).

En la figura 20, sección (a), la correlación de datos descrita en la forma de coordenadas paralelas es entendible, sin embargo, en la figura 20, sección (b) y sección (c), no se comprende la información, inclusive si los patrones de distribución en la forma de puntos de dispersión es aparentemente clara (Xiaoru Yuan et al., 2009).

Por tal razón, se integró un método radical SPPC[17], con funcionalidad de cepillado uniforme, que permite que el gestor, analice, explore y agrupe puntos y segmentos de línea de los datos de alta dimensión dentro de una presentación visual, además aporta una representación unificada y un algoritmo de escalado multidimensional acelerado por GPU[18] (Xiaoru Yuan et al., 2009).

Figura 21

Interfaz Scattering Points en Coordenadas Paralelas (SPPC)



Fuente: (Xiaoru Yuan et al., 2009).

2.1.5 Técnicas de visualización orientadas a los datos

En esta sección se describe algunas de las técnicas avanzadas de visualización, orientada a redes, árboles, a la temporalidad y espacio de los datos, que facilitan el análisis de los volúmenes de información y permiten revelar hallazgos por la percepción visual.

2.1.5.1 Árboles y Redes

La agrupación de objetos (nodos), que contienen información vinculada (atributos), e interconectados por medio de relaciones (enlaces), se denomina como una red, pudiendo existir en esta gran cantidad de nodos y enlaces. Los nodos y enlaces, pueden contener información variada, por lo que, se necesita técnicas para interpretar correctamente esta información (Castro & Luján-Ganuza, 2017g).

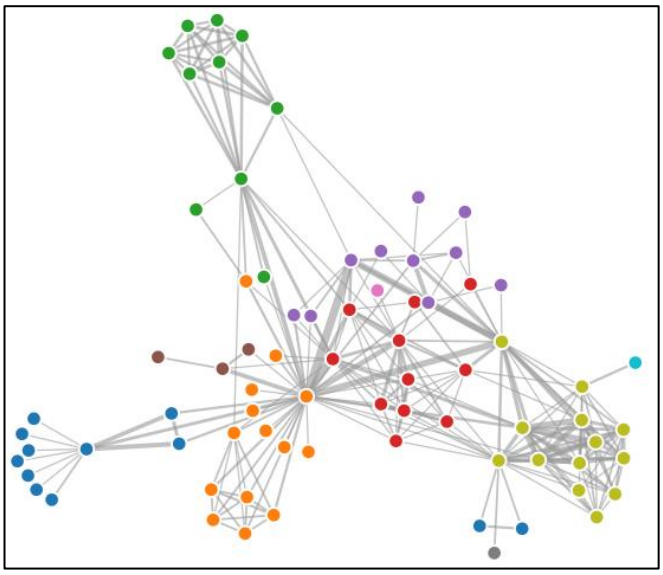
[17] SPPC: Puntos de dispersión en coordenadas paralelas

[18] GPU: Unidad de Procesamiento Gráfico, procesador especializado que ejecuta rápidamente comandos para manipular y mostrar imágenes.

Depende del tipo de objetivo, para que las redes se puedan representar en: técnicas basadas en atributos (TBA) y técnicas basadas en topología (TBT), donde se aplica tareas de localización, cuantificación y ordenamiento. Algunos ejemplos de representaciones de redes son: layout dirigido por fuerzas, redes sociales, layouts lineales, ubicando patrones, matriz de adyacencia, etc. (Castro & Luján-Ganuza, 2017g).

Los árboles son representaciones de redes, con jerarquía relacional, se pueden clasificar para su expresión visual en: Indentación; Nodo-Enlace; Encerramiento; Capas.

Figura 22
Layout dirigido por fuerza



Fuente: (d3js.org).

2.1.5.2 Datos Temporales

El tiempo es una variable de una sola dirección, en la cual no es posible un desplazamiento, por lo que una técnica orientada a los datos temporales, necesita métodos visuales, de proceso y analíticos adecuados, para que se pueda explorar y analizar los datos debido a sus características, en consecuencia, se determina que las variaciones en la línea del tiempo y criterios de temporalidades, son los principales propósitos de interés para obtener conocimiento valioso para el usuario. Para modelar los datos orientados al tiempo, se requiere aplicar tareas como: escala (ordinal, discreta, continua); alcance (información basada en puntos, o información basada en intervalos); arreglo (elementos de tiempo linear, o cíclico), vistas (ordenadas, ramificadas, y múltiples perspectivas). Algunas técnicas son: series de tiempo-tendencias, hora rueda (Time Wheel), tema río (Theme River), etc. (Aigner, Miksch, Schumann, & Tominski, 2011).

2.1.5.3 Datos Espaciales

Objetos dentro de un específico espacio geométrico; estos, generalmente son lecturas de una problemática en el mundo existente, involucra directamente a la visualización geográfica (Geovisualización), área de diseño y desarrollo de sistemas y aplicaciones, que brindan análisis dinámicos, interactivos y altamente visuales de datos espaciales. La reproducción de los datos geoespaciales en forma de imágenes se expresa a través de un mapa, y por sus características existen varias herramientas que permiten su manejo, sin embargo, la táctica básica es una visualización directa, en donde, la información espacial se mapea en una pantalla 2D sobre el cual permite agregar diferentes elementos (Castro & Luján-Ganuza, 2017a).

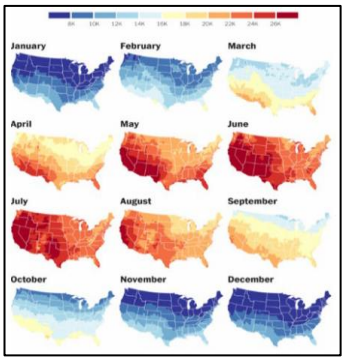
2.1.5.4 Datos Espacio Temporales

Objetos dentro de un específico espacio geométrico que varían en el tiempo, es decir, una distribución de fenómenos espacio temporales se enfoca a: cambios existenciales (eventos instantáneos), cambios espaciales (ubicación, tamaño y forma), y cambios temáticos (cambios en los valores de los atributos), y permite mostrar valores como una trayectoria en cambios espaciales, y en los cambios temáticos, se debe usar múltiples mapas, animación o herramientas interactivas (Castro & Luján-Ganuza, 2017e).

Según, Andrienko G, Andrienko N, Burch & Weiskopf (2012), enfatizó, que los aspectos temporales son cruciales: primitivas, y organización estructural, que hacen un recuento, de lo mencionado antes como tareas de modelado: escala, alcance, arreglo y las vistas, son las pautas de consideración del diseñador para su correcta selección.

Figura 23

Visualización múltiples mapas - espacio temporal



Fuente:(Wonder, 2018).

2.2 Estado del Arte

En varias investigaciones y artículos se ha mencionado que la minería de datos (Data Mining), es una disciplina que ha tenido gran repercusión en el área informática y a lo largo de los años gran acogida, enfocándose en muchos campos donde las decisiones son de vital importancia, algoritmos de: análisis de clúster, clasificación, regresión y asociación, permiten identificar y extraer información valiosa.

Por otro lado, día a día se genera una enorme cantidad de datos, vivimos en la era de Big Data, el procesamiento de Big Data ha crecido, con gran apogeo en la industria y comercio, y se utiliza en todo el mundo, siendo considerado un nuevo e importante campo de investigación. Dispositivos y servicios en la mayoría de las actividades cotidianas han permitido colectar datos en grandes proporciones, en donde se podrán aplicar técnicas de minería de datos, y sacar el mejor provecho (Furht & Villanustre, 2016).

En cuanto a la visualización, existen métodos para la representación de los datos, permitiendo visualizaciones avanzadas en tres dimensiones (3D), siendo esta una área de investigación bastante significativa. No obstante, la evolución de visualización de Big Data seguirá innovándose y creciendo rápidamente, lo que crea la necesidad de resultados multidimensionales debido a su tamaño y complejidad (Zhang, Huang, Wang, Lu, & Meng, 2015).

A continuación, se describirá como: La minería de datos (DM), los grandes volúmenes de datos (Big Data) y la visualización (VIZ) se han implementado, por medio de varias técnicas, métodos y procesos; a través del uso de herramientas de software ha sido posible que se realicen estudios de la contaminación del aire en distintos lugares del mundo.

Un estudio en la ciudad de Chiapas (México) realizó correlaciones entre los contaminantes atmosféricos: dióxido de azufre (SO2), dióxido de nitrógeno (NO2), ácido sulfhídrico (H2S) y material particulado (PM10); y las variables meteorológicas (temperatura, humedad relativa y la dirección del viento). Aplicando el método estadístico de análisis de regresión lineal múltiple, se halló una concentración adecuada para las temporalidades de año y mes, pero la concentración diaria no fue apropiada dentro del rango permitido (Ramos-Herrera, Bautista-Margulis, & Valdez-Manzanilla, 2010).

Al norte del continente americano, en el estado de Nueva Jersey (Estados Unidos), se realizó una metodología de limpieza y preparación para análisis de grandes volúmenes de datos ambientales, contenidos entre los años 2005 al 2012. A través de algoritmos de asociación de árboles de decisión y redes bayesianas, adicional, se ejecutó una comparativa entre las herramientas de extracción de datos (Rapidminer, Weka y Orange), que concluyó con la efectividad de Rapidminer y reflejó, sus resultados por una gráfica de dos dimensiones desarrollada en la herramienta DPlot, que contuvo la distribución del contaminante atmosférico temperatura (Morreale, Goncalves, & Silva, 2015).

En Isfahán (Irán), se estableció una base de datos regional del sistema de información de contaminación del aire gestionada por un Sistema de Información Geográfica (SIG), por medio del método de combinación lineal ponderada, una capa por factor, y se determinó que el 55% de la provincia tiene un rango alto ha moderado por factores meteorológicos inadecuados (Karimi, Soffianian, Mirghaffari, & Soltani, 2016).

Otro estudio en São Paulo (Brasil), aplicó la correlación entre los contaminantes atmosféricos (monóxido de carbono, dióxido de carbono, dióxido de nitrógeno y ozono), y las variables meteorológicas (temperatura y humedad), luego se evaluó el comportamiento promedio de cada variable y se clasificó en un Sistema de Informaión Geografico (SIG), valiéndose de la técnica de mapas auto-organizados. Se encontró como resultado que las áreas urbanas son las afectadas en mayor cantidad. (de Oliveira R., Carneiro, de Almeida, de Oliveira B., Nunes & dos Santos, 2018).

En Dehradun de Uttarakhand (India), se realizó un SIG, basado en un método de mapeo que utilizó un algoritmo de red neuronal de avance para la visualización de puntos calientes del ICA, considerando las variables meteorológicas: temperatura, humedad relativa, precipitación y la velocidad del aire, como entradas para obtener salidas de dióxido de nitrógeno y dióxido de azufre, donde se concluyó, que existieron bajos niveles de los dióxidos tratados (Gupta, Singh N., Singh B., 2017).

El estudio en la ciudad de Ben-galuru (India) desarrolló un modelo de predicción de la calidad del aire basado en la media móvil integrada autorregresiva (ARIMA), se usó la herramienta de programación R. ARIMA y se tomó en criterio las variables de contaminación: (NO2, PM10 y SO2); en donde los resultados han aseverado que es un modelo de predicción preciso (Abhilash, Thakur, Gupta, & Sreevidya, 2018).

Los enfoques de los estudios en la revisión literaria, se orientan cada vez con mayor afluencia hacia los sistemas de interacción, experimentaciones en las diversas ciudades han otorgado resultados satisfactorios en sus representaciones.

En la ciudad de Hong Kong (China), a través de un estudio se interpretó el comportamiento de material particulado (PM2.5) y la correlación con las variables meteorológicas (temperatura, humedad relativa, velocidad viento, dirección del viento), entre el año 2007 y 2008 creando agrupaciones diarias (0-24 horas), y de tipo estacionarias (invierno, verano, otoño y primavera); como resultado se encontró una variación en ciertas horas de las mañanas, y en las horas pico del día, a consecuencia del tráfico vehicular. Así también, se concluyó que existe mayor cantidad de este contaminante PM2.5 en el periodo de invierno (Shi, Wong, Wang, & Zhao, 2012).

En Hangzhou (China) se implementó un sistema visual analítico, este sistema utiliza escalamiento multidimensional (MDS), la agrupación jerárquica de la cual se obtiene clústeres espaciales y un diagrama de Voronoi, para visualizar e interactuar con los mismos. Por medio de mapas de árbol visualizó las estructuras jerárquicas y los atributos detallados de los grupos. Con una tabla de navegación se ilustró cómo los clústeres cambian sus contenidos con la variación del tiempo, permitiendo la apreciación que subconjunto existen menos valores (Zhou,Ye,Liu, Liu, Tao, Su, 2017).

En Beijing (China) un proyecto estableció un método de visualización llamado mapa de calor circular, el cual garantiza rapidez, y comprensión de la información espacio-temporal de los datos en series de tiempo (Li, Fan, & Mao, 2016).

De igual manera, otro estudio en esta ciudad, realizó análisis referente a visualización, una visión múltiple para el análisis visual de los problemas de smog. Se gestionó los datos intervinientes, y resumió en cuatro visualizaciones interrelacionadas; cada una especializada en un análisis diferente, logrando determinar donde se pueden localizar fuentes con técnicas avanzadas VIZ (J. Li, Xiao, Zhao, Meng, & Zhang, 2016).

Otro interesante estudio en esta localidad, realizó la recopilación de información en tiempo real de 23 puntos de monitoreo, aplicó Keyhole Markup Language (KML) basado en el estándar (XML) y soportado por Google Earth, permitiendo formular el etiquetado geográfico para la visualización de los datos monitorizados, este resaltó que la interacción y los aspectos colorimétricos resaltan visualmente los niveles de calidad del aire (Chen, 2019).

También, un estudio de gran relevancia en esta ciudad de Beijing (China); desarrolló, un sistema de análisis visual web del ICA, llamado AirVIS. Se realizó, el procesamiento y análisis de los datos de calidad del aire con información trimestral comprendida entre el 2013 y 2014, con el objeto de determinar peculiaridades espacio-temporales y gráficas de coordenadas paralelas multidimensionales de las agrupaciones monitorizadas de la calidad del aire (Liao, Peng, Li, Liang, & Zhao, 2014).

Dicho sistema, determinado como AirVIS, presentó resultados por medio de tres paneles de visualización: el primero referente a vistas SIG, donde se aprecia las estaciones de captura de los datos; el segundo, un diagrama de dispersión, con valores por hora, de los ICA y el tercer panel, expuso la vista de coordenadas paralelas, entre la hora y seis tipos de contaminantes para la evaluación multidimensional (Liao et al., 2014).

La importancia de que varios estudios se enfocan en esta zona, tiene relación a la cantidad muy significativa de habitantes que tiene China, en consecuencia, es uno de los principales países del mundo que tiene un agudo índice de contaminación atmosférica y también patrones de comportamiento específicos que determina que el índice de la calidad del aire en este país no sea bueno, pudiendo observar una nubosidad oscura que representa el smock, sobre algunas de sus ciudades.

En otra parte del mundo, dentro de la región norte Malasia, un estudio realizado por los autores Thomas, Lokanathan & Jothi (2017), implementó un método de coordenadas paralelas a través de la "descripción personalizable del espacio de parámetros" (APVT) y minería de datos, en donde se visualiza las dimensiones de los factores que contaminan el aire, debido a que el algoritmo de visualización puede trazar los conjuntos de datos multidimensionales, clasificarlos por medio de colorimetría, y representar a través de un filtrado de los datos.

Realizando un enfoque de análisis, hacia las implementaciones multidimensionales, se encontró, que Zhang et al. (2015), desarrolló un estudio implementando coordenadas paralelas con dimensiones 5W (Qué; Quiénes; Dónde; Cuándo; Porque), como los ejes paralelos más la densidad de envío y la densidad de recepción para visualizar Big Data. Aplicados para distintos tipos de datos y ramas de interés, donde demostró atributos y patrones de volúmenes de datos y minimizó superposición, en este proceso obtuvo un 80% de patrones de datos sin pérdida.

Además, Engel, Greff, Garth, Bein, Wexler, Hamann y Hagen (2012), un grupo que realizó un análisis de datos, estableció un marco visual altamente interactivo que redujo la dimensionalidad a los datos, por medio de factorización de matriz no negativa, expuso resultados de gran relevancia. Por ejemplo, transformaciones de registros almacenados en valores tridimensionales, físicamente correctos, con velocidad y gran facilidad.

Por otro lado, Paulovich, Eler, Poco, Botha, Minghim y Nonato (2011), grupo de investigación que desarrolló una interesante técnica de proyección multidimensional, Piecewise Laplacian-based Projection (PLP), que consistió en un mecanismo para interactuar con la información tratada, realizando cambios para modificar la proyección resultante, en este se modifica mapas de proyección dinámicamente siendo esta una característica particular de esta técnica.

Dentro de nuestra ciudad (Cuenca-Ecuador), se han realizado algunos estudios relacionados con el análisis de los contaminantes atmosféricos y variables meteorológicas; algunos con el enfoque del tratamiento de los datos y otros con el propósito de encontrar patrones de comportamientos, tomando en consideración la variabilidad meteorológica que se torna vulnerable en esta zona andina.

Por parte de Sellers (2013), se registró, procesó y publicó datos de los contaminantes atmosféricos colectados por sensores del Municipio de la ciudad de Cuenca, también se validó por medio de estándares normalizados, integración de datos, y visualización web, las observaciones estadísticas de variación de contaminantes en periodos de tiempo. Superó limitaciones de presentación de resultados y realizó la propuesta de una herramienta para la gestión ambiental.

En el trabajo que realizó Ortega-Guamán (2018), se analizó, evaluó y determinó el mejor algoritmo de minería de datos aplicado a los contaminantes del aire (O3, CO, SO2, NO2 y PM2.5), se concluyó que K-Means fue el mejor algoritmo de agrupamiento, y el comportamiento del O3 influye sobre el resto en mayor medida.

Otro estudio se dirigió a la evaluación de las herramientas de minería de datos (Rapidminer, KNime[19] Weka, IBM SPSS Modeler[20] y SAS Enterprise Miner [21]) aplicadas a los contaminantes atmosféricos (O3, CO, SO2, NO2 y PM2.5), a través, del algoritmo K-Means, lo que generó una matriz comparativa, finalmente se comprobó que Rapidminer obtuvo los valores más referentes de los indicadores como funcionalidad y rendimiento (Matute Rivera, 2018).

Adicional, por parte Cabrera-Lituma (2017), se estableció un análisis estadístico de contaminantes atmosféricos registrados por la estación de monitoreo de la EMOV-EP. Conjuntamente con la UDA, se publicó los resultados del ICA por medio de la web, se aplicó tablas y gráficas de los comportamientos, basadas en normativas internacionales.

El autor Andrade-Durazno (2018), se enfocó en la minería de datos considerando contaminantes atmosféricos y variables meteorológicas, donde se usó algoritmos de agrupamiento para el procesamiento de los datos y se concluyó que el algoritmo DBSCAN a partir de la metodología CRISP-DM, fue el mejor tanto en precisión y con mayor eficacia en los datos.

[19] KNIME: "Konstanz Information Miner" plataforma de análisis de código libre, escrita en Java (KNIME).
[20] IBM SPSS Modeler: software de minería de datos y aprendizaje automático (IBM).
[21] SAS Enterprise Miner: herramienta avanzada de minería para crear modelos de predicción (SAS Institut).

3. MÉTODO

Dentro de la presente sección, se describe el desarrollo del método. Se aplicó el método con un ejemplo de verificación, procesos de experimentación y desarrollo para un determinado número de datos colectados. Se implementó la metodología CRISP-DM, con los algoritmos (X-Means, K-Means y DBSCAN) a través de la herramienta Rapidminer para la creación de un modelo de minería de datos y la visualización avanzada por medio de la técnica visual de Mapas de Coordenadas Paralelas. Con respecto a los contaminantes atmosféricos y las variables meteorológicas se exponen los resultados en una aplicación web que contiene un gráfico dinámico desarrollado con D3.js.

3.1 Zona de Estudio

Dentro del análisis, se examinó la zona del estudio, para establecer el alcance del desarrollo del trabajo de graduación. El proyecto es definido con base a las relevantes investigaciones realizadas previamente en la ciudad de Cuenca, Ecuador (Andrade Durazno, 2018; Ortega Guamán, 2018). La misma se encuentra ubicada en la región sierra al sur del país a una altitud de 2550 metros sobre el nivel del mar y con un clima regional andino de 16 grados celsius en promedio.

La conjunta operabilidad entre la Universidad del Azuay y la Municipalidad de la Ciudad a través de su organismo de control "EMOV-EP", ha proporcionado valiosos resultados previos dentro de la zona de estudio. Estos resultados han sido logrados debido a la gestión del proyecto ICA dirigido por el Departamento de Investigación del Instituto de Estudios de Régimen Seccional del Ecuador (IERSE) perteneciente a la UDA, por lo que de manera consecuente, se analizó la información con un nuevo enfoque de gestión visual y presentar una información oportuna para un beneficio a la sociedad (EMOV EP, 2017; IERSE, 2017).

Figura 24
Mapa de la Ciudad de Cuenca-Ecuador

Fuente: (Google Maps, 2019)

3.2 Datos del Estudio

Los datos para el desarrollo de la investigación han sido recolectados por una estación automática de monitoreo del aire a cargo de la EMOV EP, la cual, realiza la medición de los contaminantes atmosféricos dentro de la ciudad; dichas mediciones, se transfieren al departamento de investigación IERSE, que almacena los registros en una base de datos que contiene grandes cantidades de información. Luego, este departamento actúa como gestor y compartió el set de datos de interés de la investigación.

Según los estudios realizados por IERSE y los informes anuales del proyecto ICA, en colaboración con la EMOV EP (2017), los gestores han determinado un conjunto de variables principales que intervienen en el análisis de contaminantes ambientales, las cuales son: el Ozono (O3), Dióxido de nitrógeno (NO2), Dióxido de Azufre (SO2), Monóxido de Carbono (CO), y Material Particulado (PM2.5) (IERSE, 2017).

Las variables meteorológicas consideradas para sus permanentes estudios, que son parte también de la presente investigación son: Temperatura del Aire (TEMPAIRE), Humedad Relativa (HR), Punto del rocío (DP), Presión Atmosférica (PATM), Radiación Global (RADGLOBAL), Precipitación (PRECIP_SUM), Velocidad del Viento (WINDSPEED), Dirección del Viento (WINDDIR), Radiación UVA (UVA), Radiación UVE (UVE) (EMOV EP, 2017; IERSE, 2017).

Algunos estudios realizados anteriormente, donde intervienen las variables seleccionadas, han servido de base para el desarrollo en los hallazgos cognitivos de interés. Sellers (2013) analizó el comportamiento de los contaminantes en una temporalidad determinada, así mismo Ortega-Guamán (2018), analizó, evaluó y determinó el mejor algoritmo de minería de datos. Matute Rivera (2018), realizó un similar análisis, enfocándose a la evaluación de las herramientas de minería de datos, finalmente Cabrera-Lituma (2017), estableció un análisis estadístico publicado en la web, todos estos, de las anteriormente mencionadas variables de contaminación.

Andrade-Durazno (2018), además de considerar las variables contaminantes, también consideró las variables meteorológicas para aplicar la metodología CRISP-DM y algoritmos de agrupamiento. Por estas razones previas, todas estas variables de contaminantes y variables meteorológicas son parte del desarrollo experimental del presente trabajo de graduación, específicamente, grandes volúmenes de información y el enfoque prioritario a la visualización de sus comportamientos.

3.3 Requerimientos del método CRISP-DM

La minería de datos transforma los conjuntos de datos en conocimiento, es indispensable realizar una estructura de procesos que garantice el cumplimiento de los objetivos trazados, y resolver la problemática establecida. En esta sección se detallan los requerimientos para la aplicación del método CRISP-DM y de las herramientas preliminarmente seleccionadas para el desarrollo del estudio.

Se aplicó la metodología CRISP-DM, característica por su flexibilidad en sus etapas y su fácil personalización. A través de la revisión literaria del marco teórico se sustenta que ha sido el método más utilizado en los últimos años, además, se aplicó esta metodología en el estudio antecesor realizado por Andrade Durazno (2018), lo que se consideró como criterio prioritario para la ejecución del estudio.

Para el desempeño de esta metodología, conjuntamente con el análisis de Big Data, se necesita seleccionar e instalar una herramienta software de gestión; razón por la que, se seleccionó a la herramienta Rapidminer como el software de gestión para el presente trabajo de graduación. En varios estudios mencionados en el estado del arte, esta herramienta software ha sido sometido a varios estudios comparativos, resultando la más efectiva en funcionalidad y rendimiento (Matute Rivera, 2018).

Rapidminer ofrece costo gratuito para el uso de 10.000 registros de análisis de datos. Para un número mayor de registros se requiere una licencia, misma que varía entre 2.500 hasta 10.000 dólares de costo según el número de registros límite por un periodo anual. Sin embargo, se puede solicitar una licencia educativa libre de costo, por ejemplo, para desarrollo proyectos de investigación vinculados a una institución educativa, por un periodo de tiempo determinado (Rapidminer, 2019).

Los registros que se analizan con la herramienta Rapidminer deben ser ingresados en las diferentes etapas del método dentro de la experimentación, por lo que se requiere contar con un conjunto de datos inicial, datos que hayan sido recolectados en un archivo con extensión CSV o XLSX. Con este requerimiento se creó un modelo para el flujo de datos, y la aplicación de algoritmos de agrupación. Posterior al tratamiento de toda esta información se obtuvo un archivo con la misma extensión que contiene los valores estadísticos resultantes para su presentación.

En cuanto a los resultados, estos deben ser presentados de forma clara y entendible, es así que se ha determinado el uso de herramientas dinámicas, versátiles y sencillas para la construcción de una aplicación web, donde se han exhibido los resultados encontrados.

