ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENIERÍA DE SISTEMAS

UNIDAD DE TITULACIÓN

DISEÑO DE UNA ARQUITECTURA DE DATOS MASIVOS PARA EL MAPEO GEOGRÁFICO DE LA CONTAMINACIÓN DEL AIRE EN EL DMQ ENTRE LOS AÑOS 2005 AL 2020

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL GRADO DE MAGÍSTER EN SISTEMAS DE INFORMACIÓN, MENCIÓN EN INTELIGENCIA DE NEGOCIOS Y ANÁLITICA DE DATOS MASIVOS

MARÍA GABRIELA MORA VILLACÍS
maría.mora01@epn.edu.ec

Director: PhD. Tania Elizabeth Calle Jiménez
tania.calle@epn.edu.ec

Codirector: PhD. Lorena Katherine Recalde Cerda
lorena.recalde@epn.edu.ec

2022

APROBACIÓN DEL DIRECTOR

Como directora del trabajo de titulación “DISEÑO DE UNA ARQUITECTURA DE DATOS MASIVOS PARA EL MAPEO GEOGRÁFICO DE LA CONTAMINACIÓN DEL AIRE EN EL DMQ ENTRE LOS AÑOS 2005 AL 2020”, desarrollado por María Gabriela Mora Villacís, estudiante de la Maestría de Sistemas de Información, mención en Inteligencia de Negocios y Analítica de Datos Masivos, habiendo supervisado la realización de este trabajo y realizado las correcciones correspondientes, doy por aprobada la redacción final del documento escrito para que prosiga con los trámites correspondientes a la sustentación de la Defensa oral.

Firmado electrónicamente por:
TANIA
ELIZABETH
CALLE JIMENEZ

PhD. Tania Calle

DIRECTORA

APROBACIÓN DEL CODIRECTOR

Como codirectora del trabajo de titulación “DISEÑO DE UNA ARQUITECTURA DE DATOS MASIVOS PARA EL MAPEO GEOGRÁFICO DE LA CONTAMINACIÓN DEL AIRE EN EL DMQ ENTRE LOS AÑOS 2005 AL 2020”, desarrollado por María Gabriela Mora Villacís, estudiante de la Maestría de Sistemas de Información, mención en Inteligencia de Negocios y Analítica de Datos Masivos, habiendo supervisado la realización de este trabajo y realizado las correcciones correspondientes, doy por aprobada la redacción final del documento escrito para que prosiga con los trámites correspondientes a la sustentación de la Defensa oral.

LORENA KATHERINE RECALDE CERDA

Firmado digitalmente por LORENA KATHERINE RECALDE CERDA
Fecha: 2022.12.28 21:59:19 -05'00'

PhD. Lorena Recalde

CODIRECTORA

DECLARACIÓN DE AUTORÍA

Yo, María Gabriela Mora Villacís, declaro bajo juramento que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentada para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

La Escuela Politécnica Nacional puede hacer uso de los derechos correspondientes a este trabajo, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normatividad institucional vigente.

María Gabriela Mora Villacís

DEDICATORIA

A Dios. Por Él soy lo que soy y alcanzo lo que he logrado.
Tu voluntad sea siempre sobre la mía.

A mi ñaña María José. Desde que te conté sobre este paso, me apoyaste. Gracias por tu motivación, por alegrarte en cada meta cruzada, por ser incondicional.

A mis padres. A pesar de no entender mucho
lo que estudié, siguen a mi lado y me respaldan.
Sé que celebran este logro conmigo.

A todos los profesionales que creen que no es posible salirse de su zona de confort. El cuarto nivel no solo sirve para especializar, sino para complementar.

Gabriela Mora V.

AGRADECIMIENTO

Gracias a Dios por darme la inteligencia para subir este escalón, a pesar de no ser mi área de estudios. Él sabe que la programación siempre me apasionó y me dio esta oportunidad de cumplir mi sueño.

Gracias ñaña, Alexis y Alex. Por ustedes di este gran paso, me impulsaron a presentar mi postulación a pesar de mi temor. Siempre confiaron en mí.

Gracias a mis padres y el resto de mi familia por sus palabras de ánimo, por compartir conmigo los momentos felices y no tan felices. Ustedes son mi motor.

Gracias Samu. Orar por mí cuando sentía que era demasiado, me daba fortaleza. Eres bálsamo, oasis y luz en mi vida.

Gracias infinitas amigos queridos, Edu y Andresito, por el conocimiento compartido y la paciencia. Las amanecidas y cafecitos virtuales en cada proyecto hicieron única esta aventura. De ustedes he aprendido tanto.

Gracias Panchito y Bri, compañeros de trabajo y amigos. Sus valiosas ideas aportaron enormemente en el desarrollo de mi tesis; gracias por ayudarme a caminar en esta nueva ruta, en este comienzo en el Big Data.

Y un agradecimiento especial a Tañita y a Lore que estuvieron durante este trayecto brindándome su conocimiento sin escatimar y aliento para culminar este trabajo. Tañita, usted es mi inspiración porque, desde que la conocí, confirmé que es posible combinar la Geografía y la Informática.

ÍNDICE DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

RESUMEN

Este trabajo de investigación tiene el propósito de integrar el procesamiento y visualización de datos geográficos relacionados a contaminación del aire, dentro de una arquitectura de datos masivos. Para alcanzarlo, se utilizaron datos de contaminantes atmosféricos de la ciudad de Quito, medidos a lo largo de la Red de Monitoreo Atmosférico, los cuales se expresaron mediante un índice. La arquitectura propuesta es open-source y se compone de un nodo máster y 2 workers, que consiste en un sistema computacional de análisis unificado en Spark, administrado por Yarn y enlazado a una interfaz gráfica proporcionada por Zeppelin; mismo que almacena los datos en HDFS, los procesa y los muestra visualmente mediante mapas geográficos. Para medir su eficiencia, se realizó una comparativa de tiempos de respuesta con y sin el empleo del sistema. Sin su uso, el proceso de interpolación geográfica tomó un tiempo de 4.52 segundos; mientras que el sistema propuesto mostró un tiempo de ejecución de 2.0 segundos, obteniendo una reducción del 56%, mejorando así el camino tradicional de interpolación y visualización de mapas y generando una nueva alternativa open-source con optimización de recursos y tiempo, además de contribuir en la toma de decisiones estratégicas mediante una nueva forma de análisis las problemáticas ambientales.

Palabras clave: Geo-computación paralela, arquitectura de Big Data, contaminación del aire, mapeo geográfico.

ABSTRACT

The aim of this research is to integrate the processing and visualization of geographic data related to air pollution, within a massive data architecture. To achieve it, air pollutant data from Quito, Ecuador were used; it is measured by Atmospheric Monitoring Network of the city and were expressed by an index. The proposed architecture is open-source and is made up of a master node and two workers: it consists of a unified analysis computational system in Spark, managed by Yarn and linked to a graphical interface provided by Zeppelin; which stores data in HDFS, processes it and displays it visually through geographic maps. To measure its efficiency, a response time comparison was made with and without the system. Without its use, geographic interpolation took 4.52 seconds; while the proposed system showed an execution time of 2.0 seconds, obtaining a reduction of 56%. This shows an improvement in the traditional interpolation and map visualization processes and generating a new open-source alternative with resources and time optimization; in addition to contribute to making strategic decisions through a new way of analyzing environmental problems.

Keywords: Parallel geo-computation, Big Data architecture, air pollution, geographic mapping.

1. INTRODUCCIÓN

La contaminación del aire es un problema que ha traído consecuencias negativas a las poblaciones urbanas desde hace siglos, siendo las partículas y gases presentes en el aire los más nocivos para la salud humana [1]. Según la Organización Mundial de la Salud (OMS), 9 de cada 10 personas en todo el mundo respiran aire con altos niveles de contaminantes (90% de la población) y aproximadamente 7 millones de personas mueren al año por contaminación de aire ambiente y doméstico [2]. El material particulado, especialmente las partículas menores o iguales a 2.5 micrones de diámetro, es el más peligroso para la salud humana, ya que causa enfermedades cardiovasculares, respiratorias y cáncer [3].