Para el desarrollo de la aplicación web, se requiere seleccionar las correctas tecnologías para sus instalaciones previas, en este análisis del estudio se consideró el lenguaje de marcado HTML conjuntamente con el editor de texto de código fuente, Sublime Text, debido a su rendimiento y eficiencia.

Los comportamientos de las variables que intervinieron en el estudio, dentro de los resultados obtenidos, se representan por medio de una técnica de visualización avanzada, esta técnica fue desarrollada por medio de la herramienta D3.js, por lo que se necesitó la instalación de esta librería que permite la creación de gráficos altamente dinámicos e interactivos con el uso de las tecnologías y materiales CSS, SVG y DOM.

La técnica de visualización seleccionada, el proceso, diseño y la explicación a detalle de estos contenidos, se describirá en el apartado “Implantación”, que se encuentra dentro del desarrollo de la metodología CRISP-DM aplicada al ejemplo de verificación.

En la sección “Visualización”, se describe la funcionalidad, interactividad y control de la herramienta web, también se detalla todas las especificaciones técnicas necesarias para la creación de la misma y los detalles de la interfaz de manejo de la herramienta gestora de los resultados y hallazgos de conocimiento relevante.

La clasificación y filtrado dinámico de temporalidad también fue necesaria para la estructura de los resultados visuales, esta temporalidad está basada en las características de los datos relacionados con fechas de recolección de las muestras. Se crearon variables que permitieron realizar agrupaciones de análisis, por ejemplo: agrupaciones por horas, días, semanas, meses, trimestres, etc. Esto permite una adecuada interpretación de los resultados.

Para todos los requerimientos detallados para el desarrollo del presente trabajo se encontró soporte técnico suficiente. Una gran ventaja en cuanto a las herramientas que se utilizaron para los procesos fue la facilidad de acceso y soporte, de tal forma se pudo evitar inconvenientes dentro de la evaluación experimental, lo que garantizó el cumplimiento del propósito general y los objetivos específicos.

3.4 Desarrollo de la Metodología CRISP-DM (Ejemplo de verificación)

En esta sección, se desarrolló la metodología basándonos en la experimentación, se describe paso a paso la implementación de la metodología propuesta, al ejemplo de verificación y los datos seleccionados, se realizaron dos evaluaciones experimentales con dos conjuntos de datos que varían en tamaño y tiempos de recolección por el IERSE.

3.4.1 Comprensión del Negocio

Considerando que la percepción visual de los grandes volúmenes de datos es una ciencia innovadora de la cual se puede obtener grandes beneficios, se realizó el enfoque de los objetivos a las técnicas de visualización avanzada.

Según la valoración actual, en el medio local no existe un estudio que analice el comportamiento entre contaminantes del aire y los factores meteorológicos, a través de técnicas de visualización avanzada, la representación de los resultados por medio de una técnica de visualización especializada. Lo que permite establecer como propósito el diseño y desarrollo de una gráfica analítica interactiva para descubrir aspectos y patrones de comportamiento basados en temporalidades dentro de la ciudad, facilitando la representación de los análisis complejos multidimensionales con dinamismo e interactividad en una plataforma web.

Los planes de acción, la gestión atmosférica ambiental y la evolución de los hallazgos de conocimiento establecidos por el estudio como resultados, permite facilitar la toma de decisiones por parte del gestor y las entidades involucradas en el medio de interés (IERSE, LIDI, EMOV, etc). También, proporciona bases para análisis y estudios a largo plazo.

3.4.2 Comprensión de los Datos

Las tareas que se realizaron dentro de esta etapa se describen a continuación:

3.4.2.1 Recolección de datos.

El set de datos que facilitó el IERSE fue sometido a procesos de tratamiento, estos procesos fueron realizados en fases anteriores, específicamente en la segunda fase donde se involucró variables de contaminación y factores meteorológicos, dichos procesos han aportado para la evolución del proyecto Índice de Calidad del Aire (ICA). Como resultado se obtuvo un nuevo conjunto de datos para la gestión que paso a cargo del Laboratorio de Investigación y Desarrollo en Informática (LIDI) de la Universidad del Azuay.

En este proceso de recolección de datos en una primera etapa de experimentación, se colectó una proporción parcial de los datos totales. LIDI proporcionó un total de 3.936 datos recolectados con intervalos de 10 minutos. Los registros se levantaron durante 28 días, esta información fue proporcionada en el archivo "dataAtmCon.csv" para su primera esquematización.

3.4.2.2 Descripción de los datos.

Al determinar la descripción de los datos relevantes dentro de la investigación, se realiza una clasificación simple entre los contaminantes atmosféricos y las variables meteorológicas:

3.4.2.2.1 Contaminantes Atmosféricos

Los contaminantes atmosféricos descritos en el set de datos hacen referencia a los considerados en las etapas anteriores del proyecto. Estos son determinados según la necesidad del Instituto de Estudios de Régimen Seccional del Ecuador (IERSE).

Tabla 1
Contaminantes Atmosféricos

Contaminante Atmosférico	Abreviatura	Unidad
Ozono	O3	Partes por billón
Monóxido de Carbono	CO	Partes por billón
Dióxido de Nitrógeno	NO2	Partes por billón
Dióxido de Azufre	SO2	Partes por billón
Material Particulado	PM2.5	Microgramos por metro cúbico

Distribución de los contaminantes atmosféricos, su abreviatura y su unidad de medida.

Los contaminantes ambientales generalmente provienen de fuentes naturales, es decir, actividades que no involucran ninguna actividad humana (erupciones volcánicas, erosión del suelo, etc.), sin embargo, también son generados de manera artificial provocada por varias actividades realizadas por personas, entre ellas, las actividades de fuentes fijas que desarrollan los procesos industriales y comerciales de las áreas petroleras, químicas, entre otras. Además de las fuentes móviles que hace referencia a todos los automotores en circulación, a través de las emisiones comunes de los escapes (Sellers, 2013).

Los contaminantes del aire, en un contexto general tienen graves efectos en la salud de las personas, afectan a los organismos y provoca vulnerabilidades como: los sistemas respiratorios, afecciones dermatológicas tópicas severas, desarrollo de alergias por sensibilidad en los sistemas inmunológicos, y problemas oftalmológicos agudos, entre otros (EMOV EP, 2017).

Ozono (O3)

Es un gas, establecido por tres átomos de oxígeno (O3), característicamente incoloro e inoloro, que se produce en escenarios de presión y temperatura adecuados por la mezcla del óxido de nitrógeno (NO) y compuestos orgánicos volátiles (COV), sirve de filtro contra la radiación solar ultravioleta, cuando se encuentra en la estratosfera[22], sin embargo, cuando se encuentra en la superficie troposfera[23], causa afecciones a las personas, como problemas pulmonares e irritaciones tópicas, así como a las plantaciones, cultivos y vegetación, siendo principal componente del smog o neblina de contaminación (Sellers, 2013; Martínez-Ataz & Díaz, 2004; Torres, 2002).

Monóxido de Carbono (CO)

Es uno de los principales contaminantes que influye en el índice de la calidad del aire, característicamente es un gas inodoro, incoloro, muy tóxico, que se produce cuando no existe una correcta combustión de sustancias como la gasolina o el gas, generalmente proviene de fuentes móviles a través de las emisiones de los escapes, y puede causar graves daños a la salud, inclusive la muerte, en exposición de niveles elevados que reducen oxígeno en la sangre (Sellers, 2013; Martínez-Ataz & Díaz, 2004; Brunekreef & Holgate, 2002).

[22] Estratosfera: segunda capa de la atmósfera cercana a la superficie terrestre (OZONO, 2012).
[23] Troposfera: primera capa inferior, o más próxima a la superficie terrestre (OZONO, 2012).

Dióxido de Nitrógeno (NO2)

Característicamente es un gas de color café rojizo, irritante, oxidante, altamente reactivo, y soluble, que se produce como residuo por la combustión de motores en altas temperaturas, en los sectores vehiculares, industriales y naturales, como la actividad volcánica, y por descargas eléctricas atmosféricas, también por el uso de combustibles fósiles; es un principal componente del smog y la lluvia ácida (Sellers, 2013; Martínez-Ataz & Díaz, 2004; Brunekreef & Holgate, 2002).

Dióxido de Azufre (SO2)

Es un gas, característicamente irritante, incoloro, con un olor penetrante, altamente reactivo, y denso, que se produce por la quema de combustibles fósiles sulfurosos como carbón o derivados del petróleo; generalmente proviene de fuentes urbanas e industriales, como las calefacciones y puede causar graves daños al sistema respiratorio, este gas, en contacto con la humedad del aire, forma el ácido sulfúrico, que produce efectos negativos a la flora, y también perjudica a estructuras físicas por la corrosión del material (Sellers, 2013; Fundación para la Salud Geoambiental, 2013; Brunekreef & Holgate, 2002).

Material Particulado (PM2.5)

Es la combinación compleja de partículas líquidas y sólidas, extremadamente pequeñas que se encuentran en suspensión en el aire, como nitratos, sulfatos, químicos orgánicos, entre otros. Su dimensión tiene directa relación con el efecto dañino sobre la salud, y principal mente afecta al sistema respiratorio y al desarrollo de alergias debido a la facilidad de ingreso por las vías respiratorias, lo que permite su llegada a los pulmones. Se clasifican en dos grupos, material particulado de 10um (PM10) y las partículas finas o ultrafinas determinadas por su diámetro como material particulado de 2.5um (PM2.5) (Sellers, 2013; Maté, Guaita, Pichiule, Linares, & Díaz, 2010; Fundación para la Salud Geoambiental, 2013).

3.4.2.2.2 Variables Meteorológicas

La meteorología es una rama que estudia las propiedades y los fenómenos de la atmósfera[24], se basa en el conocimiento de una serie de variables meteorológicas que varían en espacio y tiempo (FECYT, 2004). Las variables meteorológicas descritas en el set de datos, hacen referencia a los datos tomados en consideración en fases anteriores del proyecto ICA. Estos datos son determinados según la necesidad del IERSE. En el trabajo se tomó en consideración las siguientes variables.

Tabla 2
Variables Meteorológicas

Variable Meteorológica	Abreviatura	Unidad
Temperatura del Aire	TEMPAIRE	°C (Grados Centígrados)
Humedad Relativa	HR	% (Valor porcentual de agua)
Punto del Rocío	DP	°C (Grados Centígrados)
Presión Atmosférica	PATM	hPa (Hectopascal)
Radiación Global	RADGLOBAL	w/m2 (Vatios por metro cuadrado)
Precipitación	PRECIP	mm (Milímetros de agua)
Velocidad del Viento	WINDSPEED	m/s (Metros por segundo)
Dirección del Viento	WINDSDIR	° (Grados Angulares)
Radiación UVA	UVA	w/m2 (Vatios por metro cuadrado)
Radiación UVE	UVE	w/m2 (Vatios por metro cuadrado)

Distribución de las variables meteorológicas, su abreviatura y su unidad de medida.

Temperatura del Aire

La temperatura es una magnitud muy recurrida, para expresar el estado de la atmósfera, es decir el movimiento las partículas que conforman la materia. La temperatura del aire varía, entre una ubicación geográfica, entre las diferentes horas del día, también entre las estaciones climáticas, y se usa un instrumento llamado termómetro para su medición, que se basa, en los cambios de las propiedades de la materia (FECYT, 2004).

Humedad Relativa

La humedad relativa, es la presencia de vapor de agua en el aire, determinada por un valor y su variabilidad depende de varios factores del entorno, como flora, clima, y ubicación geográfica, etc. Se expresa de manera porcentual, y determina el nivel máximo tolerado (saturación) de vapor de agua que una masa de aire soporta (FECYT, 2004).

[24] Atmósfera: capa gaseosa que rodea la tierra, en ella comprenden cinco subcapas (EL OZONO, 2012).

Considerando la altura, la variación de la humedad determina, que el aire a menor altura presenta una mayor saturación, consecuentemente, el aire a mayor altura presenta poca humedad (Basantes & García, 2018).

Punto del Rocío

El punto de rocío, "temperatura de punto de rocío" es la temperatura requerida para la condensación, se refiere al vapor de agua que se encuentra en un gas; basado en el comportamiento de los gases, la concentración de vapor de agua tiene gran variabilidad, y para determinar el volumen, es necesario que sea medida a través de la condensación, produciendo un efecto de nubosidad (Vaisala, 2013).

Presión Atmosférica

Esta variable, se refiere a la fuerza que ejerce el aire sobre los cuerpos o la superficie, debido a la gravedad, tiene dependencia con diferentes variables, como situación geográfica, temperatura, humedad y condiciones meteorológicas; pero mayor incidencia con la variable altitud en la atmósfera; a mayor altitud menor presión, y a menor altitud mayor presión que el aire despliegue sobre un cuerpo (FECYT, 2004).

Radiación Global

La radiación global, es la energía que se transfiere del sol a la tierra, por medio del espacio, en forma de ondas, para la atmósfera es casi irrelevante esta radiación, sin embargo, la superficie terrestre y los cuerpos sí la absorben, se mide mediante un piranómetro (FECYT, 2004).

Precipitación

Se refiere a la cantidad de agua que se encuentra en la superficie terrestre que proviene de la humedad atmosférica, ya sea en estado líquido, como lluvia, o estado sólido como nieve o granizo; es un proceso de vital importancia para la Hidrología sin embargo, la humedad no garantiza la precipitación, ya que depende de algunos factores para llegar a la saturación (Universidad de Piura, 2015).

Velocidad del Viento

El viento se genera por el movimiento de las partículas aire en el espacio atmosférico, entre dos sitios de diferente de presión o temperatura, la velocidad del viento se produce por la energía del movimiento, la cual, cerca de la superficie terrestre o suelo, es menor, y cuando la altura alcanza niveles más grandes, la velocidad del viento aumenta, se mide a través de la interpretación de kilómetro por hora o metro por segundo (FECYT, 2004).

Dirección del Viento

Dentro de la atmósfera, está presente una directa relación entre presión y viento, lo que permite determinar medidas del viento, esta variable describe, la procedencia geográfica del viento o del movimiento del aire (FECYT, 2004).

Radiación UVA

El sol transmite energía en forma de ondas, estas, se clasifican en el espectro electromagnético, que comprende desde el rango del ultravioleta (UV), hasta las ondas de radio. La capa de ozono, es la encargada de filtrar la radiación ultravioleta, sin embargo, la radiación UVA traspasa en menor medida y produce efectos adversos a la salud de las personas; la radiación UVA puede penetrar profundamente en la epidermis y la dermis de la piel, causando foto-envejecimiento prematuro de la piel, fotosensibilidad y daño en la retina, afecta a los vasos sanguíneos dérmicos (González, Tamayo, & Sánchez, 2009).

Radiación UVE

Cuando la energía solar, atraviesa la atmósfera, algunos gases contenidos esta capa, absorben toda la radiación UVC, por lo que la radiación que llega a la superficie terrestre se compone de UVA y UVB (Cañarte, 2007). La radiación puede llegar a sus niveles críticos, o sobrepasar el límite de tolerancia, por lo que UVE (Dosis Eritémica), se presenta como una nueva determinación de la radiación UV, que comprende la radiación UVB en su totalidad y sobre el 40 % de la radiación UVA (Lavorato, Moscatelli, Pagura, Zanin, Lacomi & Cesarano, 2010).

3.4.2.3 Exploración de los datos.

La exploración de los datos, permite determinar que los valores de los contaminantes atmosféricos y las variables meteorológicas que han sido relevantes para la investigación se encuentran descritos en valores promedio, se consideró, que con base a la experimentación que se ha tenido como barrido inicial de lectura de los datos, estos valores están descritos en tipos de datos, de clase enteros y reales; y dependiendo de la dimensión en algunos de los casos constan con números decimales de 4 o 5 dígitos.

3.4.2.4 Verificación de calidad de los datos.

El 16% de los valores del set de datos inicial fueron valores nulos contenidos en la columna “HORADEC”, variable considerada para la estructura de la temporalidad en la que se basa el análisis. Además, la herramienta de minería excluye los errores referentes al tipo de datos y los interpreta como datos nulos. Sin embargo, la cantidad de datos es considerable, ya que el 84% de los datos no presentan valores nulos en el set de datos inicial de la primera experimentación realizada. El set de datos de la investigación inicial consta de valores distribuidos en registros guardados cada 10 minutos, dicho set de datos inicial fue el resultado de la gestión y tratamiento que realizó Andrade (2018), de tal forma que la comprensión es fácil para el modelado y depuración.

En una segunda etapa se cuenta con un total de 54.000 registros de un periodo de 365 días proporcionados para la segunda experimentación, donde posterior al proceso de la depreciación de valores nulos se disponen 34.231 registros útiles para el análisis, correspondiendo al 64% del total de la información. El set de datos con mayor volumen de datos fue proporcionado en el archivo “Base_contaminantes.csv” para su posterior esquematización. En las dos etapas de la experimentación se pudo disponer de un porcentaje de datos importante. En síntesis, la calidad de los datos fue sufiente y efectiva para los dos procesos experimentales realizados.

3.4.3 Preparación de los Datos

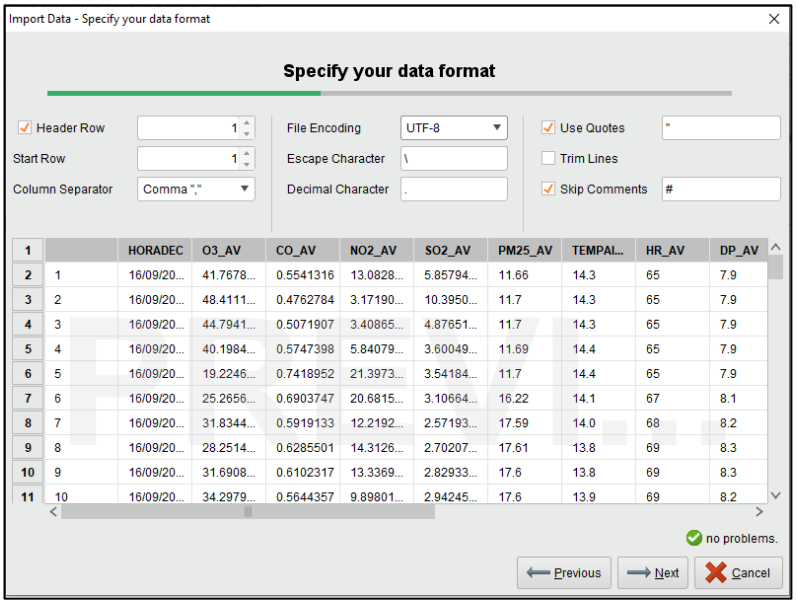
3.4.3.1 Selección de datos.

La selección de los datos parte desde el proceso de carga dinámica del archivo “Import Configuration Wizard” dentro de la herramienta Rapidminer, se puede destacar las características del archivo y la familiarización de la estructura comprendida en el set de datos.

En la figura 4, se observa el ingreso del formato de nuestro set de datos, en donde se determina características de cabecera e inicio de fila, también el separador, y algunos otros parámetros considerados relevantes para la selección de los datos.

Figura 25

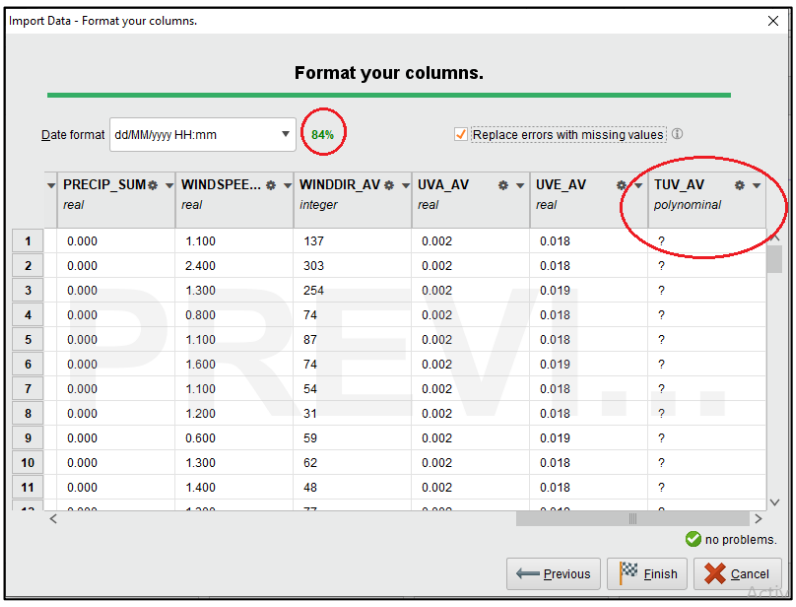
Especificaciones de formato de los datos



Continuando con el proceso de selección de los datos, no se consideró la columna total del contaminante TUV_AV, debido a que no existió datos de esta variable en el conjunto de datos. En la figura 26 se observa, que la columna no fue considerada, de tal manera que se depuró en la incorporación inicial para la gestión de los datos.

Figura 26

Selección de datos - Exclusión de datos



Es importante recalcar que la distribución del set de datos inicial para la investigación, se alcanzó, como resultado de la investigación previa realizada en la segunda etapa de análisis de los contaminantes atmosféricos y variables meteorológicas, donde se verificó una similitud entre dos variables (HORADEC y FECHOR), y se alineó valores a la temporalidad requerida, generándose el set de datos unificado base, considerando a la fecha y hora de colección de los valores (Andrade, 2018).

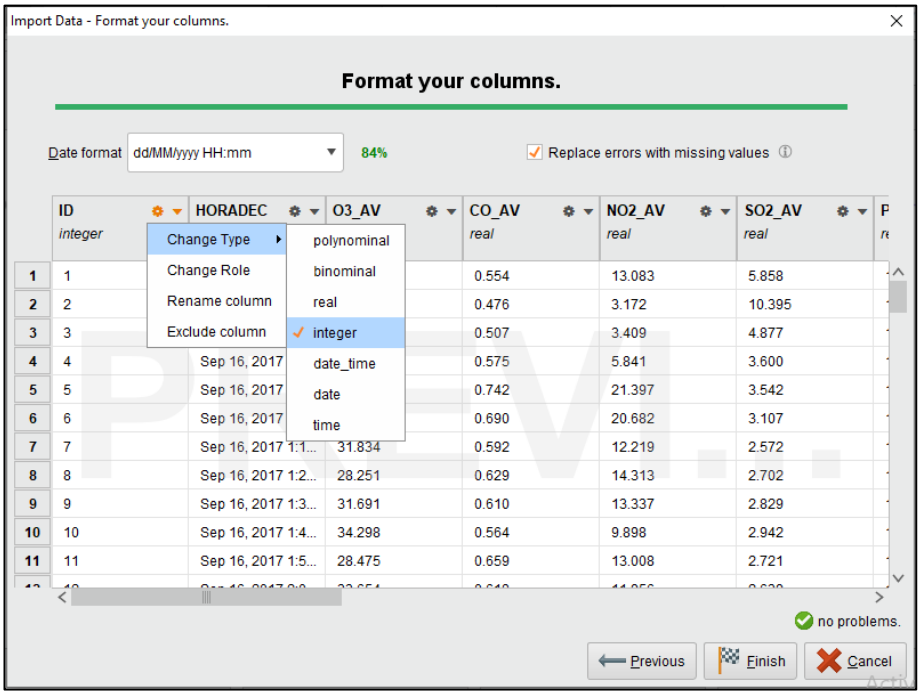
3.4.3.2 Limpieza de datos.

Dentro del proceso de limpieza de los datos se determinó un formato de fecha para la columna correspondiente a HORADEC, de donde se pudo extraer datos aislados con el propósito de añadir columnas, lo que facilitó una estructura de una temporalidad deseada para el análisis de los contaminantes atmosféricos y las variables meteorológicas. El 84% de los datos se adapta al formato haciendo excepción de un 16% de los valores nulos que contiene la columna, además la herramienta reemplaza los errores de datos con datos nulos, esto como una previa limpieza para su posterior exclusión del análisis dentro de la minería de los datos.

El proceso de "Import Configuration Wizard", permite determinar parámetros o cambios algunos como: el tipo de dato, rol que asumirá dentro de la minería de los datos, también permite renombrar la columna o excluirla del set de datos según sea necesario.

Figura 27

Formato de columnas (Limpieza de datos)



3.4.3.3 Estructura de datos.

La etapa de preparación de los datos fue la más importante dentro de la metodología aplicada, razón por la cual se esquematizó por diferentes subprocesos en donde se destacan cinco fases. Estas fueron establecidas para obtener el set de datos final, que permite evaluar de la manera puntual y pertinente la información prescrita para los objetivos planteados.

Figura 28

Fases de la estructura de los datos

3.4.3.4 Normalización de los datos.

Luego de varios intentos experimentales con el tratamiento de los datos, se determinó estrictamente necesario realizar una normalización al set de datos, ya que la distribución de los datos se apreciaba de manera general por la diversidad de las unidades de todas las variables consideradas. Al realizar las evaluaciones de prueba, se observó que la distribución hacia los diferentes clústers debía ser proporcional, de tal manera que se obtengan los patrones de comportamiento eficientes.

Se implementó y configuró dentro del modelo de minería la normalización de todos los datos del set, de tal forma que la distribución de los clústers fue basada en esta característica del modelo. Posteriormente se obtuvo la distribución de los clústers, y se retornó a los datos principales donde se consideró las unidades originales.

El proceso fue necesario para crear una columna que determinó el clúster donde se encontraba el registro, para que ese atributo fuese considerado en aspectos colorimétricos y relevantes para la presentación gráfica.

Una observación que se realizó dentro de la experimentación con la normalización, fue la cantidad de clústers que se generaban dentro de las evaluaciones de los diferentes algoritmos. Se esperó una distribución de hasta cuatro conjuntos de datos en los resultados, para conformar la estructura del set de datos final. Para ello se realizó un largo periodo de experimentación con 34.231 registros de datos.