Un estudio de la OMS muestra que para el año 2012, en el Ecuador, fallecieron 1771 personas por enfermedades relacionadas directamente con la contaminación del aire, 86 de los cuales son niños; siendo la cardiopatía isquémica el padecimiento que destaca [4]. Estas estadísticas son preocupantes. En particular, en el Distrito Metropolitano de Quito (DMQ), la situación sigue la misma tendencia. En el 2018 hubo 63 días en los que los quiteños respiraron aire perjudicial para su salud, puesto que los contaminantes encontrados en el ambiente superaron los límites permisibles; siendo la principal razón la circulación de autos y buses que utilizan combustibles fósiles [5].

En ese sentido, el DMQ tiene como competencia la elaboración de su Plan de Ordenamiento Territorial (PDOT), el cual incluye, entre varias temáticas, un diagnóstico de la situación ambiental de la ciudad. En dichos diagnósticos, se observa que los indicadores están representados de manera “plana”, es decir, mediante tablas y gráficos estadísticos (líneas, barras), como se aprecia en la Figura 1. Esto pone en evidencia el desconocimiento de la importancia de los mapas geográficos como complemento para visualizar la información, además del apoyo que ofrecen en la toma estratégica de decisiones para el desarrollo de políticas públicas que promuevan una disminución en la contaminación del aire y la mejora en la salud de los habitantes.

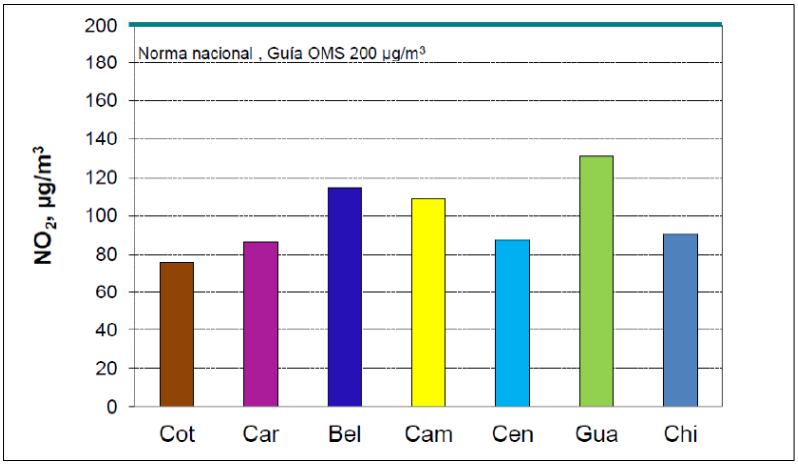


Figura 1 - Ejemplo de presentación de datos de contaminación del PDOT [6].

Por otro lado, el volumen de datos que las organizaciones generan, tanto públicas como privadas, hace imposible su análisis manual [7], resultando útil el uso de arquitecturas de datos masivos, basados en el procesamiento paralelo o distribuido. Éstos se han convertido en elementos clave para dar soporte en cuanto a la gestión eficiente de la información y la toma estratégica de decisiones [8], siendo éste último un tema de investigación emergente en las últimas décadas y considerado como una actividad organizacional primordial [9]. En ese sentido, la institución pública conocida como Red Metropolitana de Monitoreo Atmosférico de Quito (REMMAQ) genera a diario, desde el año 2004, cientos de datos por hora de 13 diferentes variables en aproximadamente 11 estaciones distribuidas en toda la ciudad; información que resulta cada vez más compleja de almacenar, manejar y analizar.

Por tales razones, en el presente trabajo se propone el aprovechamiento de los sistemas distribuidos implementando una arquitectura open-source de Big Data provista por Apache, que consiste en un sistema computacional de clústeres abiertos y análisis unificado en Spark, administrado por Yarn y enlazado a una interfaz gráfica proporcionada por Zeppelin. El sistema propuesto almacena los datos de contaminación del aire en el DMQ (entre 2005 y 2020) en Hadoop Distributed File System (HDFS), los procesa y los muestra visualmente mediante mapas geográficos. Los contaminantes considerados para el análisis fueron monóxido de carbono (CO), dióxido de nitrógeno (NO_2), ozono (O_3), dióxido de azufre (SO_2) y material particulado ($PM_{2.5}$), extraídos de las estaciones de monitoreo localizadas en Belisario, Carapungo, Centro y Cotocollao. Adicionalmente, se emplearon variables climáticas para complementar el estudio como dirección del viento, humedad, precipitación y presión, obtenidos de las estaciones de Belisario, Carapungo, Cotocollao, El Camal, Los Chillos y Tumbaco. Este trabajo de

investigación permitió comprender cómo la representación espacial puede apoyar el proceso de toma de decisiones a nivel gubernamental e introducir nuevas tecnologías a la resolución de problemas actuales como las arquitecturas de Big Data.