En conclusión, la estructura de los datos dentro de la fase de tratamiento de la información, así como en la fase de clasificación de la información se adicionó columnas que representan dimensiones adicionales, algunas como: año, mes, día, hora. Esto permitió representar las agrupaciones temporales, de tal forma que el set de datos concuerde con los requerimientos para cumplir con el propósito del trabajo de titulación.

3.4.3.5 Integración de los datos.

Con las etapas previamente analizadas y con el propósito final claramente establecido, el conjunto de datos resultante para el análisis de los contaminantes y variables meteorológicas contiene las variables necesarias para la construcción de las observaciones de análisis, sumarizando un total de 21 atributos o columnas. Se adapta e integra correctamente a la estructura de datos requerida por la herramienta web para el resultado visual, permitiendo mostrar al gestor 21 paralelas que se refiere a la técnica de visualización seleccionada desarrollada con las herramientas HTML5 y D3.js.

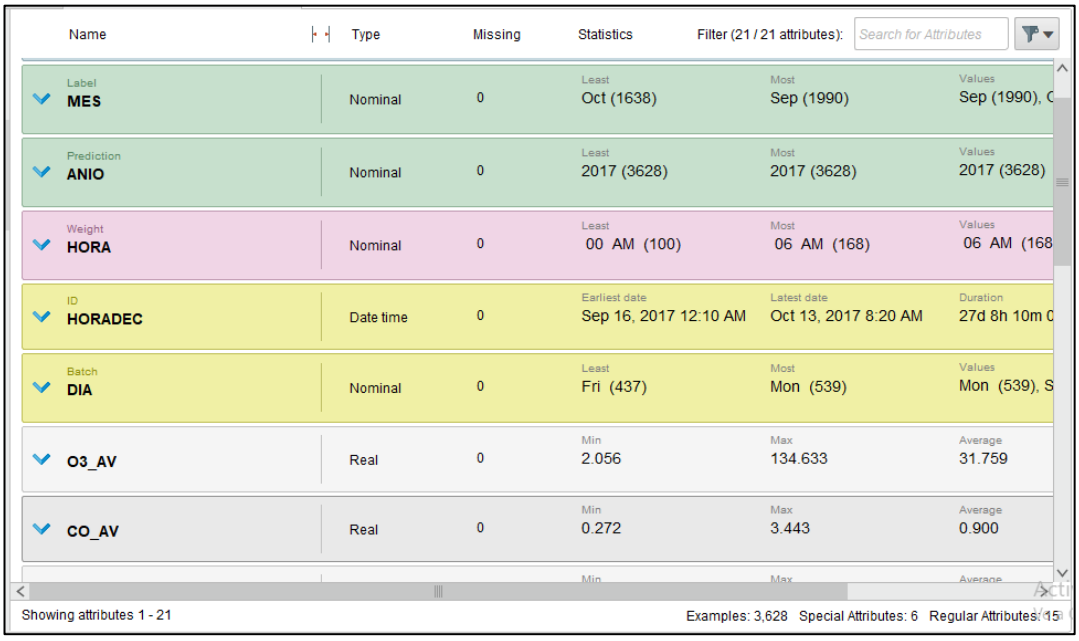
Luego de la aplicación de la metodología hasta esta etapa, ya se tiene el set de datos final. La sección detalle estadístico permite apreciar estos datos, como máximos, mínimos, promedios, y valores referentes a clasificaciones, de esta forma en la etapa del modelado el set de datos se adapta a los procesos de clusterización.

Para este set de datos se eliminó del mismo los valores nulos, ya que los algoritmos de minería de datos (X-Means, K-Means, DBSCAN) no permiten su ejecución de datos nulos en sus atributos.

El número total de elementos nulos que se descartó en la primera etapa de experimentación, fue de 308 de un total de 3936 datos considerados, del mes de septiembre a octubre del año 2017. En la segunda etapa de la experimentación, ascendió a 18.330 datos nulos, de un total de 52.560 datos considerados, comprendidos de enero a diciembre del año 2018, resultó un total de 34.231 registros considerados para el gestor ambiental. Estos datos se depuraron debido a errores en los datos y datos nulos que no contenían información, por lo que la herramienta de minería usada excluyó estos valores.

Figura 29

Set de datos Final (Estadística)



3.4.4 Modelado de datos

Posterior al análisis, la depuración y limpieza de los datos, es imprescindible definir y construir el modelo que se utilizará para la minería. Dentro de esta sección, el modelo que se utilizó fue el agrupamiento de datos (Clustering), ya que permitió juntar los valores de los contaminantes con los agentes meteorológicos, este agrupamiento resultó ser el más efectivo, sin embargo, el agrupamiento resultante con los algoritmos de evaluación (X-Means, K-Means, y DBSCAN), es bastante generalizado, debido a la naturaleza de los datos.

El modelado se desarrolló, con las distintas utilidades de la herramienta Rapidminer, con su funcionalidad se estableció el tipo de modelo y se construyó orientado al set de datos final que se obtuvo en las etapas anteriores de la metodología.

Cuando se habla de clusterización o agrupamiento, se refiere a una técnica de minería de datos que el objetivo principal es buscar similitudes de agrupamiento, de tal forma que los objetos en un grupo sean similares entre ellos y diferentes a otro grupo (Puolamäki, Papapetrou, & Lijffijt, 2010).

Un aspecto importante que se tomó en cuenta, fue contar con los distintos algoritmos de agrupamiento (X-Means, K-Means, DBSCAN) dentro de un set de datos resultante, de tal manera que pueda ser una dimensión de consideración para determinar patrones basándose en a la cantidad de datos que contiene cada clúster del algoritmo aplicado.

En el desarrollo del modelo de los datos, interfieren algunas de las utilidades:

3.4.4.1 Read CSV

Lee una fuente externa al software en formato CSV. Esta utilidad se usó de manera implícita en las etapas anteriores, lo que facilitó la gestión a los datos en un entorno intuitivo. La información nula o valores en blanco es añadida para la gestión, sin embargo, se menciona que no se pudo procesar algoritmos o funciones de minería con datos nulos, debido los aspectos contemplados en la metodología. Esto se solucionó por medio de la lectura "Read CSV" que corrigió la información que ingresó, pero además existe una tarea específica para omitir valores nulos.

3.4.4.2 Filter Nulos

Con esta utilidad se estableció un filtro a la información. El proceso comprende el filtrado de los valores nulos. Se asigna el nombre de la variable y la asignación del filtro "is not missing", el resultado proporciona que el modelo pueda procesar la información sin errores.

3.4.4.3 Generate Attributes

Herramienta de utilidad para generar un atributo basado en la información que pueda contener otro atributo. Las funcionalidades de Generate Attributes permitió aislar y adicionar columnas, que fueron necesarias para la correcta interpretación, a través de función "date_str_custom" se obtuvo parámetros importantes para el set de datos final.

La operación "function expressions", permite establecer el valor de un atributo con base en una expresión lógica, de comparación, información de texto, funciones matemáticas, funciones trigonométricas, etc.

3.4.4.4 Set Role

Este componente se utiliza para establecer un peso o fijar un rol al atributo. En el modelo de datos desarrollado, se consideró asignar un rol a los atributos (Año, Mes, Día, y Hora) de tal manera se obtenga una clasificación sobre la base de una temporalidad.

3.4.4.5 Componente de Algoritmo de Evaluación (X-Means)

Utilidad que realiza el tratamiento de la información mediante un algoritmo de minería de datos, en este componente se estableció el algoritmo X-Means basado en una variable "k" mínima de 2 y una variable "k" máxima 20.

3.4.4.6 Componente de Algoritmo de Evaluación (K-Means)

Realiza la agrupación de los datos basada en el algoritmo K-Means, en este intrumento de la herramienta de minería se estableció el algoritmo basado en una variable "k" de 4.

3.4.4.7 Componente de Algoritmo de Evaluación (DBSCAN)

Utensilio de la herramienta de minería de datos, donde se estableció el tratamiento de los datos por medio del algoritmo DBSCAN, con un valor de épsilon de 25 con un mínimo de puntos 10.

3.4.4.8 Generate Attributes (X-Means K-Means DBSCAN)

Establecidos para generar un atributo basado en la información que pueda contener otro atributo. En el modelo se usan para fijar los valores de cada registro de los algoritmos de evaluación que intervienen, tomando en consideración, el atributo clúster que se generó al procesar la información.

3.4.4.9 Join (Algoritmos)

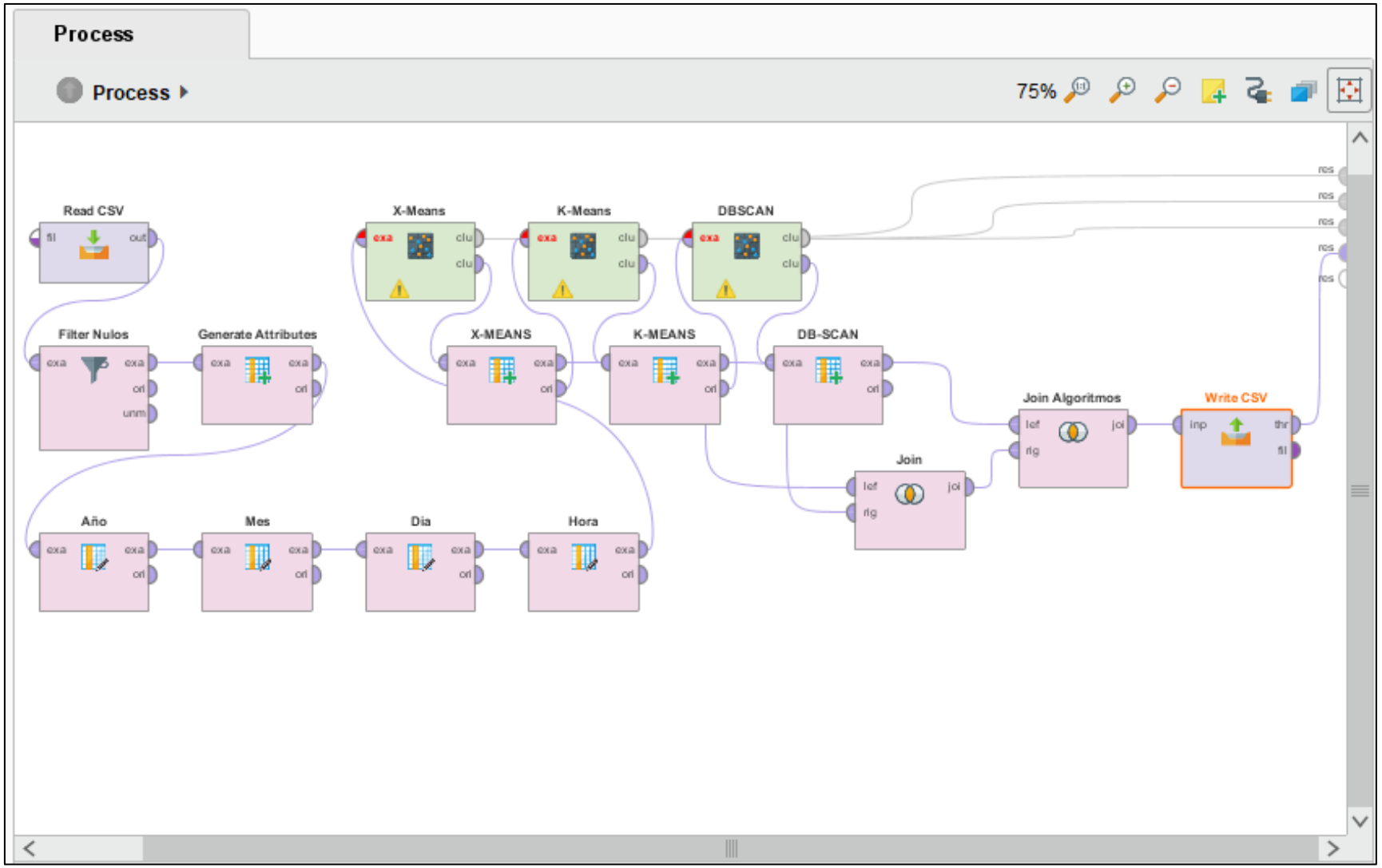
Esta herramienta se usó para agrupar toda la información del modelo, en donde constan todos los algoritmos de evaluación.

3.4.4.10 Write CSV

Escribe los datos resultantes, dentro de un archivo físico, externo al software, en formato CSV. Los datos del modelo de datos desarrollado se grabaron en el archivo "Evaluacion_Algoritmos.csv".

En la figura 16, se aprecia el modelo de datos que se realizó, en donde se describen todos los procesos realizados con las herramientas utilitarias de Rapidminer.

Figura 30
Modelado de datos



3.4.5 Evaluación

Para realizar una correcta evaluación del modelo se procedió a realizar una prueba, la misma que consta de una comparativa de los diferentes algoritmos de evaluación de la minería, su tiempo de ejecución y su efectividad según los requerimientos del objetivo.

En la tabla 3, se puede observar los diferentes algoritmos de evaluación y los resultados de su procesamiento de las dos experimentaciones descritas.

Tabla 3
Algoritmos de evaluación (datos 1 mes) - Cantidad de clústers y registros

Algoritmo	Número Clúster	Clúster 0	Clúster 1	Clúster 2	Clúster 3	Clúster 4	TOTAL
X-Means	4	920	419	1936	353		3628
K-Means	4	2300	413	659	256		3628
DBSCAN	5	681	2880	40	13	14	3628

Distribución de los clústers, número de clúster por algoritmo y cantidad de registros

En los resultados de la agrupación, se identifica una distribución general. Se observa una generalización por la naturalidad de los datos y las multidimensiones existentes. Sin embargo, con la evaluación descrita y la clasificación de los clústers en el set de datos resultante, se puede establecer un parámetro descriptivo útil para la percepción visual. En la tabla 4, se puede apreciar los diferentes algoritmos de evaluación y los resultados de su procesamiento de la segunda experimentación, con un espectro más amplio del número de registros previamente normalizados.

Tabla 4
Algoritmos de evaluación (datos 1 año) - Cantidad de clústers y registros

Algoritmo	Número Clúster	Clúster 0	Clúster 1	Clúster 2	Clúster 3	Clúster 4	TOTAL
X-Means	3	7761	14838	11631			34230
K-Means	4	7761	8594	8055	9820		34230
DBSCAN	4	27761	1032	652	4786		34230

Distribución de los clústers, número de clúster por algoritmo y cantidad de registros

3.4.6 Implantación

Luego de los procesos anteriores, en donde se realizó la construcción del modelo de gestión de datos, la información tratada en la presente fase se convierte en conocimiento.

Para que el usuario consiga una comprensión completa del conocimiento generado, se desarrolló una visualización intuitiva con gran dinamismo e interactividad, debido a que no se disponía de una herramienta altamente visual que facilite las observaciones del gestor, la misma que ha permitido aportar a la toma decisiones óptimas. En este desarrollo se pudo tratar los datos directamente en tiempo real dentro de la gráfica de la aplicación web; lo que permitió identificar los patrones relacionales del comportamiento de los datos variables de la contaminación del aire, permitiendo así proponer recomendaciones y operaciones con sustento en los resultados.

Esta sección se describe en detalle en el apartado “Visualización”, que permite desarrollar procesos y técnicas estructuradas de visualización, dentro de la implantación del ejemplo de verificación. Además, la metodología CRISP-DM utilizada para el desarrollo del ejemplo de verificación de este trabajo de graduación, propone crear un espacio de resultados, de tal manera que los hallazgos de la experimentación serán especificados dentro de este apartado en la mencionada área del documento.

3.5 Visualización

La visualización ha garantizado la efectividad de los estudios realizados en las diferentes áreas de estudio, para el área referente a la contaminación ambiental, la visualización ha sido clave para la comprensión y la interpretación de los datos y los intereses de propósito. Basados en la revisión de la literatura, la aproximación inicial a la técnica de visualización propuesta en este trabajo ha sido previamente difundida en el sistema AirVIS en el documento de Liao, Peng, Li, Liang, & Zhao, (2014).

Como se mencionó dentro de los requerimientos de la metodología implementada, la aplicación web combina tecnologías, protocolos e intérpretes, por lo que se consideró el lenguaje de marcado HTML y Sublime Text 3 para la edición de texto.

La representación de los datos y comportamientos se realizaron por medio de la técnica de visualización de coordenadas paralelas, a través de un proceso de implementación que permitió adaptarse al conjunto de datos; la herramienta D3.js engloba el uso de (CSS, SVG y DOM) para la interacción.

En esta sección, se presenta las especificaciones técnicas, instalación de herramientas de software, implementación del diseño y técnica de visualización, también la interfaz visual de forma completa y con un gran nivel de detalle para su interpretación.

3.5.1 Descripciones técnicas y herramientas de software

Las descripciones técnicas de la computadora en donde se realizó los procesos de visualización, se listan a continuación:

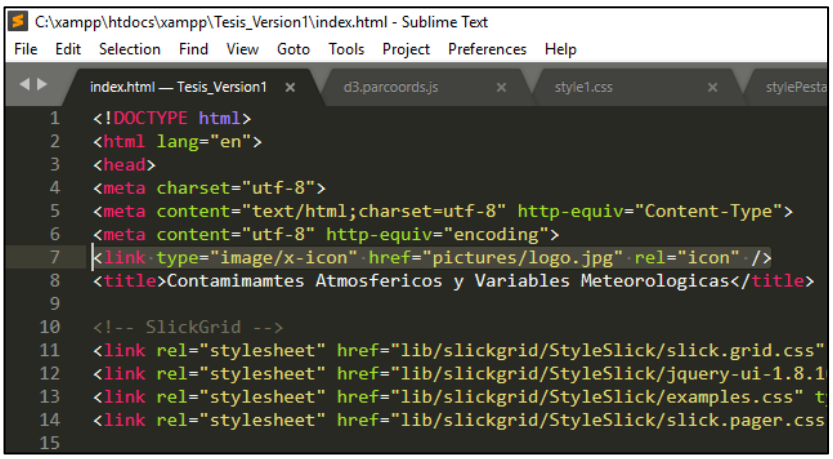
- Procesador: Intel(R) Core (TM) i5-4210U CPU @ 1.70 GHz.
- Memoria instalada (RAM): 6.00 GB.
- Sistema Operativo: Windows 10 Pro N © 2018 Microsoft Corporation.
- Tipo de Sistema: Sistema operativo de 64 bits, procesador x64.

Se instalaron las herramientas Sublime Text 3 y la librería D3.js gestionada a través del editor de texto para el gráfico de resultados del proceso, con 34.231 registros.

3.5.1.1 Sublime Text 3

Con el propósito del desarrollo web y la fácil edición del código fuente dentro del editor de texto, se usó esta herramienta de edición de texto. La instalación se realizó desde la página oficial de Sublime Text donde se seleccionó el archivo correspondiente que se adaptó a nuestro sistema operativo.

Figura 31
Entorno Sublime Text 3



A través del módulo Package Control se gestionó e instaló plugins[25] (ofrecidos como software libre), se realizó implementando el código fuente Python del módulo en la consola del editor de texto dentro del menú Ver > Mostrar (Package Control, 2019).

[25] Plugin: componente de código, para ampliar las funciones de una herramienta.

Figura 32
Código Package Control

SUBLIME TEXT 3

```
import urllib.request,os,hashlib; h =
'6f4c264a24d933ce70df5dedcf1dcaee' +
'ebe013ee18cced0ef93d5f746d80ef60'; pf = 'Package
Control.sublime-package'; ipp =
sublime.installed_packages_path();
urllib.request.install_opener( urllib.request.build_opener(
urllib.request.ProxyHandler()) ); by = urllib.request.urlopen(
'http://packagecontrol.io/' + pf.replace(' ', '%20')).read();
dh = hashlib.sha256(by).hexdigest(); print('Error validating
download (got %s instead of %s), please try manual install' %
(dh, h)) if dh != h else open(os.path.join( ipp, pf), 'wb'
).write(by)
```

Fuente: (Package Control, 2019)

3.5.1.2 D3.js

Para el desarrollo de la parte dinámica del gráfico se usó la librería D3.js, librería que permite la interacción y fluctuación dinámica de los valores del set de datos final. La instalación se realizó con la descarga del archivo "d3.zip", desde la página oficial de la librería versión (5.9.7), este archivo se colocó dentro de la carpeta que almacena todos los archivos necesarios para la aplicación web y se direcciona desde el "index.html", el script[26] que incluye a la librería, para su uso.

También, se lo puede direccionar hacia un archivo almacenado en la nube con un script disponible en la página oficial de D3 (d3js.org, 2019). Algunas librerías que se usó para el desarrollo visual fueron: d3.min.js, d3.parcoords.js, d3.svg.multibrush.js, divgrid.js, jquery.min.js, entre otras.

Figura 33
Script vinculación de librería D3.js

```
<script src="https://d3js.org/d3.v5.min.js"></script>
```

Fuente: (d3js.org, 2019)

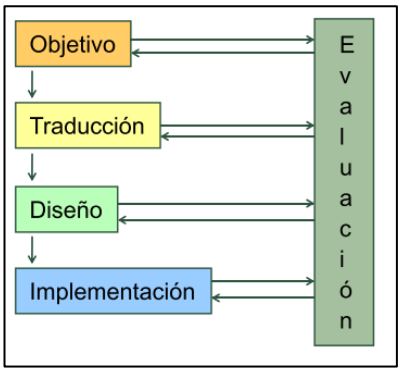
[26] Script: archivo de órdenes de procesamiento de texto plano.

3.5.2 Desarrollo e implementación del diseño de visualización

Dentro del desarrollo y diseño de la visualización, según la revisión literaria y basados en las analogías de H. Pfister & Miriah Meyer citadas por (Castro & Luján-Ganuza, 2017c), se implementa el esquema descrito en la figura 34.

Figura 34

Diseño de visualización



Fuente: (Castro & Luján-Ganuza, 2017c)

3.5.2.1 Objetivo

Se planteó como objetivo la interacción del gestor ambiental, donde se pudo desarrollar tareas de dominio; se analizó subconjuntos de datos de los contaminantes y variables meteorológicas comprendidos en temporalidades y conformes a condiciones específicas de selección o cepillado (brush), se identifica específicamente los registros en la limpieza de los datos que se visualizan.

3.5.2.2 Traducción

La traducción de los datos se implementó a través de dimensiones paralelas multidimensionales, basados en la técnica de visualización de coordenadas paralelas y las amplias funcionalidades de interacción.

3.5.2.3 Diseño

Se consideró la codificación visual con aspectos colorimétricos de orden y percepción, se realizó el modelo apropiado que facilite toda la información del detalle de los registros que intervienen en el análisis atmosférico, además evaluaciones estadísticas para fácil comprensión.

3.5.2.4 Implementación

Se tomó en cuenta aspectos generales de la interfaz, y se realizó prototipos y diseños de interacción con algunas herramientas útiles para un minado gráfico, conjuntamente con la aprobación del gestor y del director del trabajo de titulación.

3.5.2.5 Evaluación

Todas estas tareas interactuaron con una parte de evaluación experimental, esto facilitó la retro alimentación del proceso, pudiendo ser flexible para cambios de desarrollo dentro de todas etapas. El modelo del diseño permitió flexibilidad para cumplir con el propósito principal.

3.5.3 Desarrollo e implementación de la técnica de visualización

La técnica de visualización que se implementó fue la técnica de coordenadas paralelas, siendo altamente descriptiva y aplicada en varios estudios según la revisión literaria. Una técnica óptima debido a su gran integración y comprensión. La técnica se adaptó fácilmente a los datos multidimensionales típicos que determinan la calidad del aire.

Las dimensiones que la gráfica dinámica contiene son 21 paralelas en total. Se cuenta con 15 paralelas que se refieren a las dimensiones de los contaminantes atmosféricos y variables meteorológicas. También, cuatro dimensiones que indican los aspectos temporales, por ejemplo: año, mes, día, y hora; adicional se integró dos dimensiones, de los códigos de los registros y el identificador del clúster en donde se encontraba contenido el dato.

El potente análisis multidimensional permitió que las dimensiones sean representadas por ejes verticales paralelos; y cada registro se representó por una línea poligonal, esta línea interceptó cada paralela según el valor de cada dimensión respectiva. De lo expuesto se pudo detectar, correlaciones entre diferentes contaminantes, variables meteorológicas y el impacto de la concentración de cada contaminante en el ICA.

Para evitar el desorden visual la codificación de color de las líneas se corresponde con la interpolación, basada en la selección de una dimensión generalmente clasificada por los clústers (Liao et al., 2014).

La visualización de las coordenadas paralelas permite múltiples interacciones. Se puede realizar la selección de segmentos, y multisegmentación de los datos. También, permite el reordenamiento de unidades a lo largo del eje de cada dimensión, pudiendo ser un orden ascendente o descendente según se requiera. El desplazamiento de la paralela de las dimensiones a lo largo de la gráfica, pudiendo ser modificado el orden y juntar variables de análisis que en primera instancia estén separadas por paralelas intermedias. Los barridos de filtrado, permiten extraer parte de los datos de una manera dinámica e intuitiva.

Los filtros mediante cepillado (brush), son un conjunto de métodos dinámicos. Al realizar un cepillado en la gráfica de resultados se puede obtener un resaltado, una sombra, modos de pintura y algunos otros efectos que diferencian los datos seleccionados de todos los registros. Por medio de un rectángulo llamado el pincel, que se dibuja con el ratón, se puede visualizar los cambios en tiempo real de las específicas colecciones generadas. La gráfica contiene funciones técnicas de pincel como: enlace único y agrupamiento (2D-strums), condiciones en una variable (1D-axes), condiciones en dos variables (1D-axes-multi), subconjuntos de variables y serie de tiempo (angular) (Becker & Cleveland, 1987).

El objetivo es facilitar los análisis de patrones y encontrar comportamientos por parte del gestor de estos datos; todas estas funciones se implementaron en la gráfica del trabajo para que sean útiles y de sencillo manejo. Por ejemplo, en una prueba se pudo ajustar el orden de las coordenadas paralelas de dimensiones relacionadas, donde se posicionó a los ejes paralelos de forma adyacente para mejorar percepción visual entre las dimensiones.

3.5.4 Interfaz Visual

La interfaz visual se expresó en una gráfica de coordenadas paralelas, con 21 dimensiones de ejes verticales paralelos de los diversos materiales presentados en la atmósfera. Además, dimensiones que permiten los análisis basados en el tiempo y paralelas de los identificadores de cada registro de la información.

La gráfica contiene 34.231 registros de polilínea conectados entre los ejes dimensionales. También cuenta con parámetros intuitivos visuales (posición, orientación, zoom y filtros). La representación interactiva dentro de los parámetros visuales categorizó los aspectos de colorimetría, por medio de una selección de la dimensión; en primera instancia se consideró la agrupación de los clústers para este aspecto (Ver figura 35).

Además, se tomó en cuenta parámetros de datos (mapeo de los datos, resaltado y sombras de los grupos de datos), lista y detalle de registro (resaltado de dato único, cepillado o brush, y configuración). La aplicación web cuenta con distribución de secciones en la pantalla:

3.5.4.1 Encabezado

En esta sección se encuentra el logo de la universidad, el tema de resumen de la gráfica presenta y el departamento de desarrollo (LIDI), que impulsó la investigación.

3.5.4.2 Gráfica

Esta sección contiene el resumen del set de datos y las características mencionadas anteriormente por la gráfica de coordenadas paralelas.

3.5.4.3 Barra de estado

En esta sección, se encuentra algunas funciones:

Botones de filtrado

Keep (permite un acercamiento o zoom a la selección de los datos por el pincel), Exclude (excluye datos de que se encuentran seleccionados por el pincel) y Export (permite extraer el set de datos).

Selector de colorimetría

Dimensión a la que se basa el color de los registros.

Botón de descarga

Permite descargar la gráfica en un archivo de extensión jpg.

Botón de exportar datos

Permite descargar los datos un archivo de extensión csv.

Contador

Permite visualizar el estado de carga y el número de registros.

Porcentaje de opacidad

Permite visualizar el porcentaje de opacidad de las líneas.

Botones de Brillo

Ocultar Valor Ejes / Mostrar Valor Ejes (permite ocultar o mostrar las unidades de los ejes), Oscuro/Claro (permite seleccionar el fondo de pantalla para realce visual).

Figura 35

Gráfica de Visualización en modo Claro.

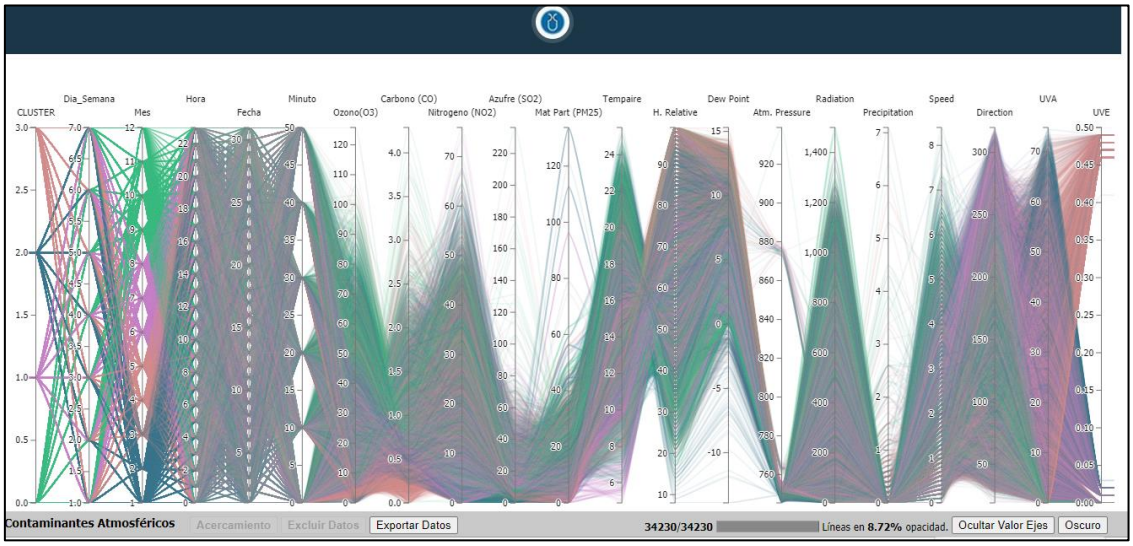


Figura 36

Gráfica de Visualización en modo Oscuro.

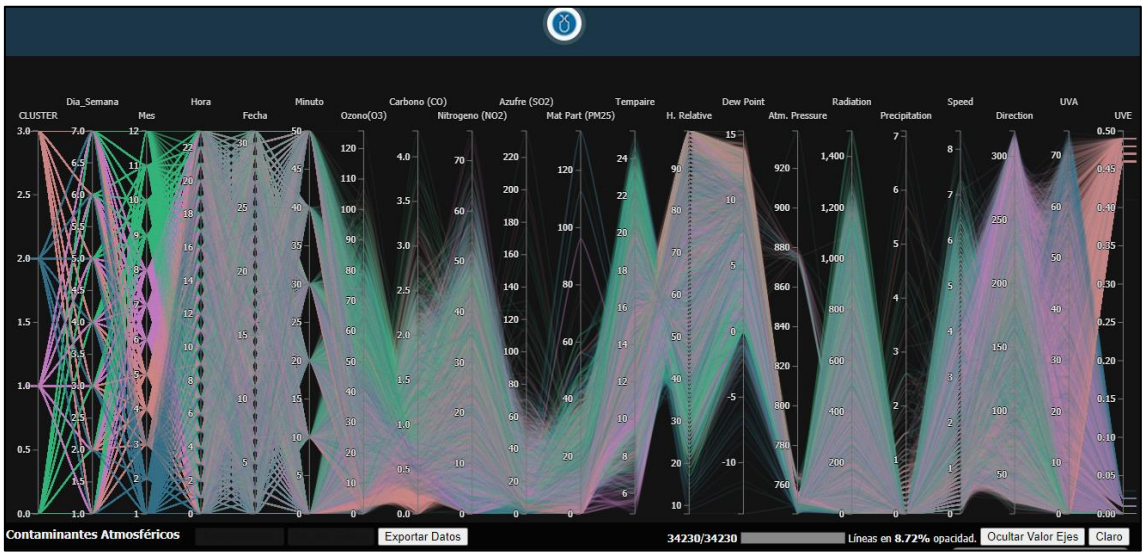
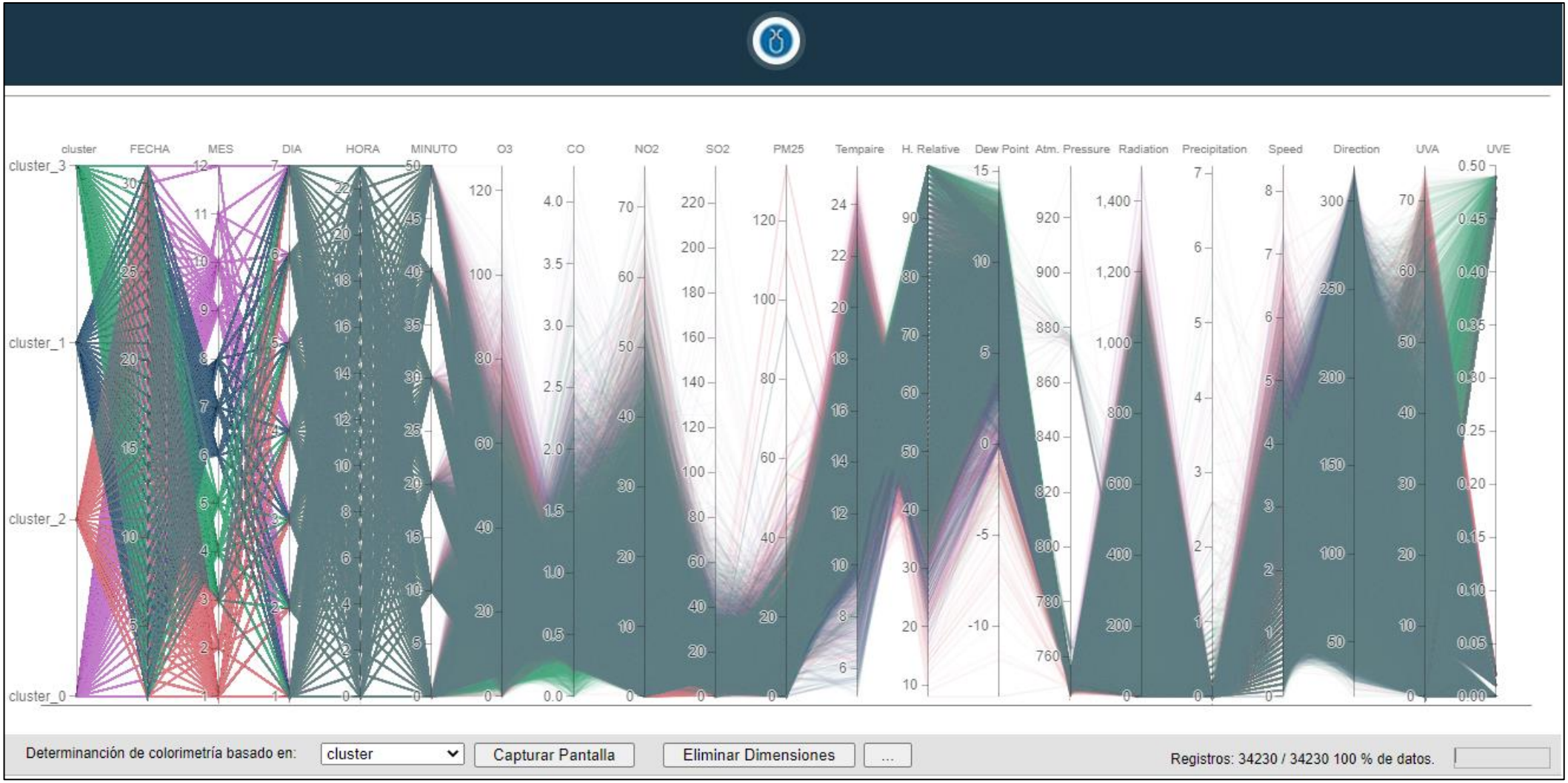


Figura 37

Gráfica Coordenadas Paralelas - Contaminantes Atmosféricos y Variables Meteorológicas



3.5.4.4 Menú gráfico de complementos visuales

En esta sección, se encuentra algunas funciones que complementan la visualización:

3.5.4.4.1 Opción Detalle Lista

Opción permite visualizar en la parte inferior de la gráfica de coordenadas paralelas la información detallada del registro, por medio de una grilla. Si se ubica el puntero del ratón sobre el registro, se permite visualizar ese la línea registro de forma resaltada, además, la grilla permite ordenar los valores en sentido ascendente y descendente.

3.5.4.4.2 Gráfica de barras

Contiene dos selectores, el primero se refiere a las dimensiones presentes en el gráfico general, y el segundo a la temporalidad; al realizar estas selecciones se genera una gráfica de barras lateral con los valores máximos y mínimos distribuidos por la temporalidad y la dimensión selectas.

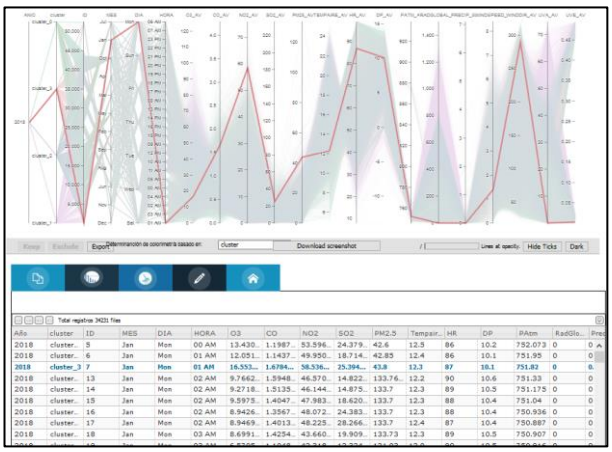
3.5.4.4.3 Cepillado (Brush)

Opción contiene dos selectores, el primero se refiere al modo de brush o cepillado que realiza el filtrado, y el segundo selector permite anexar información al filtro, por medio de los operadores lógicos, OR o AND. También, contiene un botón de reseteo de brush y la información del porcentaje de información contenida en el filtrado.

3.5.4.4.4 Configuración y Créditos

Opción que permite seleccionar el algoritmo de agrupación de evaluación del gráfico (X-Means, K-Means y DBSCAN). La pestaña “Creditos” contiene la información del laboratorio de desarrollo, tema de investigación y autor de la aplicación.

Figura 38
Gráfico de complementos visuales - Opción Detalle Lista

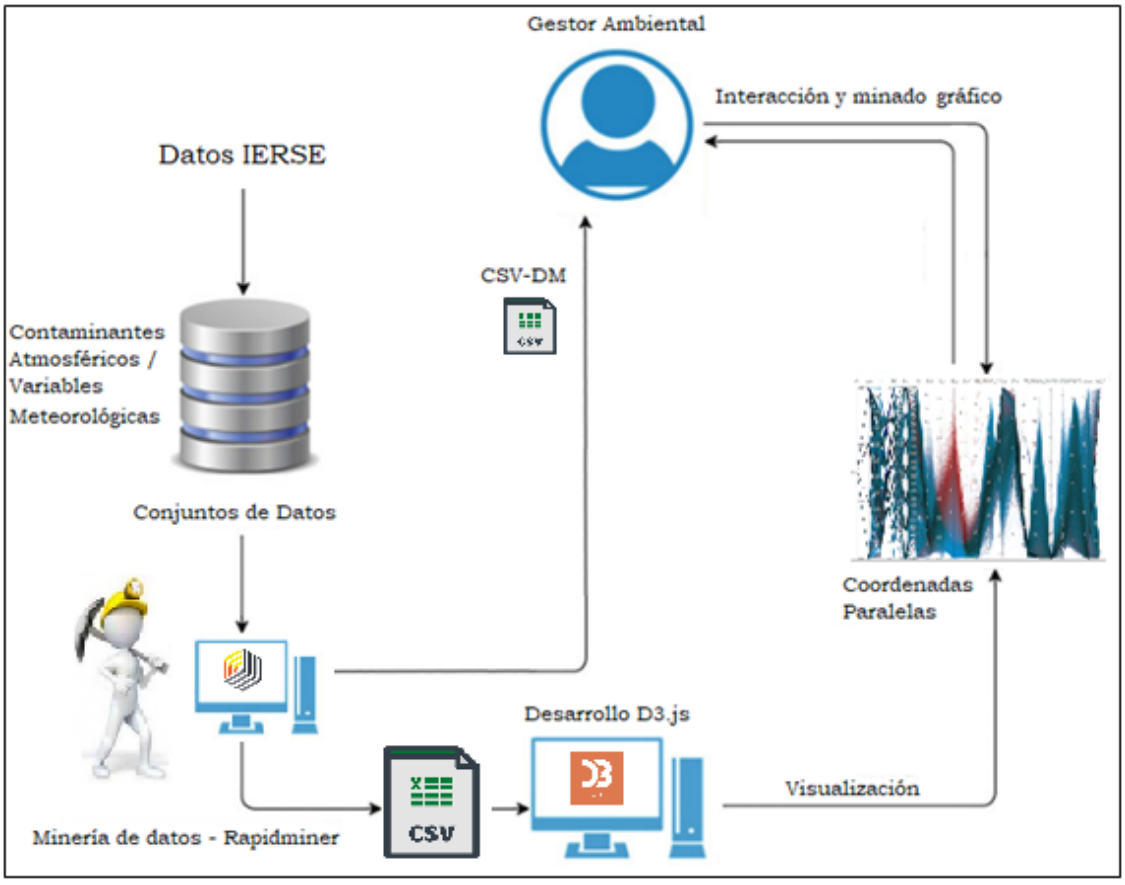


3.5.5 Esquema General Implementado

Al cumplir el desarrollo e implementación del diseño visual, la técnica de visualización y la interfaz de gestión, el sistema desarrollado corresponde al esquema general descrito en la figura 37. Aquí, se describe de forma general todo el proceso que se generó dentro de la metodología CRISP-DM; así también, dentro de la esquematización y desarrollo de la visualización altamente interactiva.

Figura 39

Esquema General Minería y Visualización de los Contaminantes Atmosféricos y Variables Meteorológicas (IERSE)



4. RESULTADOS

Luego de realizar las diferentes etapas del procesamiento y evaluación de los datos de contaminantes y variables atmosféricas con la metodología CRISP-DM y los diferentes algoritmos de evaluación (X-Means, K-Means y DBSCAN), se logró representar los resultados con un método de visualización avanzada, que permite la interacción y la representación de las distintas asociaciones entre los contaminantes y las variables meteorológicas. El presente trabajo generó los resultados por medio del diseño y la técnica de visualización de coordenadas paralelas, se consiguió determinar el comportamiento de las distintas dimensiones consideradas en el estudio.

Se analizó los esquemas resultantes de los diferentes valores, agrupados, filtrados y depurados dentro de la misma aplicación web en la sección del gráfico de las coordenadas paralelas. Se aplicó algunas alternativas de control (listado de detalles de registro, aspectos colorimétricos, barrido de información seleccionada, etc.). La comprensión y entendimiento de dichas relaciones es efectivamente entendida e interpretada por el gestor ambiental. El uso de la aplicación web de la colección de los datos analizados, realizó minería o filtrados de datos desde el propio gráfico, lo que permitió verificar las singularidades específicas que se presentaron entre los contaminantes atmosféricos y las variables meteorológicas.

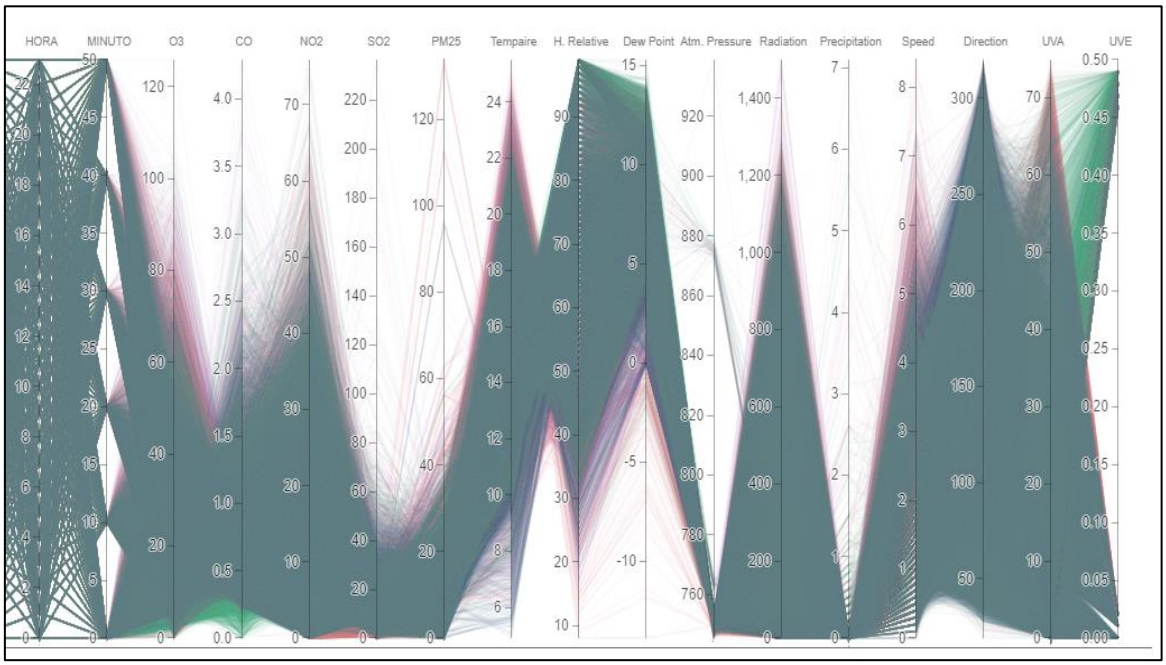
4.1 Visualización General

Al explorar de manera general la aplicación web, en la sección del gráfico, se determinó una lógica de correspondencia y relación entre dimensiones descritas, se pudo observar que los valores dimensionales de los contaminantes atmosféricos y variables meteorológicas tuvieron integridad con respecto a los estudios previos.

En la figura 40, se observa que la mayor cantidad de registros del ozono (O3) se ubican en un rango entre 10 y 60 partes por billón (ppb) correspondiendo al 68% de registros, el monóxido de carbono (CO) se concentra entre 0.5 y 2.0 partes por millón (ppm) con el 73 % de los registros, el dióxido de nitrógeno (NO2) se concentra entre 10 y 40 partes por billón (ppb) con el 69% de datos; así también, el dióxido de azufre (SO2) se concentra entre 0 y 40 partes por billón (ppb) que corresponde al 99% de los registros; y el material particulado (PM2.5) se encuentra entre 0 y 30 microgramos / metro cúbico (ug/m3), correspondiente al 95% de los datos totales.

Figura 40

Resultados Generales



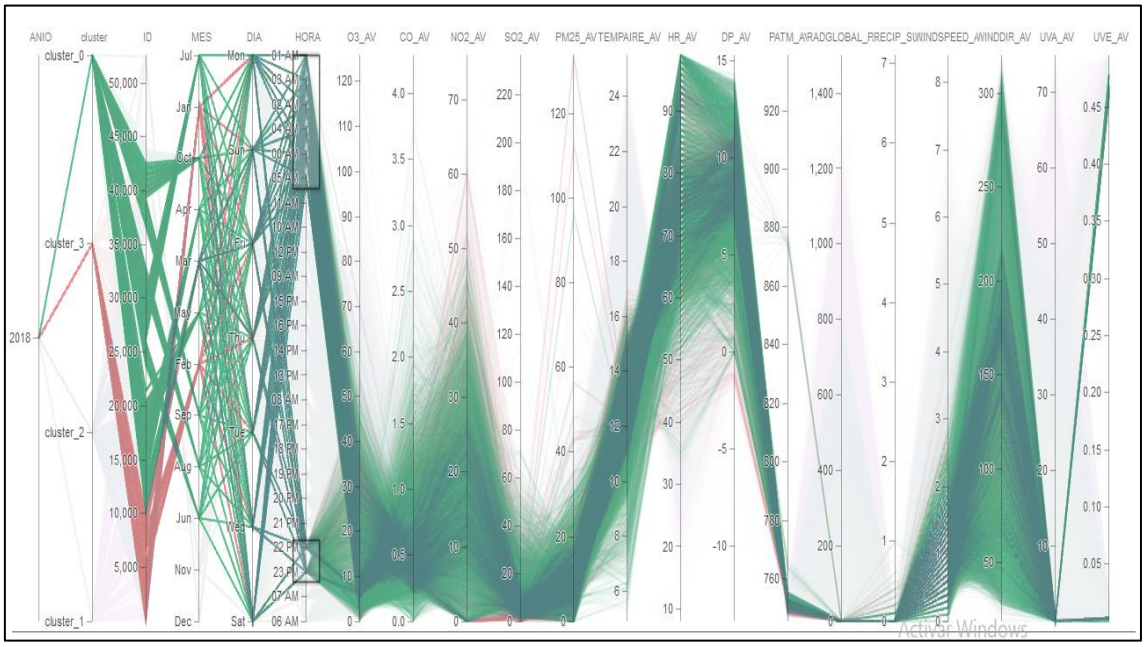
Por parte de las variables meteorológicas que en el gráfico también las aborda; la temperatura (TEMPAIRE) concentra el 84% de los datos entre los valores de 12 y 22 grados centígrados (°C), la humedad relativa (HR) comprende el 72% de sus valores entre 45 y 85 de porcentaje de agua, el punto del rocío (DP) engloba el 86% de los datos entre los valores de 5 y 12 °C. La presión atmosférica (PATM) entre 745 y 755 hecto pascales (hPa) correspondiente al 99% de los datos, en cuanto a la radiación global (RADGLOBAL) su distribución se encuentra entre 0 y 600 vatios por metro cuadrado (w/m2) que corresponde al 88% de los datos. La precipitación (PRECIP) contiene sus valores entre 0 y 0.3 milímetros de agua. La velocidad del viento (WINDSPEED) comprende sus valores entre 1 y 6 metros / segundos (m/s) lo que corresponde al 73% de los datos, la dirección del viento (WINDDIR) se encuentra entre 100 y 250 grados (°). La radiación UVA tiene el 90% de datos que se encuentran entre 0 y 35 w/m2 y la radiación UVE concentra sus valores entre 0 y 0.05 w/m2 con el 82% de los registros. Con este análisis general se puede buscar relaciones y asociaciones específicas mediante el cepillado o filtrado, esto permite encontrar los patrones de comportamiento; por ejemplo, los valores entre 0 y 10 ppb del ozono (O3), tienen una correlación con los valores entre 70 y 99 de porcentaje de agua de la humedad relativa (HR).

4.2 Temporalidad

Tomando en consideración una evaluación visual que se basa en la temporalidad, se planteó un escenario donde se pudo observar que en horas de la madrugada el índice de contaminación se reduce. En la figura 41 se puede observar un patrón de comportamiento en relación con la selección de horas comprendidas entre las 10:00 pm y 05:00 am, donde el porcentaje concentrado de datos, corresponde al 29,78% con un total de 10.197 registros.