1.1. Objetivo general

Diseñar una arquitectura de datos masivos para el mapeo geográfico de la contaminación del aire en el DMQ entre los años 2005 al 2020.

1.2. Objetivos específicos

- Realizar una revisión sistemática de literatura relacionada al tópico a estudiar.
- Efectuar la descarga, revisión y limpieza de datos sobre contaminantes y variables meteorológicas.
- Calcular un índice que exprese la contaminación de aire medida en cada una de las estaciones de monitoreo.
- Diseñar una arquitectura para almacenamiento y procesamiento de datos masivos de contaminación de aire e integrarla al proceso de representación mediante mapas geográficos.
- Evaluar la situación del DMQ en los últimos años respecto a la contaminación del aire mediante un análisis de los resultados obtenidos.

1.3. Alcance

Este proyecto plantea implementar un clúster para procesamiento distribuido de datos tabulares relacionados a contaminación del aire y su visualización en una interfaz gráfica a través de superficies continuas o mapas geográficos. No incluye algoritmos de Machine Learning para limpieza de datos o predicción de contaminación, ni evaluación de métodos de interpolación geoespacial.

1.4. Marco Teórico

1.4.1. Sistema distribuido

Es un sistema que consta de un conjunto de computadoras conectadas por una red, llamadas generalmente nodos, que coordinan actividades y comparten recursos, de tal forma que el usuario percibe un cómputo integrado, aunque dichas máquinas se localicen en diferentes lugares [10]. Las ventajas de utilizar este tipo de sistemas son:

escalabilidad, ya que puede expandirse fácilmente añadiendo cuantas computadoras sean necesarias; *redundancia,* puesto que los servicios e información se almacenan en todas las máquinas, y si una de ellas no está disponible, el trabajo no se detiene, lo cual se traduce en *disponibilidad* [11]; mejoran el rendimiento de las aplicaciones a través del procesamiento múltiple; *confiabilidad*, *extensibilidad* y *portabilidad* a través de la modularidad; y *rentabilidad* a través de recursos compartidos y sistemas abiertos [12].

En la era de la información masiva, los sistemas distribuidos han tomado fuerza y se ha popularizado su uso entre las organizaciones, ya que permiten manejar de mejor manera la gran cantidad de datos que generan y solventar las crecientes necesidades de rendimiento de las aplicaciones [13].

1.4.1.1. Apache Spark

Es un motor de análisis unificado, basado en computación distribuida (o en clúster) de código abierto, utilizado en el procesamiento de datos masivos o Big Data [14]. Hoy en día se ha convertido en una herramienta de amplio uso debido a las ventajas que ofrece, entre ellas se puede mencionar que es muy rápido y confiable para análisis de datos, ejecución de programas y escritura de datos; proporciona interfaces de programación de aplicaciones (APIs) en diferentes lenguajes como Java, Scala, Python y R; está diseñado, tanto para el procesamiento en memoria, como para el procesamiento en disco; es sencillo de instalar y tiene la capacidad de reunir conjuntos de datos provenientes de múltiples y diversas fuentes [15].

Apache Spark en modo clúster está controlado por un administrador (Cluster Manager) y se compone de maestro, -instancia que contiene el Driver Program- y varios esclavos, nodos o workers, -instancias que contienen a los ejecutores. Los workers pueden instalarse en un mismo nodo (un servidor) o en diferentes nodos (clúster EMR con múltiples instancias EC2) [16]. Esta arquitectura se conoce como master-slave y se puede observar en la Figura 2.