Figura 41

Comportamiento de variables – Temporalidad de horas (madrugada)

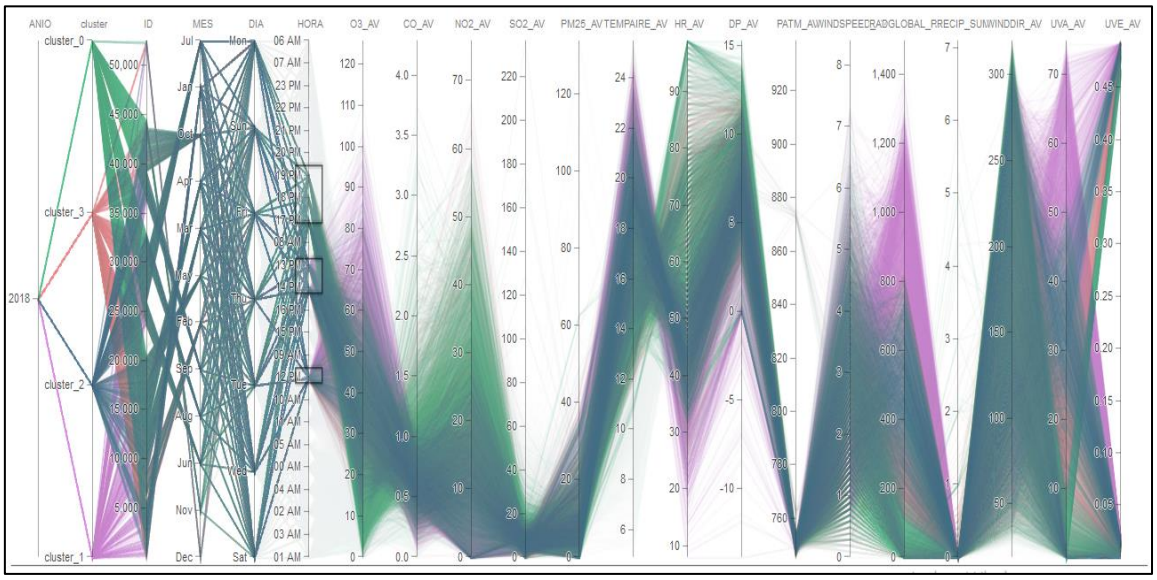


El patrón lógico de comportamiento de los contaminantes atmosféricos ozono (O3), monóxido de carbono (CO), dióxido de nitrógeno (NO2), dióxido de azufre (SO2), y material particulado (PM2.5), determinó que dichas dimensiones, en estas horas comprendidas en este rango de la madrugada, tuvieron la tendencia a reducir sus valores. Se consideró como aspecto relevante que en estas horas no existe la fluidez vehicular, tampoco existe la labor industrial en exceso, y aquellos serían los factores que afectan directamente al incremento de los índices de la contaminación ambiental. No obstante, se observó que los índices de ciertas variables meteorológicas (HR, DP), incrementaron dentro de esta temporalidad, lo que mostró una relación inversamente proporcional a las variables inicialmente descritas, quizá por aspectos climatológicos que se caracterizan de manera común a estas horas dentro de la ciudad de Cuenca.

En la figura 42 se observa un patrón de comportamiento según la selección de horas pico, escenario de análisis con horas comprendidas entre el rango de: 12:00 am a 14:00 pm, y 17:00 pm a 19:00 pm. Coincidentemente son las horas de mayor fluctuación de tráfico dentro de la ciudad de Cuenca, donde el porcentaje concentrado de datos corresponde al 26,38% con un total de 9.032 registros de datos.

Figura 42

Comportamiento de variables – Temporalidad de horas (horas pico)

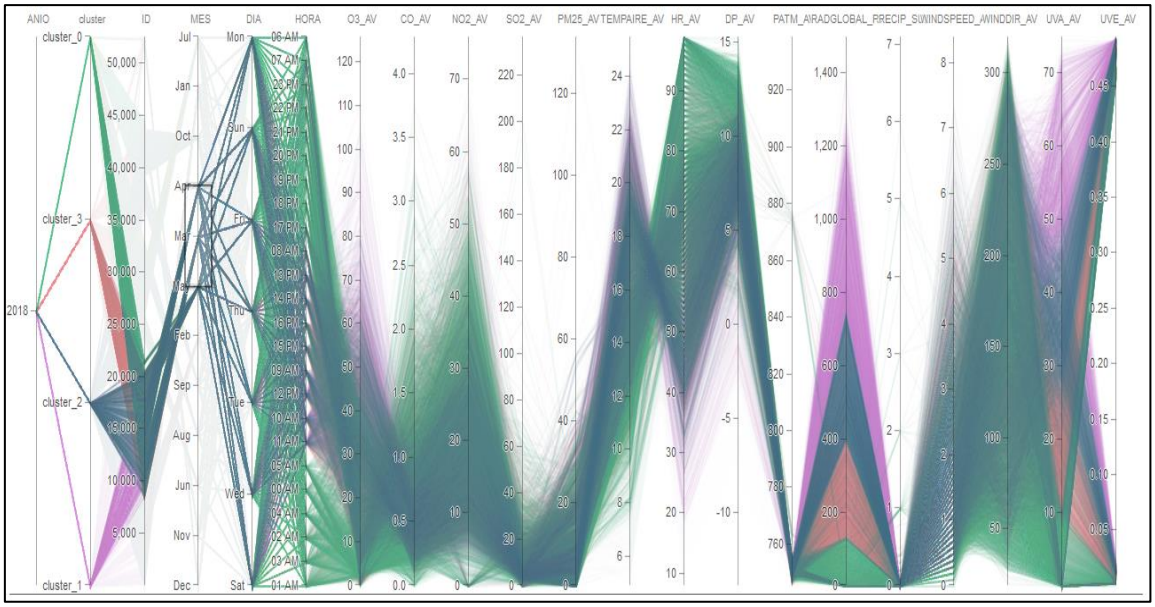


La figura 42 se visualiza, el comportamiento de los contaminantes atmosféricos: ozono (O3), monóxido de carbono (CO), y dióxido de nitrógeno (NO2), encontrando que dichas dimensiones en estas horas pico comprendidas dentro del rango señalado, tuvieron tendencia a incrementar sus valores, se consideró como aspecto de gran relevancia la fluidez vehicular, la gestión industrial, entre otros aspectos, que afectan directamente al incremento de los índices de la contaminación ambiental en la ciudad. Por otra parte, el dióxido de azufre (SO2) y el material particulado (PM2.5) conservaron sus valores de concentración como en el análisis general. Sin embargo, también se determinó que ciertas variables meteorológicas como: la humedad relativa (HR), el punto de rocío (DP), y la temperatura (TEMPAIRE), incrementaron sus índices dentro de estas condiciones, se visualiza una relación directamente proporcional con los aspectos climatológicos que se caracterizan de manera común a estas horas dentro de la ciudad de Cuenca.

En otro escenario, se observa temporalidades según las estaciones climatológicas comprendidas en trimestres. En la figura 43, se identifica un patrón de comportamiento con base a la selección de una temporalidad estacionaria de primavera, comprendida entre los meses de marzo, abril y mayo, donde el porcentaje concentrado de datos fue correspondiente al 31,87% con un total de 10.911 registros.

Figura 43

Comportamiento de variables – Temporalidad de meses (primavera)



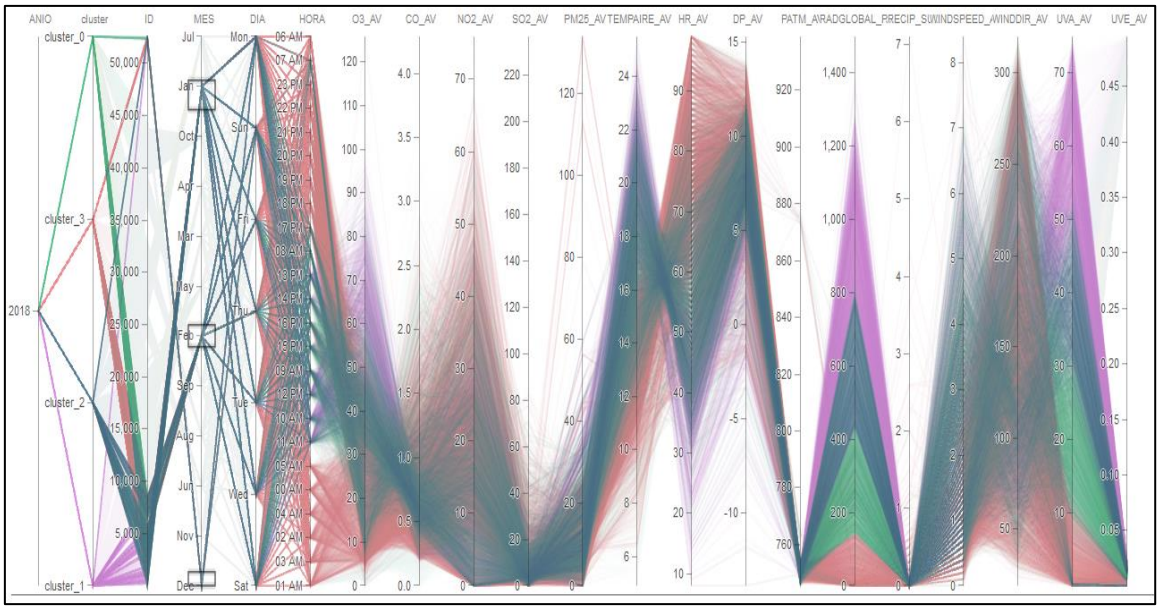
En este análisis, el comportamiento del ozono (O3) determinó como resultado que el 75% de sus registros de valores se encontraron entre 0 y 30 partes por billón (ppb) en los meses comprendidos para la primavera, lo que significó que los niveles de ozono son bajos, esto como una observación general dentro del escenario planteado.

Por otro lado, también se encontró que las variables meteorológicas como: humedad relativa (HR), punto de rocío (DP), alcanzaron los índices de sus valores como los más altos, es decir que para la humedad relativa (HR), el 69% de los datos filtrados se agruparon entre valores de 60 a 100 de porcentaje de agua, y para el punto de rocío (DP) se comprendieron los valores entre 0 y 10 °C, lo que fue correspondiente al 70% de los registros analizados; se encontró en síntesis, que estos aspectos climatológicos se caracterizan por el incremento de lluvias dentro de este escenario.

En la figura 44, se observa un patrón de comportamiento con base a la selección de una temporalidad estacionaria de invierno, comprendida entre los meses de diciembre, enero y febrero, donde el porcentaje concentrado de datos corresponde al 21,18% con un total de 7.252 registros de los datos.

Figura 44

Comportamiento de variables – Temporalidad de meses (invierno)



El comportamiento del ozono (O3) determinó que el 42% de este contaminante se encuentra entre 30 y 90 partes por billón (ppb) en estos meses. El 35% de valores filtrados para el contaminante dióxido de nitrógeno (NO2) se encontró entre 20 y 50 partes por billón (ppb); el dióxido de azufre (SO2) y el material particulado (PM2.5), conservaron sus valores de concentración como en el análisis general.

También, en la evaluación se pudo determinar que las variables meteorológicas: temperatura (TEMPAIRE), radiación ultravioleta (UVA), dirección del viento (WINDDIR), y radicación global (RADGLOBAL), concentraron sus valores entre rangos de índices altos.

4.3 Clústers

Otro aspecto de evaluación para obtener resultados fue la cantidad de datos contenidos en los clústers de agrupamiento. Este análisis se lo realizó con dependencia del algoritmo de evaluación seleccionado, permitió verificar el porcentaje de datos y la distribución en los clústers que fueron considerados dentro de la gráfica, lo que facilitó la validación de las cantidades de datos, es decir, se determinó el clúster que contiene la mayor cantidad de datos y su importancia.

En la figura 45, se observa, el porcentaje de datos en el “cluster_0” seleccionado, este agrupa a 17.760 registros, que corresponden a un 51,88 % de los datos totales, el gráfico permite apreciar la relación entre las diferentes dimensiones dentro del clúster.

Figura 45

Selección de clúster de agrupamiento (clúster 0)

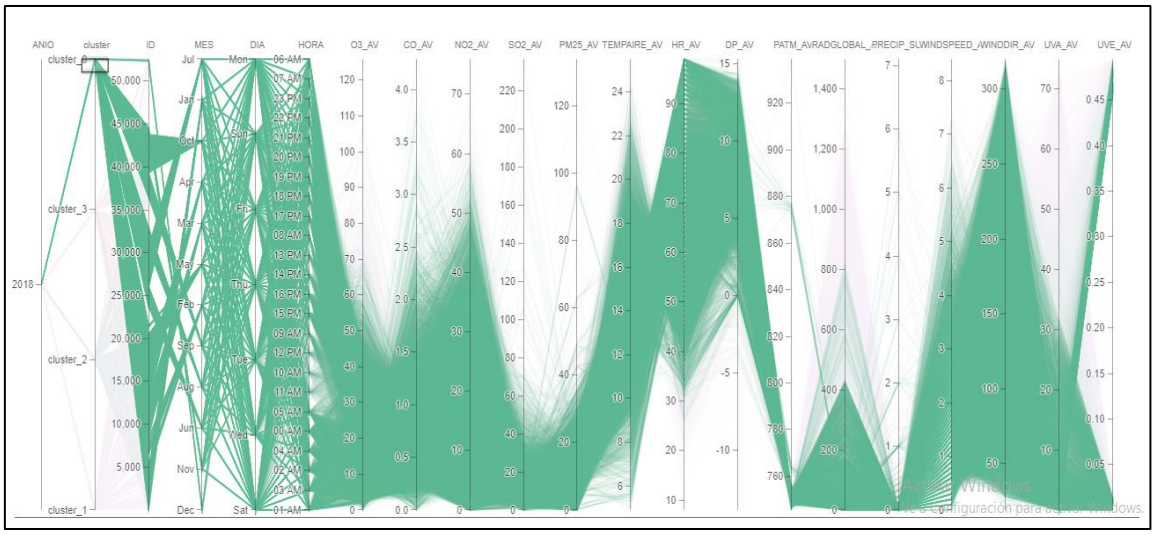
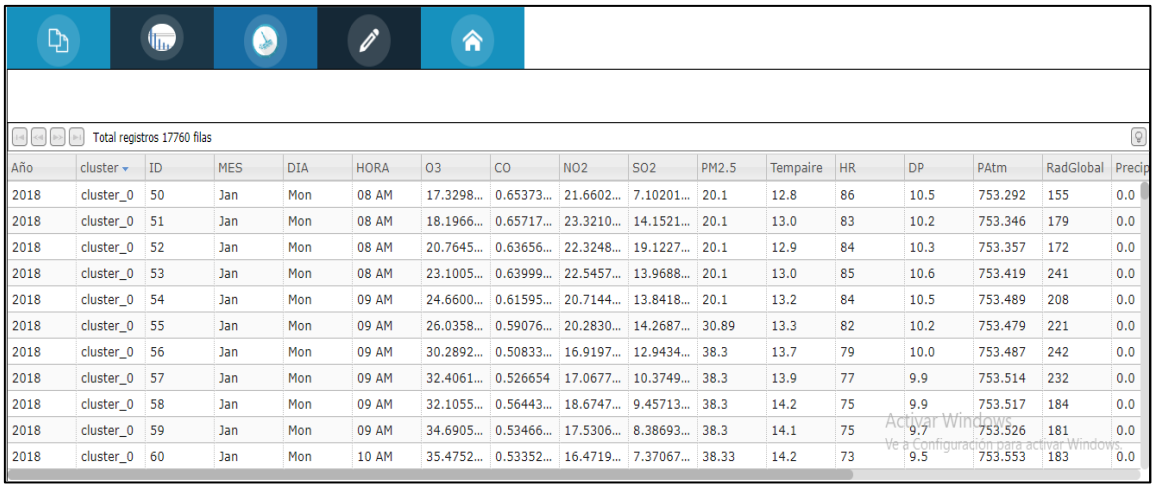


Figura 46

Listado en relación con la selección de agrupamiento (clúster 0)



Total registros 17760 filas

Año	cluster	ID	MES	DIA	HORA	O3	CO	NO2	SO2	PM2.5	Tempaire	HR	DP	PAtm	RadGlobal	Precip
2018	cluster_0	50	Jan	Mon	08 AM	17.3298...	0.65373...	21.6602...	7.10201...	20.1	12.8	86	10.5	753.292	155	0.0
2018	cluster_0	51	Jan	Mon	08 AM	18.1966...	0.65717...	23.3210...	14.1521...	20.1	13.0	83	10.2	753.346	179	0.0
2018	cluster_0	52	Jan	Mon	08 AM	20.7645...	0.63656...	22.3248...	19.1227...	20.1	12.9	84	10.3	753.357	172	0.0
2018	cluster_0	53	Jan	Mon	08 AM	23.1005...	0.63999...	22.5457...	13.9688...	20.1	13.0	85	10.6	753.419	241	0.0
2018	cluster_0	54	Jan	Mon	09 AM	24.6600...	0.61595...	20.7144...	13.8418...	20.1	13.2	84	10.5	753.489	208	0.0
2018	cluster_0	55	Jan	Mon	09 AM	26.0358...	0.59076...	20.2830...	14.2687...	30.89	13.3	82	10.2	753.479	221	0.0
2018	cluster_0	56	Jan	Mon	09 AM	30.2892...	0.50833...	16.9197...	12.9434...	38.3	13.7	79	10.0	753.487	242	0.0
2018	cluster_0	57	Jan	Mon	09 AM	32.4061...	0.526654	17.0677...	10.3749...	38.3	13.9	77	9.9	753.514	232	0.0
2018	cluster_0	58	Jan	Mon	09 AM	32.1055...	0.56443...	18.6747...	9.45713...	38.3	14.2	75	9.9	753.517	184	0.0
2018	cluster_0	59	Jan	Mon	09 AM	34.6905...	0.53466...	17.5306...	8.38693...	38.3	14.1	75	9.7	753.526	181	0.0
2018	cluster_0	60	Jan	Mon	10 AM	35.4752...	0.53352...	16.4719...	7.37067...	38.33	14.2	73	9.5	753.553	183	0.0

Detalle de los valores de las dimensiones, valores de clasificación por temporalidad (mes, día, hora).

En la figura 47, el porcentaje contenido en el “cluster_1” seleccionado corresponde a un 7,85% de los datos totales, agrupando a 2.688 registros, el gráfico permite apreciar la correspondencia entre las diferentes dimensiones, dentro de este clúster.

Figura 47

Selección de clúster de agrupamiento (clúster 1)

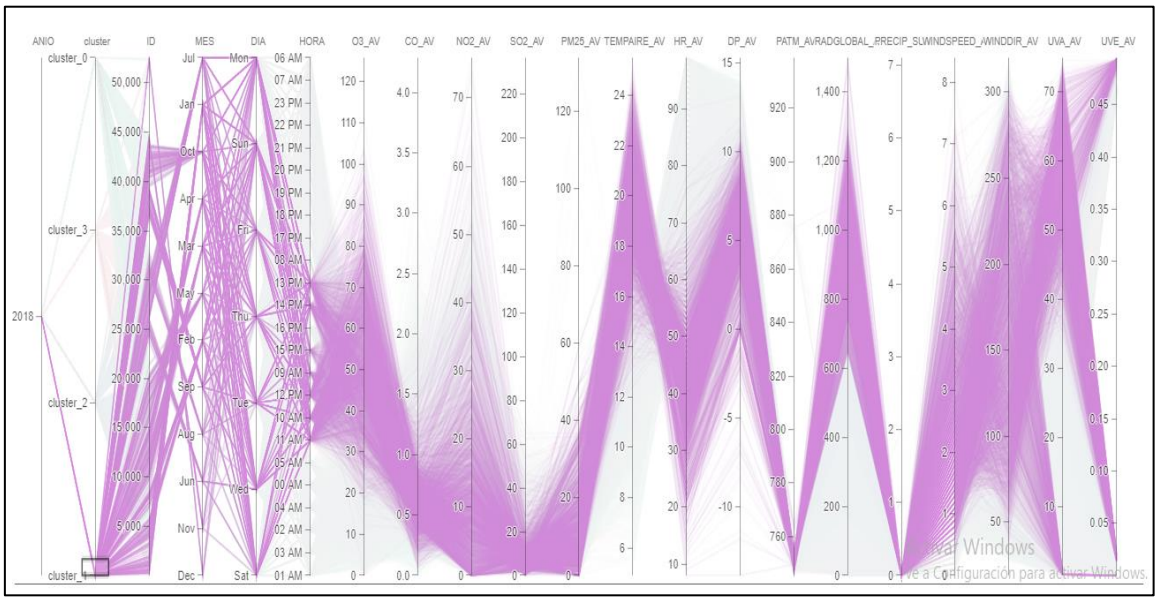
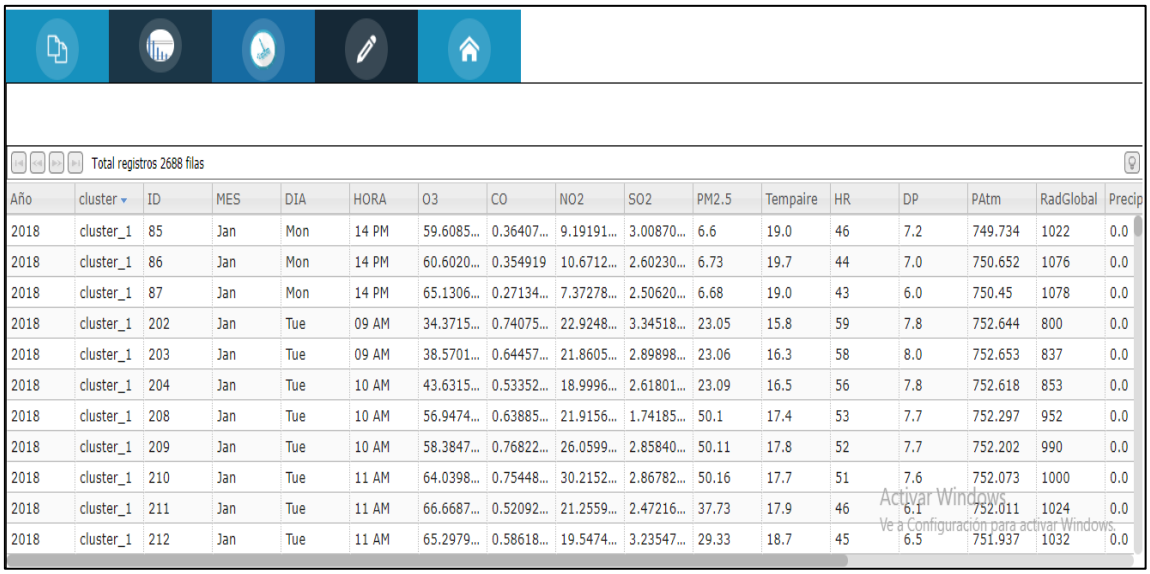


Figura 48

Listado en relación con la selección de agrupamiento (clúster 1)



Total registros 2688 filas

Año	cluster	ID	MES	DIA	HORA	O3	CO	NO2	SO2	PM2.5	Tempaire	HR	DP	PAtm	RadGlobal	Precip
2018	cluster_1	85	Jan	Mon	14 PM	59.6085...	0.36407...	9.19191...	3.00870...	6.6	19.0	46	7.2	749.734	1022	0.0
2018	cluster_1	86	Jan	Mon	14 PM	60.6020...	0.354919	10.6712...	2.60230...	6.73	19.7	44	7.0	750.652	1076	0.0
2018	cluster_1	87	Jan	Mon	14 PM	65.1306...	0.27134...	7.37278...	2.50620...	6.68	19.0	43	6.0	750.45	1078	0.0
2018	cluster_1	202	Jan	Tue	09 AM	34.3715...	0.74075...	22.9248...	3.34518...	23.05	15.8	59	7.8	752.644	800	0.0
2018	cluster_1	203	Jan	Tue	09 AM	38.5701...	0.64457...	21.8605...	2.89898...	23.06	16.3	58	8.0	752.653	837	0.0
2018	cluster_1	204	Jan	Tue	10 AM	43.6315...	0.53352...	18.9996...	2.61801...	23.09	16.5	56	7.8	752.618	853	0.0
2018	cluster_1	208	Jan	Tue	10 AM	56.9474...	0.63885...	21.9156...	1.74185...	50.1	17.4	53	7.7	752.297	952	0.0
2018	cluster_1	209	Jan	Tue	10 AM	58.3847...	0.76822...	26.0599...	2.85840...	50.11	17.8	52	7.7	752.202	990	0.0
2018	cluster_1	210	Jan	Tue	11 AM	64.0398...	0.75448...	30.2152...	2.86782...	50.16	17.7	51	7.6	752.073	1000	0.0
2018	cluster_1	211	Jan	Tue	11 AM	66.6687...	0.52092...	21.2559...	2.47216...	37.73	17.9	46	6.1	752.011	1024	0.0
2018	cluster_1	212	Jan	Tue	11 AM	65.2979...	0.58618...	19.5474...	3.23547...	29.33	18.7	45	6.5	751.937	1032	0.0

Detalle de los valores de las dimensiones, valores de clasificación por temporalidad (mes, día, hora).

En la figura 49, el porcentaje contenido del “cluster_2” seleccionado corresponde a un 12,89% de los datos totales esto agrupa a 4.414 registros. Se aprecia la distribución de las líneas a través de las diferentes dimensiones que contiene este clúster, se puede observar el vínculo de los contaminantes y las variables meteorológicas.

Figura 49

Selección de clúster de agrupamiento (clúster 2)

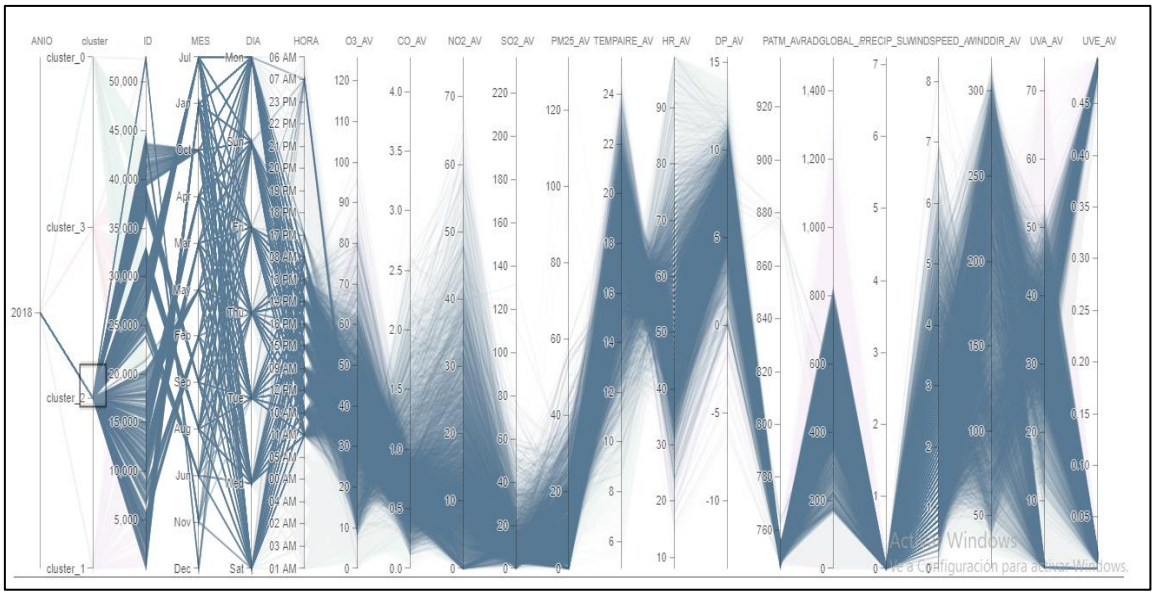
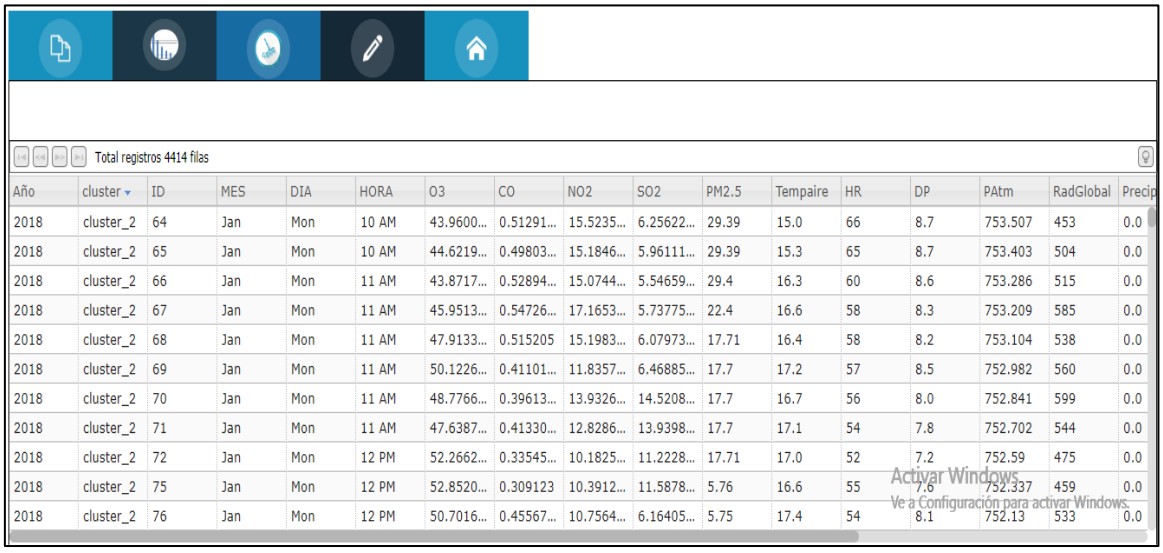


Figura 50

Listado en relación con la selección de agrupamiento (clúster 2)



Total registros 4414 filas

Año	cluster	ID	MES	DIA	HORA	O3	CO	NO2	SO2	PM2.5	Tempaire	HR	DP	PAtm	RadGlobal	Precip
2018	cluster_2	64	Jan	Mon	10 AM	43.9600...	0.51291...	15.5235...	6.25622...	29.39	15.0	66	8.7	753.507	453	0.0
2018	cluster_2	65	Jan	Mon	10 AM	44.6219...	0.49803...	15.1846...	5.96111...	29.39	15.3	65	8.7	753.403	504	0.0
2018	cluster_2	66	Jan	Mon	11 AM	43.8717...	0.52894...	15.0744...	5.54659...	29.4	16.3	60	8.6	753.286	515	0.0
2018	cluster_2	67	Jan	Mon	11 AM	45.9513...	0.54726...	17.1653...	5.73775...	22.4	16.6	58	8.3	753.209	585	0.0
2018	cluster_2	68	Jan	Mon	11 AM	47.9133...	0.515205	15.1983...	6.07973...	17.71	16.4	58	8.2	753.104	538	0.0
2018	cluster_2	69	Jan	Mon	11 AM	50.1226...	0.41101...	11.8357...	6.46885...	17.7	17.2	57	8.5	752.982	560	0.0
2018	cluster_2	70	Jan	Mon	11 AM	48.7766...	0.39613...	13.9326...	14.5208...	17.7	16.7	56	8.0	752.841	599	0.0
2018	cluster_2	71	Jan	Mon	11 AM	47.6387...	0.41330...	12.8286...	13.9398...	17.7	17.1	54	7.8	752.702	544	0.0
2018	cluster_2	72	Jan	Mon	12 PM	52.2662...	0.33545...	10.1825...	11.2228...	17.71	17.0	52	7.2	752.59	475	0.0
2018	cluster_2	75	Jan	Mon	12 PM	52.8520...	0.309123	10.3912...	11.5878...	5.76	16.6	55	7.6	752.337	459	0.0
2018	cluster_2	76	Jan	Mon	12 PM	50.7016...	0.45567...	10.7564...	6.16405...	5.75	17.4	54	8.1	752.13	533	0.0

Detalle de los valores de las dimensiones, valores de clasificación por temporalidad (mes, día, hora).

En la figura 51, el total de registros de la selección agrupa a 9.369 registros, siendo el porcentaje contenido del “cluster_3” el 27,36 % de los datos totales. Se observa como fluctúan a través de todo el gráfico los datos seleccionados contenidos en el “cluster_3” en dónde se consigue verificar el comportamiento y la tendencia del subconjunto de datos.

Figura 51

Selección de clúster de agrupamiento (clúster 3)

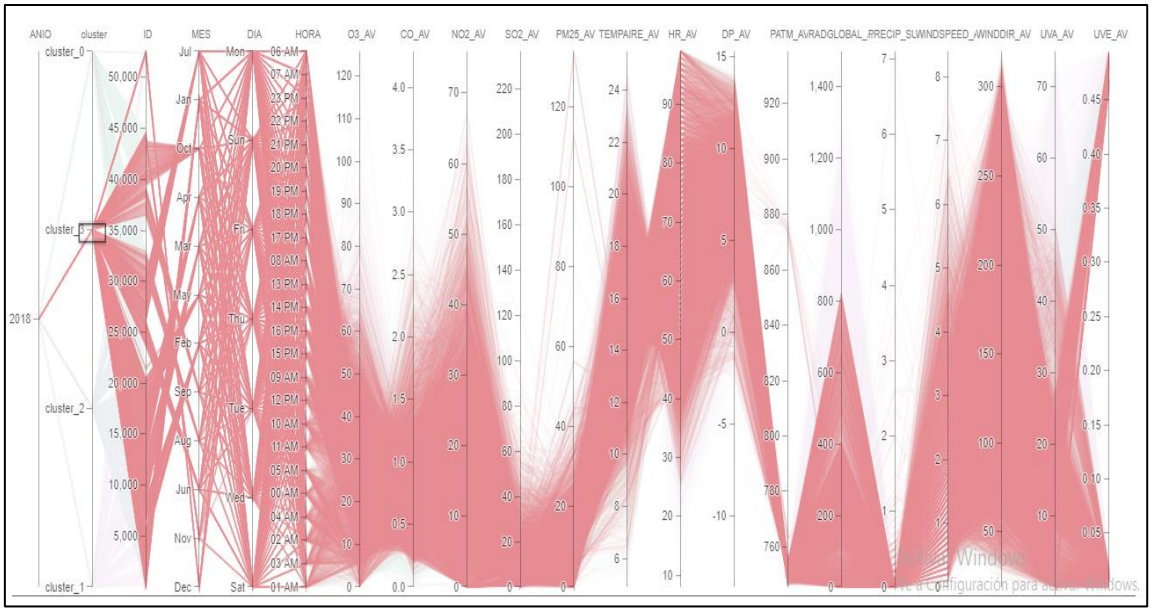
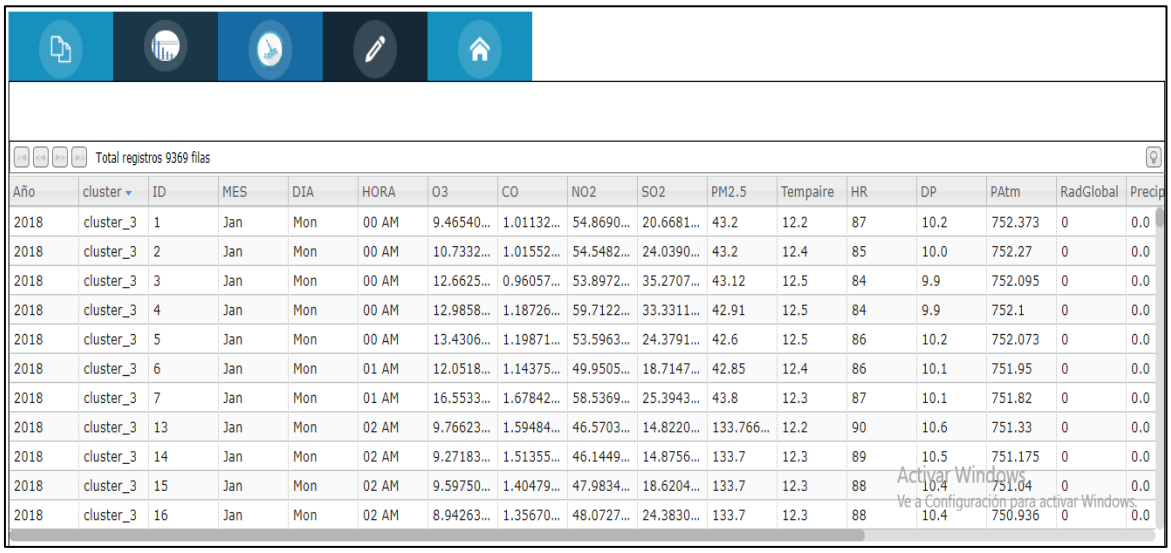


Figura 52

Listado en relación con la selección de agrupamiento (clúster 3)



Total registros 9369 filas

Año	cluster	ID	MES	DIA	HORA	O3	CO	NO2	SO2	PM2.5	Tempaire	HR	DP	PAtm	RadGlobal	Precip
2018	cluster_3	1	Jan	Mon	00 AM	9.46540...	1.01132...	54.8690...	20.6681...	43.2	12.2	87	10.2	752.373	0	0.0
2018	cluster_3	2	Jan	Mon	00 AM	10.7332...	1.01552...	54.5482...	24.0390...	43.2	12.4	85	10.0	752.27	0	0.0
2018	cluster_3	3	Jan	Mon	00 AM	12.6625...	0.96057...	53.8972...	35.2707...	43.12	12.5	84	9.9	752.095	0	0.0
2018	cluster_3	4	Jan	Mon	00 AM	12.9858...	1.18726...	59.7122...	33.3311...	42.91	12.5	84	9.9	752.1	0	0.0
2018	cluster_3	5	Jan	Mon	00 AM	13.4306...	1.19871...	53.5963...	24.3791...	42.6	12.5	86	10.2	752.073	0	0.0
2018	cluster_3	6	Jan	Mon	01 AM	12.0518...	1.14375...	49.9505...	18.7147...	42.85	12.4	86	10.1	751.95	0	0.0
2018	cluster_3	7	Jan	Mon	01 AM	16.5533...	1.67842...	58.5369...	25.3943...	43.8	12.3	87	10.1	751.82	0	0.0
2018	cluster_3	13	Jan	Mon	02 AM	9.76623...	1.59484...	46.5703...	14.8220...	133.766...	12.2	90	10.6	751.33	0	0.0
2018	cluster_3	14	Jan	Mon	02 AM	9.27183...	1.51355...	46.1449...	14.8756...	133.7	12.3	89	10.5	751.175	0	0.0
2018	cluster_3	15	Jan	Mon	02 AM	9.59750...	1.40479...	47.9834...	18.6204...	133.7	12.3	88	10.4	751.04	0	0.0
2018	cluster_3	16	Jan	Mon	02 AM	8.94263...	1.35670...	48.0727...	24.3830...	133.7	12.3	88	10.4	750.936	0	0.0

Detalle de los valores de las dimensiones, valores de clasificación por temporalidad (mes, día, hora).

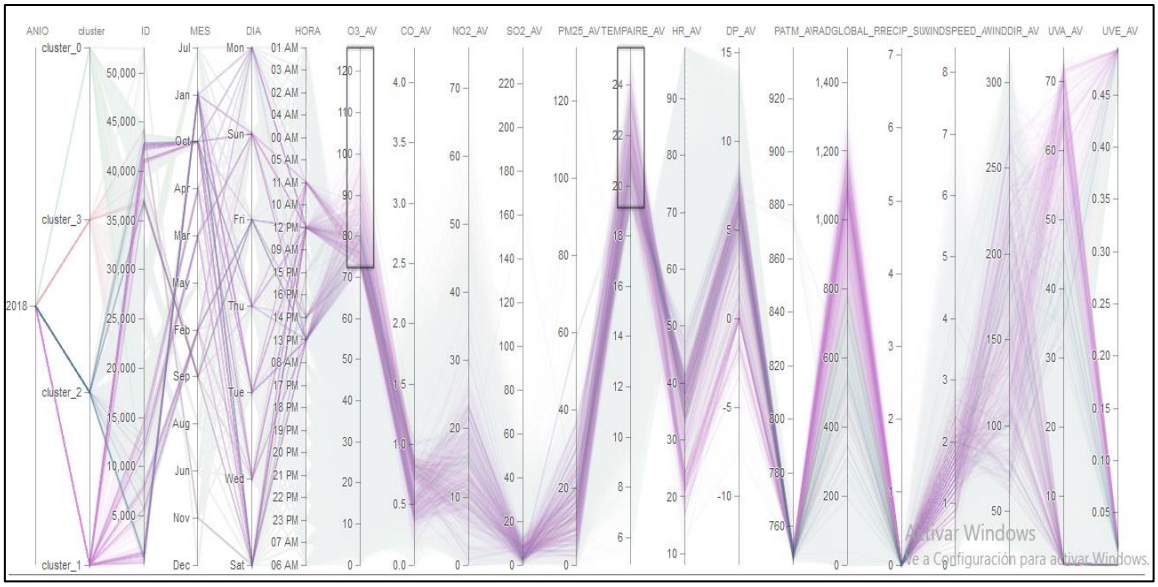
4.4 Relación entre variables

En el caso del ozono (O3) y la temperatura del aire (TEMPAIRE) presentaron una correlación muy alta entre sus valores, el comportamiento del ozono tiene una tendencia directamente proporcional a la temperatura, corroborando los resultados planteados en la fase anterior del estudio realizado, según Andrade Durazno, (2018).

En la figura 53, se observó el patrón de comportamiento y fluctuación de las variables mencionadas, además tomando en consideración otras variables meteorológicas que directamente se encontraron involucradas de manera proporcional con esta analogía, las variables meteorológicas: Radiación Global (RADGLOBAL), Radiación UVA, Radiación UVE, resultaron ser directamente proporcional al ozono (O3) y a la temperatura (TEMPAIRE).

Figura 53

Relación Ozono(O3) y Temperatura (Tempaire)



Otro importante resultado que se encontró entre el ozono (O3) y la velocidad del viento (WINDSPEED), donde existe una correlación proporcional. Andrade (2018), reveló que existe una correlación negativa entre las horas de 10:00 am y 12:00 pm, como se puede ver en la figura 54, posterior a estas horas, el comportamiento entre el ozono (O3) y la velocidad del viento (WINDSPEED) tiene una correlación positiva directamente proporcional como se observa en la figura 55.

Figura 54

Ozono(O3) y Velocidad del Viento (WINDSPEED) - 10:00 a 12:00

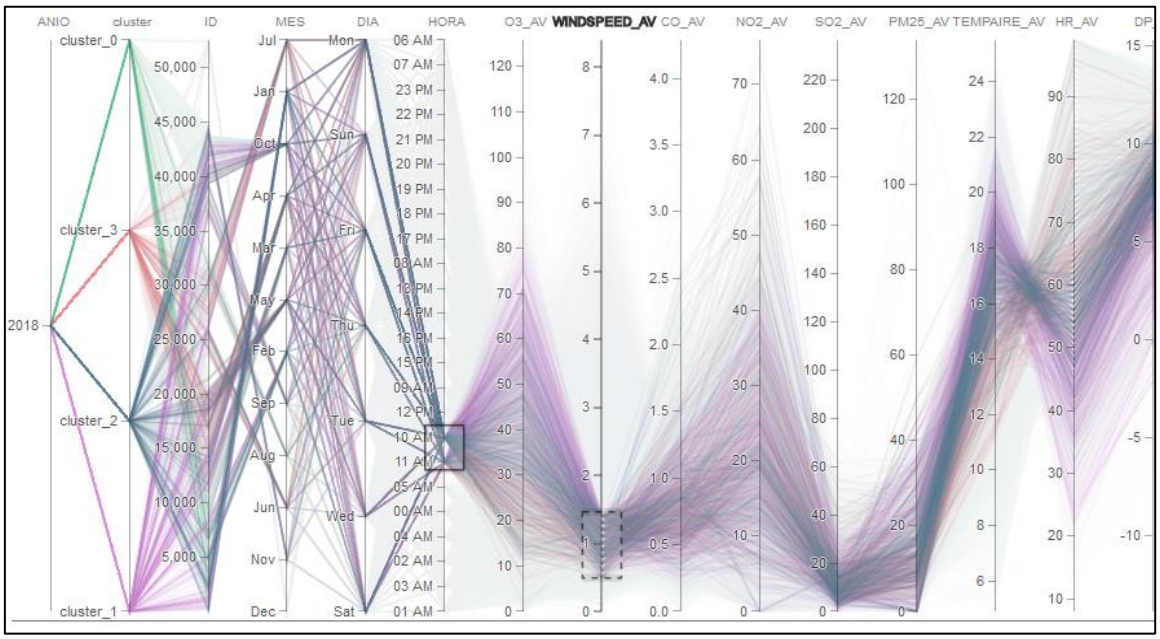
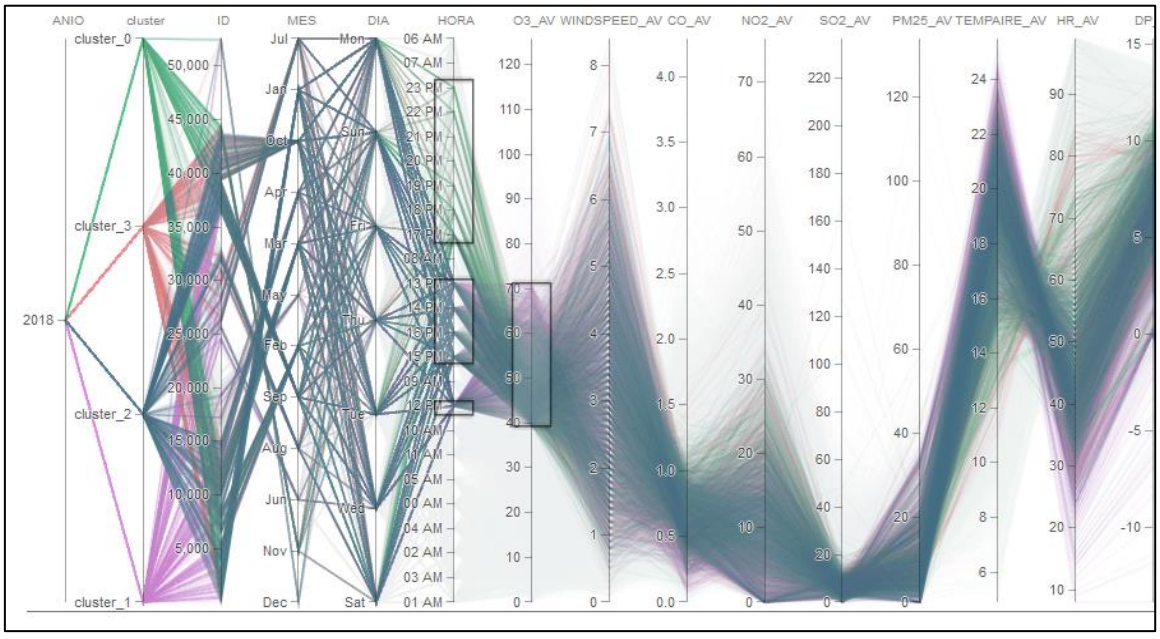


Figura 55

Relación Ozono(O3) y Velocidad del Viento (WINDSPEED)



Además, entre el ozono (O3) y la presión atmosférica (PATM) se determinó que existe una correlación inversamente proporcional, como se aprecia en la figura 56. Andrade (2018). Existe también una correlación negativa como se observa en la gráfica, sin embargo, a las 14:00 pm alcanza un valor pico positivo, y se verifica el resultado, también que existe fluctuación en las diferentes horas del día, como se pudo analizar en la figura 57.

Figura 56

Relación Ozono(O3) y Presión Atmosférica (PATM)

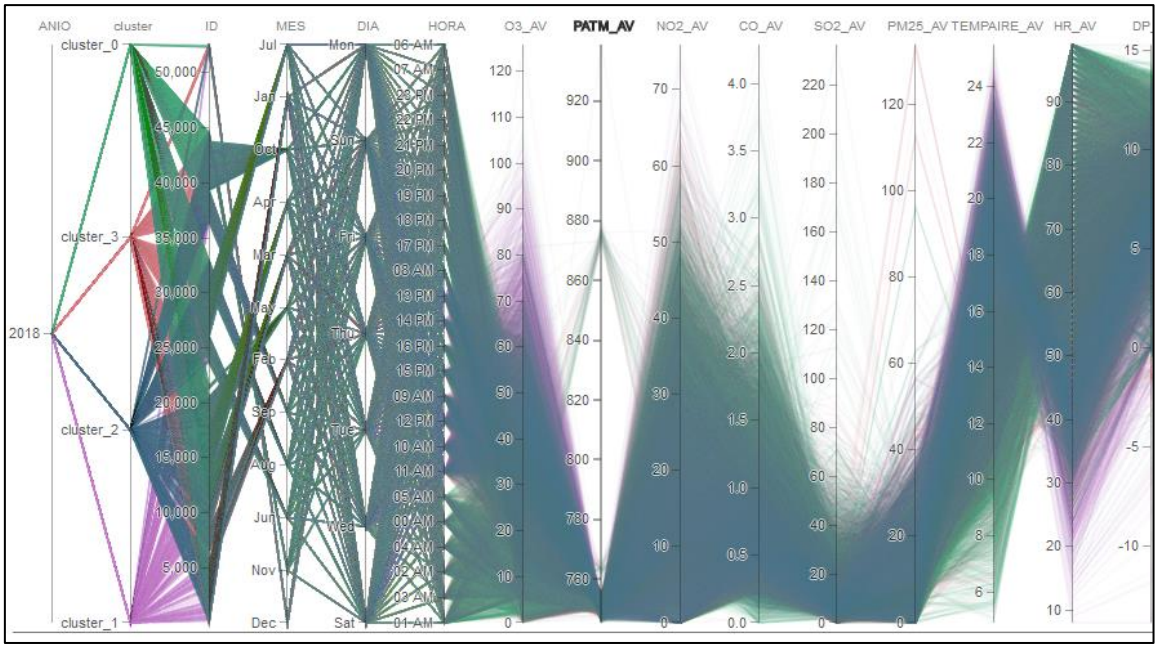
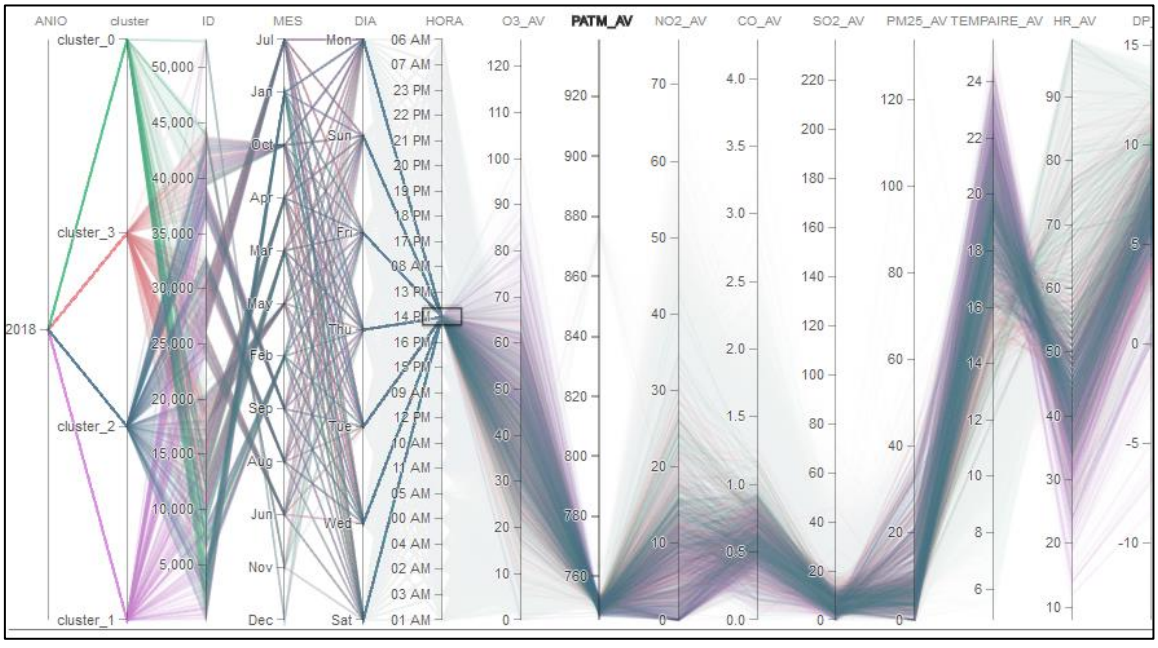


Figura 57

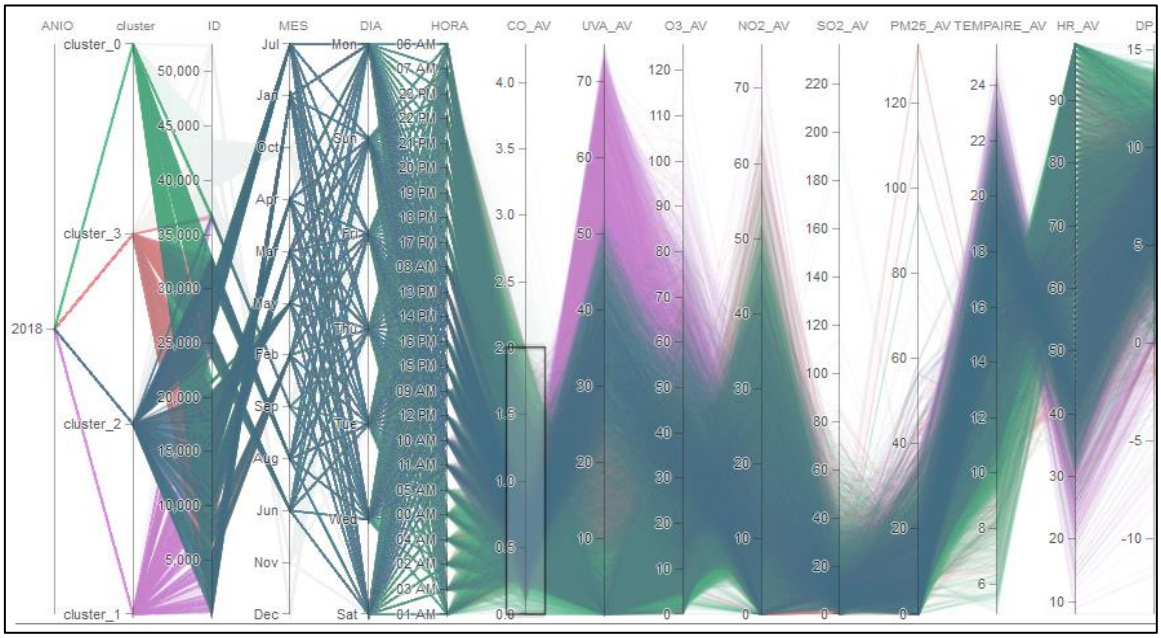
Relación Ozono(O3) y Presión Atmosférica (PATM) - 14:00



Por otro lado, entre el monóxido de carbono (CO) y la radiación ultravioleta (UVA) existe una correlación inversamente proporcional, como se observa en la figura 58. Entre menores sean los valores del CO, mayores serán los valores de la radiación UVA.

Figura 58

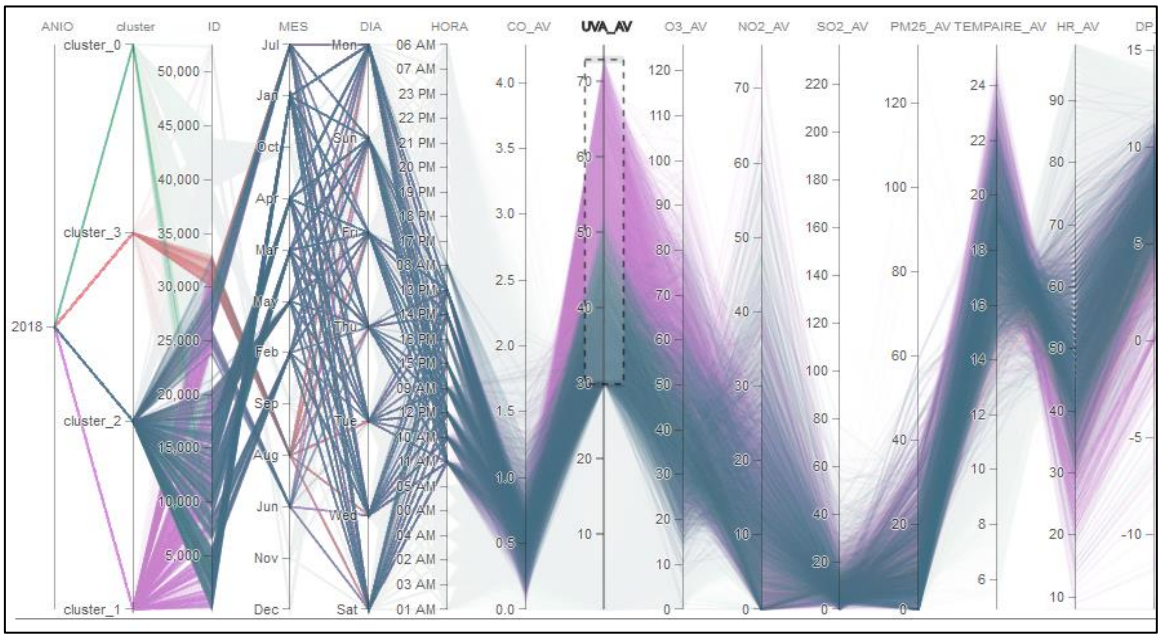
Relación Monóxido de Carbono (CO) y la Radiación Ultravioleta (UVA)



Se determinó que posterior a las horas entre 07:00 am y 09:00 am, existe un incremento inverso de manera proporcional de la variable meteorológica de radiación ultravioleta UVA, como se observa en la figura 59.

Figura 59

Monóxido de Carbono (CO) y la Radiación Ultravioleta (UVA) 30-75



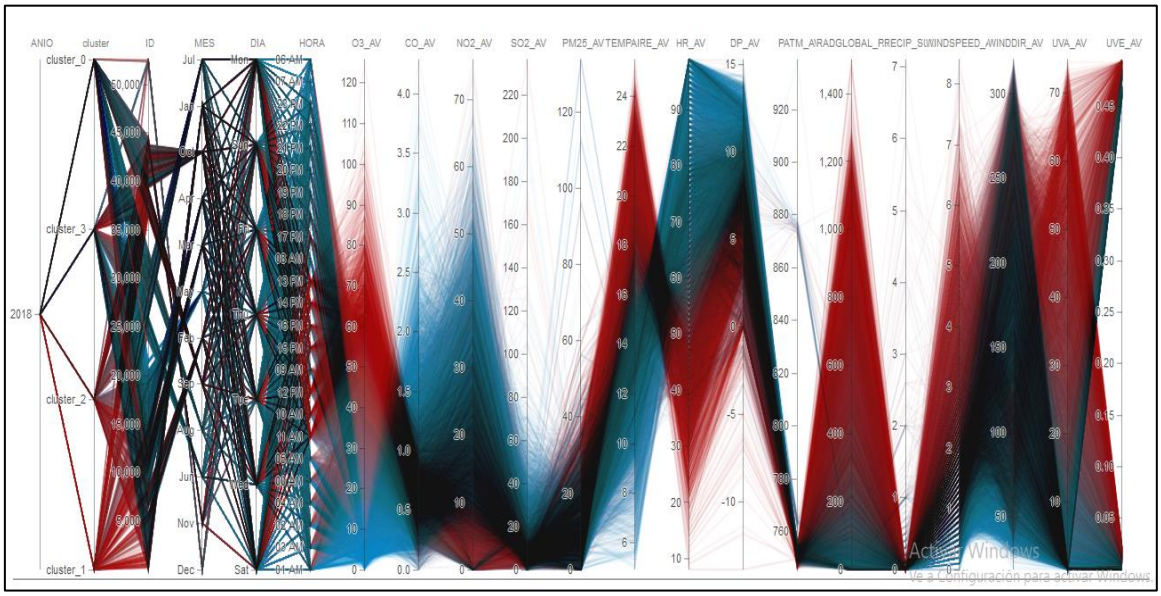
4.5 Colorimetría

Otra de las particularidades que se presentó en el gráfico resultante del comportamiento de las variables, fue el aspecto de colorimetría, es decir el color de las líneas del gráfico basado en la fluctuación de un determinado contaminante o variable meteorológica seleccionado por el usuario. Dentro del análisis que se realizó con el ozono (O3), se obtuvo un resultado relevante en donde el color azul determina bajas concentraciones del contaminante, mientras que el color rojo nos indica una alerta de altos índices de concentración del contaminante.

En la figura 60, se observa el patrón de comportamiento del ozono (O3) basado en colorimetría de advertencia y prevención. Las líneas de color rojo indican el patrón de fluctuación y relaciones con el resto de contaminantes atmosféricos y variables meteorológicas.

Figura 60

Determinación colorimetría basada en el Ozono(O3)



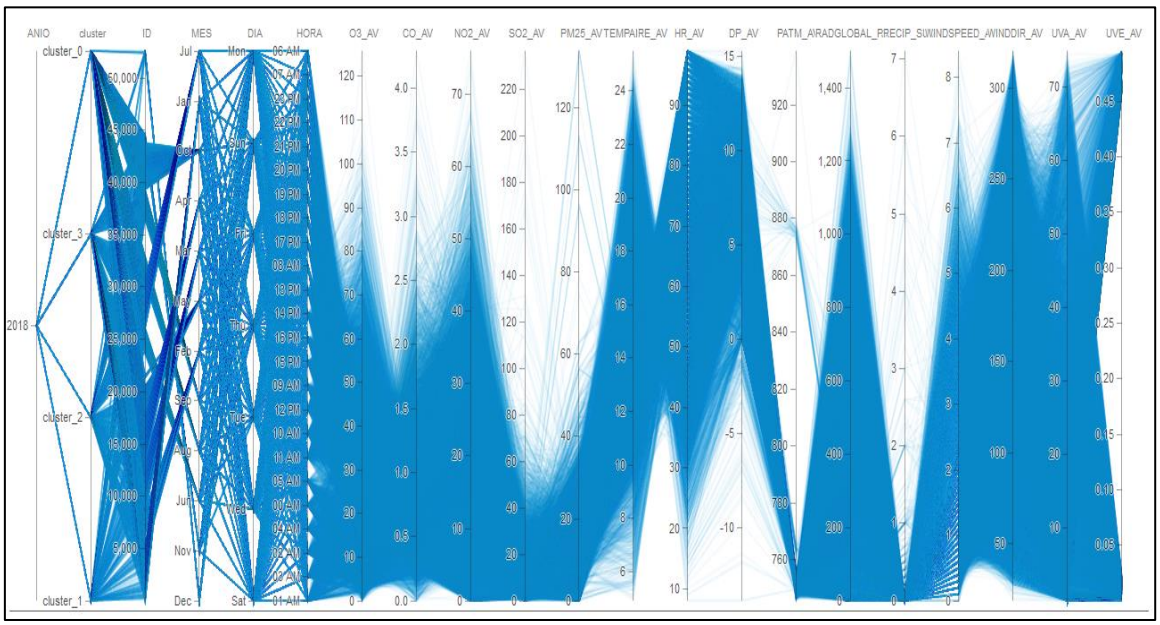
Se observó la correspondencia relacional que presentaron las variables dentro de un aspecto colorimétrico, se distinguió la relación directa y proporcional entre el ozono (O3), la temperatura (TEMPAIRE) y las variables meteorológicas (WINDSPEED, Radiación UVA, Radiación UVE). También se identificó que la variable humedad relativa (HR), disminuyó sus valores, corroborando la lógica de la relación y comportamiento en la zona de estudio (Cuenca), que, con mayor temperatura, disminuyó la humedad.

Se consideró también esta particularidad para el contaminante Monóxido de Carbono (CO), el aspecto colorimétrico en esta dimensión se puede obtener un comportamiento diferente.

En la figura 59 se observa el patrón de comportamiento del Monóxido de Carbono (CO) basado en colorimetría considerando que la mayor cantidad de datos comprende en un rango de 0 a 1.5 partes por billón (ppb).

Figura 61

Determinación colorimetría basada en el Monóxido de Carbono (CO)



Dentro del rango comprendido de 0 a 1.5 partes por billón (ppb), se encuentra un total de 32.639 registros lo que corresponde a un 95% del total de datos.

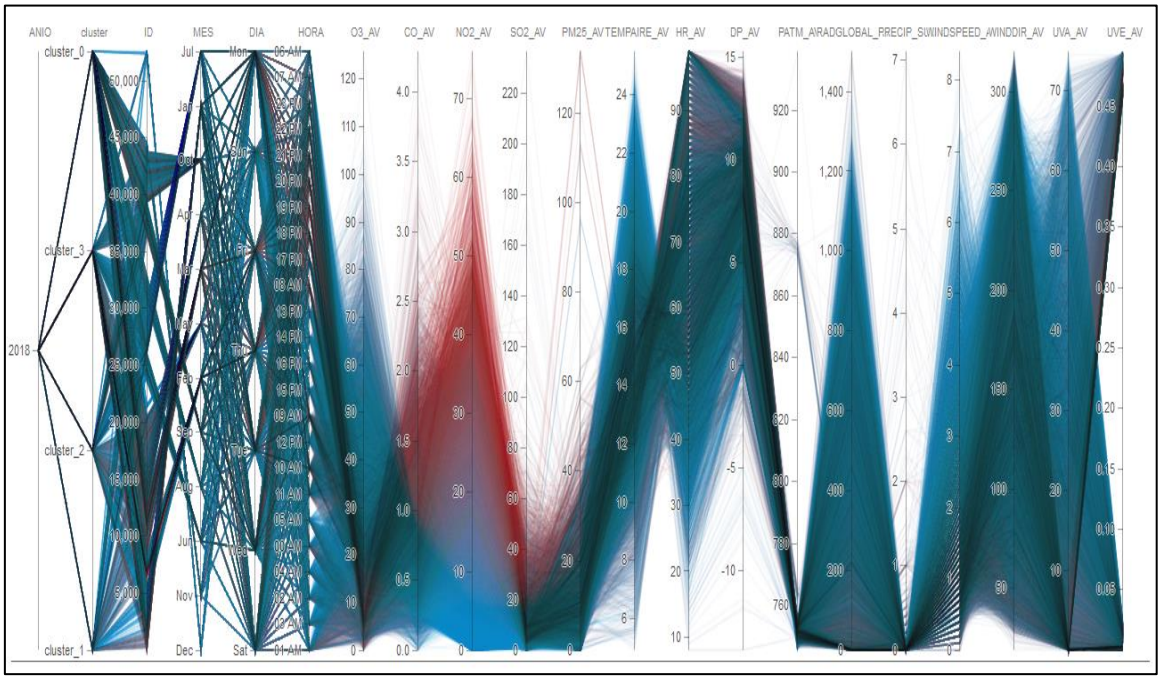
Por lo tanto, al existir una mayor concentración de datos o registros en un rango determinado y no contener un mayor despliegue dentro de la dimensión del contaminante que evalúa el aspecto colorimétrico, no existe una diferenciación visual notable. Esto sucede debido a que los datos que contienen los valores de altas concentraciones en esta dimensión, son muy pocos con respecto al total general. Se consideró, que pudieron tratarse de datos anómalos o datos que se han registrado de manera explícita por un comportamiento bastante inusual.

Algo diferente ocurre con la selección del contaminante Dióxido de Nitrógeno (NO2) para determinar el aspecto colorimétrico, en donde, se logró apreciar una variación visual en cuanto a los valores comprendidos en el conjunto de datos y su despliegue entre mínimos y máximos.

En la siguiente figura, se observa el patrón de comportamiento del Dióxido de Nitrógeno (NO2) y la relación con respecto a las diferentes dimensiones consideradas en el gráfico.

Figura 62

Determinación colorimetría basada en el Dióxido de Nitrógeno (NO2)



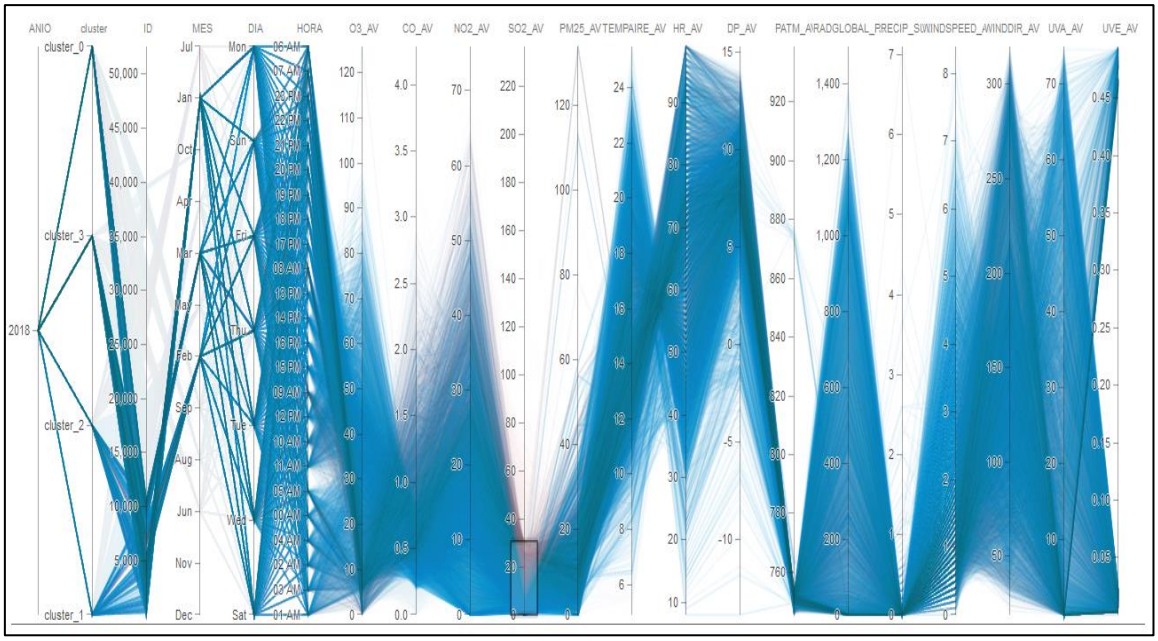
Se puede observar, una cierta relación directa con el Monóxido de Carbono (CO), el Dióxido de Azufre (SO2) y algunas de las variables meteorológicas como: humedad relativa (HR), punto de rocío (DP), y dirección del viento (WINDDIR).

Al considerar el contaminante Dióxido de Azufre (SO2) como otra variante de selección de colorimetría, se observó que el comportamiento es similar al Monóxido de Carbono (CO), en donde se concentran valores muy cercanos a la totalidad de los datos en un rango específico.

Dentro del rango comprendido de 0 a 30 partes por billón (ppb), se encuentra un total de 33.083 registros lo que corresponde a un 97% del total de datos.

Figura 63

Determinación colorimetría basada en el Dióxido de Azufre (SO2)



5. DISCUSIÓN

Se ha realizado el desarrollo del proceso de minería de datos y visualización de grandes volúmenes de información, con respecto a los contaminantes atmosféricos y variables meteorológicas recolectados por el IERSE, experimentando en datos correspondientes a un año de recolección (2018), de donde se pudo obtener varios aspectos de análisis basándose en la relación y asociación entre contaminantes y variables.

Este seguimiento demuestra una funcionalidad adecuada y acorde a los objetivos planteados desde el inicio del estudio y la correcta aplicación del ejemplo de verificación del desarrollo de la metodología CRISP-DM, así también, el seguimiento demuestra la adecuada selección y aplicación de la técnica de visualización, diseñada, desarrollada e implementada, a través de una aplicación web dinámica e interactiva que permite el análisis, evaluación y control de los patrones de comportamiento de los contaminantes atmosféricos y variables meteorológicas para la gestión y toma de decisiones por parte del gestor ambiental.

La visualización implementada ha demostrado apoyar a la percepción de visualización al igual que en el trabajo de (Liao et al., 2014), que expuso la vista de coordenadas paralelas, entre la hora y seis tipos de contaminantes para la evaluación multidimensional. Además, se obtuvo corroborar los análisis detallados por Thomas, Lokanathan & Jothi (2017), en su estudio que permitió clasificarlos por medio de colorimetría, y poder representar por filtrado de los datos.

Se desarrolló el proceso de minería de datos para 54.000 registros colectados, que se refiere al rango de meses entre enero a diciembre del año 2018 obteniendo satisfactoriamente como resultado 34.231 registros del tratamiento de minería, los mismos que se encuentran en el set de datos final que alimenta el contenido de la gráfica de la visualización por medio de la técnica de coordenadas paralelas.

Los resultados permiten en una primera instancia la evaluación general de 21 dimensiones de ejes verticales paralelos de los diversos materiales presentados en la atmósfera, donde se determinó una lógica de correspondencia y relación de las dimensiones, además se observó que los valores dimensionales referentes a los contaminantes atmosféricos y variables meteorológicas tienen integridad con respecto a los estudios previos, por ejemplo algunos de ellos realizados, Andrade (2018), Ortega (2018), Sellers (2013) entre otros.

Este resultado visual general permitió encontrar posibles errores, comúnmente que parten de los datos, o errores de proceso y desarrollo, hasta errores de interpretación; al realizar una observación inicial de amplio espectro se pudo verificar el comportamiento y las tendencias de los datos, las correlaciones, valores anómalos, etc.

En la observación general realizada se pudo identificar que la variable meteorológica de la radiación ultravioleta UVE tiene un comportamiento particular, agrupando sus valores en rangos de valores bajos y valores altos, generando una brecha amplia entre estos grupos, lo que permite suponer un probable error dentro del análisis.

Además, en comparación a la radiación ultravioleta UVA, tiene una variación en las unidades de la dimensión paralela, lo que permite plantear una hipótesis de error en la recolección o tratamiento previo, inclusive alguna inferencia al error humano dentro de algún proceso.

La temporalidad es otro de los aspectos que se pudo evaluar, se incluyó dimensiones de tiempo que permitieron plantear aspectos horarios fundamentales e importantes dentro de los resultados obtenidos, aquello sirvió de base para determinar grandes escenarios e hipótesis con base a estas agrupaciones temporales, por ejemplo: las horas, días, o meses de filtrado, la diferenciación climatológica de las estaciones de invierno, verano, otoño y primavera, entre otras, de tal forma que permite crear planes de acción que tengan provecho de interés general. Sin embargo, la interacción lograda como resultado permite realizar muchas combinaciones y muchas hipótesis que el gestor puede tomar en consideración.

Los resultados con base al porcentaje y número de registros contenidos por los clústers, permitió que se pueda validar el proceso, de los diferentes escenarios que se plantearon, siendo de gran provecho la selección del algoritmo de evaluación y la distribución de los datos por los diferentes clústers, no obstante, la naturalidad de los datos, las unidades y el escalamiento multidimensional reveló que la distribución fue bastante general, pero permite la interpretación.

Dentro de las relaciones entre variables se comparó los resultados expuestos con algunas investigaciones recientes, siendo el contaminante ozono (O3) el más influyente en varios estudios (Ortega, 2018). El ozono (O3) y la temperatura del aire (TEMPAIRE) presentaron una correlación alta, una tendencia directamente proporcional a la temperatura corroborando los resultados planteados en la fase anterior del estudio realizado, según Andrade (2018), adicional se encontró relación directa también con la Radiación Global (RADGLOBAL), la Radiación UVA, y la Radiación UVE.

Se obtuvo grandes y diferentes ventajas de la aplicación del presente trabajo de titulación, la técnica visual coordenadas paralelas, permitió representar grandes conjuntos de datos, representar valores superiores a tres dimensiones en un plano 2D, efectividad para el reconocimiento de patrones por el gestor.

Además, es altamente flexible debido a su escalabilidad individual para cada eje paralelo permitiendo combinar varias unidades de medias; y también permite la integración de herramientas efectivas (zoom y brush), para una exploración efectiva del conjunto de datos.

La principal desventaja es la conglomeración visual que se puede generar, esto debido al exceso de ejes que a mayor cantidad se acercan entre sí, también debido a la gran cantidad de datos que se visualizan. Esto se solucionó a través del reordenamiento y eliminación de ejes que puede ser posible dentro de la gráfica generada; sin embargo, puede existir dificultades y malas interpretaciones de los resultados debido a la cercanía que los ejes paralelos puedan tener si el número de variables de visualización es considerable. Por otro lado, al existir grandes volúmenes de información o registros, la gestión de la aplicación web se ralentiza, ya que se realiza el proceso de renderización de todos los registros, lo que provoca que se consuma altos recursos de procesamiento.

Un aspecto dentro de las limitaciones del trabajo, fue la implementación de una técnica visual combinatoria, es decir, no se realizó la implementación de una técnica híbrida que combine técnicas de visualización dentro del mismo plano de las coordenadas paralelas, una sección adicional refleja las barras laterales lo que mejora y facilita la interpretación.

Se podría aplicar en el futuro algunas técnicas híbridas de visualización para grandes volúmenes de datos; también, se podría aplicar a futuro técnicas matemáticas y estadísticas que no permitan depreciar cantidades considerables de datos para el proceso de minería, debido a que las herramientas de minería al tener valores nulos en las variables de análisis, los descartan de la estadística, estas técnicas podrían estar basadas en completar la información necesaria a través predicciones estadísticas.

En contexto, es importante recalcar que se cumplió con los objetivos propuestos: se identificaron todas las variables de estudio, se realizó un marco teórico de la técnica de minería de datos propuesta y de las técnicas visuales interactivas, y se desarrolló una aplicación que permite visualizar los grandes volúmenes de datos a través de una técnica de visualización multidimensional dinámica de los patrones de comportamiento de las variables de contaminación atmosférica y variables meteorológicas.

Se generó conocimiento que permite la toma de decisiones al gestor ambiental, lo que podría motivar la creación de planes de acción por parte de las entidades interesadas, además, este trabajo de graduación permite ser una base literaria para investigaciones de análisis y gestión de relaciones específicas, reduciendo el espectro del análisis enfocando con más detalle a grupos pequeños de contaminantes atmosféricos y variables meteorológicas.

6. CONCLUSIONES

En este trabajo de graduación, se cumplió satisfactoriamente con el objetivo general y los objetivos específicos propuestos: se identificaron las variables de estudio, se realizó un marco teórico de la técnica de minería de datos propuesta y de las técnicas visuales interactivas, se aplicó el ejemplo de verificación basado en la metodología CRISP-DM sobre las variables contenidas en la atmósfera, y se desarrolló una aplicación web que permite visualizar los grandes volúmenes de datos resultantes a través de una técnica de visualización multidimensional de coordenadas paralelas, determinando los patrones de comportamiento de las variables de contaminación atmosférica y variables meteorológicas que permitan la toma de decisiones de los gestores ambientales.

Se generó conocimiento, que permite crear planes de acción por parte de los gestores ambientales. Además, es importante mencionar que previamente, los gestores ambientales lograron examinar y validar los resultados establecidos a través de ciertas hipótesis y filtrados constatando la integridad en los datos y su correcta distribución.

Este método de desarrollo, a diferencia de las propuestas mencionadas en trabajos relacionados anteriores, se enfoca a la visualización, interpretación y minado de datos por medio de filtros y barridos de subconjuntos que permite la variación gráfica en tiempo real, detectando patrones de condiciones detalladas y se adapta a la alimentación de varios archivos de texto (CSV) como disponga el gestor ambiental.

Este trabajo de graduación, se concluye generando un aporte significativo literario, que sea útil para implementar técnicas híbridas de visualización. También puede servir de base para investigaciones de análisis y gestión de relaciones específicas, enfocando con más detalle a grupos pequeños, dimensiones específicas o subgrupos de contaminantes atmosféricos y variables meteorológicas.

El contaminante ozono (O3) fue el más influyente en varios estudios generando importancia para el control y gestión del mismo (Ortega, 2018), coincidiendo con los hallazgos resultantes encontrados en el presente trabajo. El ozono (O3), la temperatura del aire (TEMPAIRE), la Radiación Global (RADGLOBAL), la Radiación UVA, y la Radiación UVE, presentaron una correlación alta, en consecuencia se concluyó que existe una tendencia directamente proporcional entre estas dimensiones, corroborando los resultados planteados según Andrade (2018) y se pudo verificar la capacidad de análisis de varias dimensiones en una sola consulta.

El aspecto colorimétrico permitió determinar la variabilidad de los valores por medio de la interpolación entre rangos de colores, es decir, según la distribución los datos de cada una de las dimensiones, se pudo asignar un color específico. Esto facilitó los análisis de evaluación para los escenarios e hipótesis, sin embargo, se pudo adaptar algunas funcionalidades extras de la técnica visual para su correcta y total interpretación.

Las herramientas funcionales disponibles dentro de la aplicación web de resultados, permitieron un manejo intuitivo del usuario, esto debido a las técnicas vanguardistas vigentes de visualización con gran despliegue y escalabilidad. En la actualidad se dispone de varias herramientas informáticas que permiten el desarrollo de las técnicas visuales, y son de fácil acceso para formular conocimiento útil y de gran provecho.

Adicional, se procesó la minería de los datos, a través de los algoritmos de agrupación (X-Means, K-Means y DBSCAN) siendo el algoritmo K-Means el mejor algoritmo de procesamiento, por su gran rendimiento y adaptación al conjunto de datos. Se usó la herramienta Rapidminer, que resultó ser la herramienta óptima para el desarrollo de la metodología, procesamiento y gestión (Matute, 2018).

Se disminuyó en gran medida, las posibles limitaciones y desventajas que pudo presentar la gráfica de resultados, sin embargo, existen algunas adaptaciones que se pueden considerar para la evolución de la presente investigación.

En un trabajo futuro, se podría proponer una técnica de visualización híbrida, que combine dos o más técnicas de visualización para una mejor estimación de resultados, pudiendo ser propuesta la técnica Scattering Points en Coordenadas Paralelas (SPPC), con gran adaptación para el área de contaminación del aire. Además, se podría considerar un set de datos más grande, debido a que la mayor cantidad de datos permitiría revelar comportamientos de una forma más exacta e incluiría temporalidades referentes a comportamientos anuales.

Otra propuesta que podría estimarse, es la generación de varias pantallas de visualización en la presentación de resultados, que permita incluir varias técnicas de visualización avanzada y geovisualización, e incluso alimentar los datos de la aplicación en tiempo real, a través de la gestión de un servidor asignado según corresponda la viabilidad y disponibilidad recursos de software y hardware.

BIBLIOGRAFÍA

Abhilash, M. S. K., Thakur, A., Gupta, D., & Sreevidya, B. (2018). Time series analysis of air pollution in bengaluru using ARIMA model. *Advances in Intelligent Systems and Computing*, *696*, 413–426. https://doi.org/10.1007/978-981-10-7386-1_36

Agila-Palacios, M. V., Ramirez-Montoya, M. S., García-Valcárcel, A., & Samaniego-Franco, J. (2017). Uso de la tableta digital en entornos universitarios de aprendizaje a distancia. *RIED. Revista Iberoamericana de Educación a Distancia*, *20*(2), 255. https://doi.org/10.5944/ried.20.2.17712

Aigner, W., Miksch, S., Schumann, H., & Tominski, C. (2011). *Visualization of Time-Oriented Data*. https://doi.org/10.1007/978-0-85729-079-3

Andrade Durazno, S. P. (2018). *APLICACIÓN DE MINERIA DE DATOS EN EL ANÁLISIS DE CONTAMINANTES ATMOSFÉRICOS Y VARIABLES METEOROLÓGICAS* (Universidad del Azuay). Recuperado de: http://dspace.uazuay.edu.ec/handle/datos/8466

Andrienko, G., Andrienko, N., Burch, M., & Weiskopf, D. (2012). Visual analytics methodology for eye movement studies. *IEEE Transactions on Visualization and Computer Graphics*, *18*(12), 2889–2898. https://doi.org/10.1109/TVCG.2012.276

Baloian, N. (2016). Tutorial : Uso básico de RapidMiner. *Universidad De Chile*.

Basantes, A. C. N., & García, E. H. (2018). Altitud, variables climáticas y tiempo de permanencia de las personas en plazas de Ecuador. *Urbe. Revista Brasileira de Gestão Urbana*, *10*(2), 414–425. https://doi.org/10.1590/2175-3369.010.002.ao11

Becker, R. A., & Cleveland, W. S. (1987). Brushing Scatterplots. *Technometrics*, *29*(2), 127–142. https://doi.org/10.1080/00401706.1987.10488204

Beltran, D., Poveda, D., Bolívar, A., Corrales, S. Y., & Sarmiento, H. M. (2010). *David beltran diego poveda*. 1–47.

Bostock, M., Ogievetsky, V., & Heer, J. (2011). *D 3 : Data-Driven Documents*. *17*(12), 2301–2309.

Brunekreef, B., & Holgate, S. T. (2002). Air pollution and health. *Lancet*, *360*(9341), 1233–1242. https://doi.org/10.1016/S0140-6736(02)11274-8

Burget, R., Karásek, J., Smékal, Z., Uher, V., & Dostál, O. (2014). Rapidminer image processing extension: A platform for collaborative research. *The 33rd International Conference on Telecommunication and Signal Processing, TSP*, (September), 114–118. Recuperado de: https://www.researchgate.net/publication/260727350_RapidMiner_image_processing_extension_A_platform_for_collaborative_research?enrichId=rgreq-08d6eebe-376f-416d-b9b7-63b6ed240dae&enrichSource=Y292ZXJQYWdlOzI2MDcyNzM1MDtBUzoxMDQ2NTUyMDM4NjQ1ODVAMTQwMTk2MzE

Cabrera Lituma, M. A. (2017). *Plataforma de visualización estadística de variables atmosféricas múltiples* (UNIVERSIDAD DEL AZUAY). Recuperado de: http://dspace.uazuay.edu.ec/handle/datos/6554%0A

Cañarte, K. (2007). Radiación ultravioleta. *Ciencia & Tecnología Para La Salud Visual y Ocular*, (9), 97. https://doi.org/10.19052/sv.1520

Castro, S., & Luján-Ganuza, M. (2017a). Datos Espaciales. *Laboratorio de Visualización y Computación Gráfica Dpto. de Cs. E Ing. de La Computación - UNS*, 1–2.

Castro, S., & Luján-Ganuza, M. (2017b). Datos N Dimensionales. *Laboratorio de Visualización y Computación Gráfica Dpto. de Cs. E Ing. de La Computación - UNS*.

Castro, S., & Luján-Ganuza, M. (2017c). El Diseño de la Visualización. *Laboratorio de Visualización y Computación Gráfica Dpto. de Cs. E Ing. de La Computación - UNS*.

Castro, S., & Luján-Ganuza, M. (2017d). El Proceso de Visualización. *Laboratorio de Visualización y Computación Gráfica Dpto. de Cs. E Ing. de La Computación - UNS*.

Castro, S., & Luján-Ganuza, M. (2017e). Espacio-Temporales. *Laboratorio de Visualización y Computación Gráfica Dpto. de Cs. E Ing. de La Computación - UNS*, 1–2.

Castro, S., & Luján-Ganuza, M. (2017f). Percepción en Visualización. *Laboratorio de Visualización y Computación Gráfica Dpto. de Cs. E Ing. de La Computación - UNS*, 1–43.

Castro, S., & Luján-Ganuza, M. (2017g). Visualización de Árboles y Grafos. *Laboratorio de Visualización y Computación Gráfica Dpto. de Cs. E Ing. de La Computación - UNS*.

Castro, S., & Luján, M. (2017). Técnicas de Visualización. *Laboratorio de Visualización y Computación Gráfica Dpto. de Cs. E Ing. de La Computación - UNS*, 22. https://doi.org/10.5433/1679-0359.2012v33Supl1p3033

Castro, S., Luján, M., & Martig, S. (2006). *Aplicación de Visualización de Grafos utilizando Servicios Web*. (Cic), 4.

Chen, P. (2019). Visualization of real-time monitoring datagraphic of urban environmental quality. *Eurasip Journal on Image and Video Processing*, *2019*(1). https://doi.org/10.1186/s13640-019-0443-6

Christina, J., & Komathy, K. (2013). Analysis of hard clustering algorithms applicable to regionalization. *2013 IEEE Conference on Information and Communication Technologies, ICT 2013*, (Ict), 606–610. https://doi.org/10.1109/CICT.2013.6558166

Claessen, J. H. T., & Van Wijk, J. J. (2011). Flexible linked axes for multivariate data visualization. *IEEE Transactions on Visualization and Computer Graphics*, *17*(12), 2310–2316. https://doi.org/10.1109/TVCG.2011.201

d3js.org. (2019). Recuperado de: https://d3js.org/

de Oliveira, R. H., Carneiro, C. de C., de Almeida, F. G. V., de Oliveira, B. M., Nunes, E. H. M., & dos Santos, A. S. (2018). Multivariate air pollution classification in urban areas using mobile sensors and self-organizing maps. *International Journal of Environmental Science and Technology*, (0123456789). https://doi.org/10.1007/s13762-018-2060-9

DefinicionABC. (2019). Píxel. Recuperado de: https://www.definicionabc.com/tecnologia/pixel.php

Demchenko, Y., De Laat, C., & Membrey, P. (2014). Defining architecture components of the Big Data Ecosystem. *2014 International Conference on Collaboration Technologies and Systems, CTS 2014*, 104–112. https://doi.org/10.1109/CTS.2014.6867550

Diez, T., Dominguez, M., Martinez, J., & Sáenz, J. (2012). Creación de páginas web accesibles con HTML5. *Consultado El*, 119–128. Recuperado de: http://www.esvial.org/wp-content/files/Atica2012_pp120-129.pdf

EL OZONO. (2012). Acta Bioquímica Clínica Latinoamericana. *Federación Bioquímica de La Provincia de Buenos Aires Argentina*, 41–49. Recuperado de: http://www.redalyc.org/articulo.oa?id=53559383005

EMOV EP. (2017). *Informe_Calidad_Aire_Cuenca_2017*. 123. Recuperado de: http://gis.uazuay.edu.ec/ierse/links_doc_contaminantes/Informes Claudia Calidad del Aire/Informe_Calidad_Aire_Cuenca_2017.pdf

Engel, D., Greff, K., Garth, C., Bein, K., Wexler, A., Hamann, B., & Hagen, H. (2012). Visual steering and verification of mass spectrometry data factorization in air quality research. *IEEE Transactions on Visualization and Computer Graphics*, *18*(12), 2275–2284. https://doi.org/10.1109/TVCG.2012.280

FECYT. (2004). *Meteorología y climatología: Ciencia y la Tecnología 2004.*

Ferraiolo, J., Jun, F., & Jackson, D. (2003). Scalable vector graphics (SVG) 1.1 specification. *World Wide Web Consortium (W3C). URL …*, (September). Recuperado de: http://www.tka4.org/materials/lib/Code/FileFormats/Vector2D-SVG/REC-SVG-20010904.pdf

Fundación para la Salud Geoambiental. (2013a). Dióxido de Azufre. Recuperado de: https://www.saludgeoambiental.org/dioxido-azufre-so2

Fundación para la Salud Geoambiental. (2013b). Material Particulado. Recuperado de: https://www.saludgeoambiental.org/material-particulado

Furht, B., & Villanustre, F. (2016). Big data technologies and applications. In *Big Data Technologies and Applications*. https://doi.org/10.1007/978-3-319-44550-2

Gallardo Arancibia, J. A. (2013). *Metodología para el Desarrollo de Proyectos en Minería de Datos CRISP-DM*. *84*, 487–492. Recuperado de: http://ir.obihiro.ac.jp/dspace/handle/10322/3933

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

González, M., Tamayo, M., & Sánchez, Á. (2009). LA RADIACIÓN ULTRAVIOLETA. SU EFECTO DAÑINO Y CONSECUENCIAS PARA LA SALUD HUMANA. *Theoria: Ciencia, Arte y Humanidades*, *18*, 69–80. Recuperado de: http://www.redalyc.org/articulo.oa?id=29917006006

Google Maps. (2019). Recuperado Marzo 3, 2019, de: https://www.google.com/maps/place/Cuenca/@-2.8922671,-79.0594206,12z/data=!3m1!4b1!4m5!3m4!1s0x91cd18095fc7e881:0xafd08fd090de6ff7!8m2!3d-2.9001285!4d-79.0058965

Gore, R. W., & Deshpande, D. S. (2017). An approach for classification of health risks based on air quality levels. *Proceedings - 1st International Conference on Intelligent Systems and Information Management, ICISIM 2017*, *2017-Janua*, 58–61. https://doi.org/10.1109/ICISIM.2017.8122148

Gui-Fen, C., Guo-Wei, W., Li, M., Li, Y., Jian, J., & Hang, C. (2009). Research on spatially weighted fuzzy dynamic clustering algorithm and spatial data mining visualization. *2009 WRI World Congress on Software Engineering, WCSE 2009*, *3*, 60–66. https://doi.org/10.1109/WCSE.2009.87

Gupta, R., Singh, N., & Singh, B. (2017). Neural Network Based GIS Application for Air Quality Visualization. *Proceedings on 2016 2nd International Conference on Next Generation Computing Technologies, NGCT 2016*, (October), 169–172. https://doi.org/10.1109/NGCT.2016.7877409

Hassan-Montero, Y., & Herrero-Solana, V. (2006). Improving Tag-Clouds as Visual Information Retrieval Interfaces. *Information Sciences*, *I*(2), 25–28. https://doi.org/10.1088/0268-1242/25/2/024015

Heinrich, J., & Weiskopf, D. (2013). State of the Art of Parallel Coordinates. *Eurographics Conference on Visualization (EuroVis)*, 95–116. https://doi.org/10.2312/conf/EG2013/stars/095-116

IBM, I. B. M. (2012). Manual CRISP-DM de IBM SPSS Modeler. *IBM Corporation*, 56. Recuperado de: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/15.0/es/CRISP-DM.pdf

IERSE. (2017). Recuperado Marzo 5, 2019, de: http://gis.uazuay.edu.ec/ierse/

Inselberg, A. (1985). The plane with parallel coordinates. *The Visual Computer*, *1*(4), 69–91. https://doi.org/10.1007/BF01898350

Jiawei, H., Micheline, K., & Jian, P. (2012). Data Mining: Concepts and Techniques, Third Edition. In *Morgan Kaufmann Publishers*. Recuperado de: http://library.books24x7.com/toc.aspx?bkid=44712

JS. (2019). JavaScript. Recuperado de: https://www.javascript.com/

Kandogan, E. (2001). Visualizing multi-dimensional clusters, trends, and outliers using star coordinates. *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '01*, 107–116. https://doi.org/10.1145/502512.502530

Kandogan, E., Road, H., & Jose, S. (2018). Roundabouts and Mini Roundabouts. *U.S. Department of Transportation Federal Highway Administration*. https://doi.org/10.1.1.4.8909

Karimi, H., Soffianian, A., Mirghaffari, N., & Soltani, S. (2016). Determining Air Pollution Potential Using Geographic Information Systems and Multi-criteria Evaluation: A Case Study in Isfahan Province in Iran. *Environmental Processes*, *3*(1), 229–246. https://doi.org/10.1007/s40710-016-0136-4

Kingsy, G., Manimegalai, R., Geetha, D., Rajathi, S., Usha, K., & Raabiathul, B. (2016). Air pollution analysis using enhanced K-Means clustering algorithm for real time sensor data. *IEEE Region 10 Annual International Conference, Proceedings/TENCON*, (August 2006), 1945–1949. https://doi.org/10.1109/TENCON.2016.7848362

KNIME. (2019). Plataforma de análisis KNIME. Recuperado de: https://www.knime.com/products

Lanjewar, U. M., & Shah, J. J. (2012). Air Pollution Monitoring & Tracking System Using Mobile Sensors and Analysis of Data Using Data Mining. *International Journal of Advanced Computer Research*, (4), 2277–7970.

Lavorato, M., Moscatelli, F., Pagura, M. R., Zanin, M., & Cesarano, P. (2010). *Calibración De Un Medidor De Radiación UVE*. *22*(2), 105–108.

Li, H., Fan, H., & Mao, F. (2016). A visualization approach to air pollution data exploration-A case study of air quality index (PM2.5) in Beijing, China. *Atmosphere*, *7*(3). https://doi.org/10.3390/atmos7030035

Li, J., Xiao, Z., Zhao, H. Q., Meng, Z. P., & Zhang, K. (2016). Visual analytics of smogs in China. *Journal of Visualization*, *19*(3), 461–474. https://doi.org/10.1007/s12650-015-0338-2

Liao, Z., Peng, Y., Li, Y., Liang, X., & Zhao, Y. (2014). A web-based visual analytics system for air quality monitoring data. *Proceedings - 2014 22nd International Conference on Geoinformatics, Geoinformatics 2014*. https://doi.org/10.1109/GEOINFORMATICS.2014.6950834

Liu, D. Y. (2012). *Visualization of Multivariate Data*.

Liu, Y. (2010). Visualization of Multivariate Data. *Department of Biomedical, Industrial and Human Factors Engineering Wright State University*, 115–146. https://doi.org/10.1201/9780429192760-5

Martínez-Ataz, E., & Díaz, Y. (2004). *Contaminación atmosférica* (Vol. 45). Univ de Castilla La Mancha.

Maté, T., Guaita, R., Pichiule, M., Linares, C., & Díaz, J. (2010). Short-term effect of fine particulate matter (PM 2.5) on daily mortality due to diseases of the circulatory system in Madrid (Spain). *Science of the Total Environment*, *408*(23), 5750–5757. https://doi.org/10.1016/j.scitotenv.2010.07.083

Matute Rivera, M. A. (2018). *Evaluación de las herramientas de minería de datos en variables de contaminación atmosférica* (Universidad del Azuay). Recuperado de: http://dspace.uazuay.edu.ec/handle/datos/8203

Morreale, P., Goncalves, A., & Silva, C. (2015). *Modeling and Processing for Next-Generation Big-Data Technologies*. *4*, 211–229. https://doi.org/10.1007/978-3-319-09177-8

OMS. (2017). *¿Cuál Es El Panorama General?* 5.

Ontaneda, I. (2017). *MEDICIÓN DEL CONSUMO DE ENERGÍA EN UN NODO SENSOR INALÁMBRICO EN LA TRANSMISIÓN DE VIDEO SOBRE IPV6 EN APLICACIONES DE DOMÓTICA.*

Ortega Guamán, J. J. (2018). *Impacto de la aplicación de algoritmos de minería de datos en variables de contaminación del air* (UNIVERSIDAD DEL AZUAY). Recuperado de: http://dspace.uazuay.edu.ec/handle/datos/7800

Package Control. (2019). Installation Package Control. Recuperado de: https://packagecontrol.io/installation

Paulovich, F. V., Eler, D. M., Poco, J., Botha, C. P., Minghim, R., & Nonato, L. G. (2011). Piecewise laplacian-based projection for interactive data exploration and organization. *Computer Graphics Forum*, *30*(3), 1091–1100. https://doi.org/10.1111/j.1467-8659.2011.01958.x

Peña, M. (2018). *Visualización de grandes volúmenes de datos Silvia Castro Luján Ganuza Mario Peña*. 10.

PHP. (2019). PHP. Recuperado de: https://www.php.net/manual/es/intro-whatis.php

Puolamäki, K., Papapetrou, P., & Lijffijt, J. (2010). Visually controllable data mining methods. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 409–417. https://doi.org/10.1109/ICDMW.2010.141

Ramos-Herrera, S., Bautista-Margulis, R., & Valdez-Manzanilla, A. (2010). ESTUDIO ESTADÍSTICO DE LA CORRELACIÓN ENTRE CONTAMINANTES ATMOSFÉRICOS Y VARIABLES METEOROLÓGICAS EN LA ZONA NORTE DE CHIAPAS, MÉXICO. *División Académica de Ciencias Biológicas.*, *14*, 16.

Rapidminer, I. (2019). RapidMiner Studio. Recuperado de: https://rapidminer.com/products/studio/

SAS Institut Inc. (2019). SAS Enterprise Miner.

Sellers, C. A. (2013). *Publicación de los contaminantes atmosféricos de la estación de monitoreo en tiempo real de la ciudad de Cuenca, utilizando servicios estándares OGC* (Universidad del Azuay; Vol. 2). Recuperado de: http://dspace.uazuay.edu.ec/handle/datos/2546

Shi, W., Wong, M. S., Wang, J., & Zhao, Y. (2012). Analysis of airborne particulate matter (PM2.5) over Hong Kong using remote sensing and GIS. *Sensors (Switzerland)*, *12*(6), 6825–6836. https://doi.org/10.3390/s120606825

Sublime Text. (2017). Sublime Text 3. Recuperado de: https://www.sublimetext.com/3

Thomas, J. J., Lokanathan, R., & Jothi, J. A. (2017). *Parallel Coordinates Visualization Tool on the Air Pollution Data for Northern Malaysia*. 271–284. https://doi.org/10.1007/978-3-319-66984-7_16

Torres, W. H. (2002). <*BIOLOGÍA DE LAS ESPECIES DE OXÍGENO REACTIVAS.pdf*>. *XXVI*, 19–54.

Universidad de Piura. (2015). *Capitulo 2 : Precipitacion*. 9–26. Recuperado de: http://www.biblioteca.udep.edu.pe/bibvirudep/tesis/pdf/1_136_147_89_1257.pdf

University of Waikato. (2019). Data Mining Software in Java. Recuperado de: http://www.cs.waikato.ac.nz/ml/weka/

Vaisala. (2013). *Punto de rocío en el aire comprimido*. 1–4. Recuperado de: https://goo.gl/ats0aA

VyGLab. (2015). Information Visualization of Multivariate Data. *Laboratorio de Visualización y Computación Gráfica Dpto. de Cs. E Ing. de La Computación - UNS*, 61. Recuperado de: http://cadics.isy.liu.se/?page=informationvisualizationofmultivariatedata

W3S. (2019). W3School. Recuperado de: https://www.w3schools.com/

Wonder, U. (2018). *North America Land Data Assimilation System (NLDAS) Daily Air Temperatures and Heat Index (1979-2011) Request. Accessed on Dec. 1, 2018*.

Xiaoru Yuan, Peihong Guo, He Xiao, Hong Zhou, & Huamin Qu. (2009). Scattering Points in Parallel Coordinates. *IEEE Transactions on Visualization and Computer Graphics*, *15*(6), 1001–1008. https://doi.org/10.1109/tvcg.2009.179

Zhang, J., Huang, M. L., Wang, W. B., Lu, L. F., & Meng, Z. P. (2015). Big data density analytics using parallel coordinate visualization. *Proceedings - 17th IEEE International Conference on Computational Science and Engineering, CSE 2014, Jointly with 13th IEEE International Conference on Ubiquitous Computing and Communications, IUCC 2014, 13th International Symposium on Pervasive Systems*, , 1115–1120. https://doi.org/10.1109/CSE.2014.219

Zhao, D. (2013). *Frontiers of Big Data Business Analytics*. https://doi.org/10.1201/b14700-4

Zhou, Z., Ye, Z., Liu, Y., Liu, F., Tao, Y., & Su, W. (2017). Visual analytics for spatial clusters of air-quality data. *IEEE Computer Graphics and Applications*, *37*(5), 98–105. https://doi.org/10.1109/MCG.2017.3621228