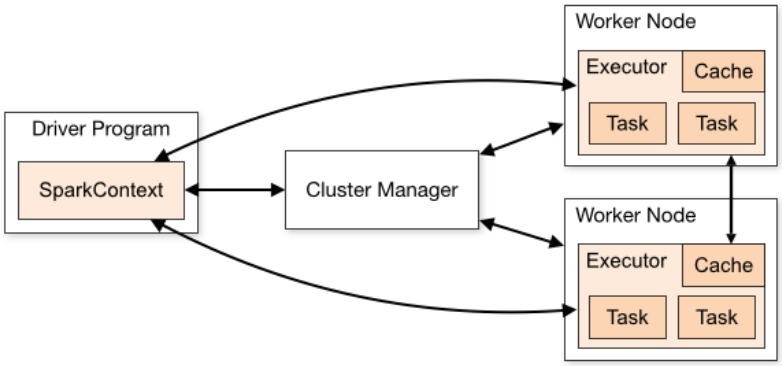


Figura 2 - Arquitectura de Apache Spark [17].

SparkContext, que actúa como un Driver Program, se conecta a uno de los administradores de clúster, cuya función es asignar recursos a las aplicaciones que se van a ejecutar. Una vez conectado, Spark obtiene ejecutores en los nodos del clúster, que son procesos que realizan cálculos y almacenan datos para la aplicación. A continuación, envía el código de la aplicación (definido por archivos JAR o Python pasados a SparkContext) a los ejecutores. Finalmente, SparkContext envía tareas a los ejecutores para que las ejecuten [17].

Los administradores del clúster podrían ser [17]:

- *Standalone:* Es el incluido en Spark.
- *Apache Mesos:* Es un administrador general que puede ejecutar Hadoop Mapreduce y aplicaciones de servicio. Actualmente, se encuentra obsoleto.
- *Hadoop Yarn:* Es el administrador de recursos propio de Hadoop 2.
- *Kubernetes:* Es un sistema open-source que automatiza la implementación, el escalado y la gestión de aplicaciones en contenedores.

1.4.2. Datos geográficos

Según IBM, los datos geográficos o geoespaciales son aquellos basados en el tiempo y que están relacionados con una ubicación específica en la superficie de la Tierra, que al procesarlos pueden aportar información sobre las relaciones existentes entre variables y revelar patrones y tendencias. La localización proporcionada puede ser estática a corto plazo, por ejemplo: la ubicación de un equipo, un terremoto, niños que viven en la pobreza; o dinámica, por ejemplo: un vehículo o peatón en movimiento, la propagación de una enfermedad infecciosa, entre otros [18].

Es importante reconocer lo valioso de este tipo de datos y el conocimiento que se puede extraer de los mismos para comprender a profundidad una gran cantidad de problemas. Los datos espaciales pueden combinarse con otros provenientes de varias fuentes para aumentar su potencial. Datos meteorológicos, de censos, imágenes satelitales, fotografías aéreas e incluso datos de redes sociales pueden contribuir a la construcción de visualizaciones que explican mejor el mundo que nos rodea [19].

1.4.3. Interpolación espacial

La interpolación espacial es una parte de la Geoestadística que permite inferir o estimar una variable ligada a puntos desconocidos a partir de valores conocidos con coordenadas, obteniendo como resultado un mapa continuo o superficie ráster [20], que vendría a ser una matriz de píxeles con un valor z determinado, como se muestra en la Figura 3.

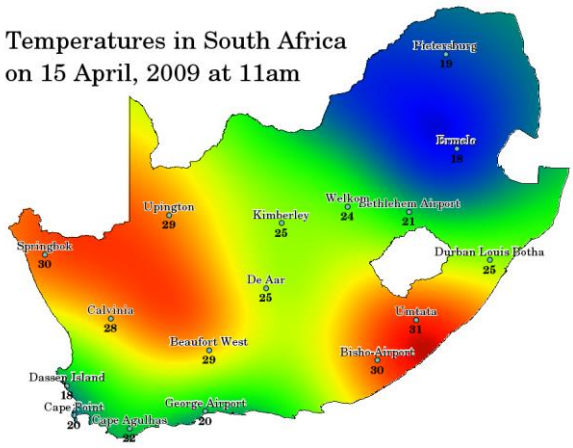


Figura 3 - Superficie continua creada con interpolación espacial [20].

Esta metodología es utilizada para modelar la distribución espacial y espacio-temporal de los fenómenos naturales, físicos y socioeconómicos [21]; además brinda la ventaja de ahorrar tiempo y recursos, ya que estos eventos requieren una recopilación de datos de campo que puede ser costosa y en ciertos lugares incluso inaccesible. Ejemplos de estos fenómenos son: precipitación, dispersión de ruido, temperatura, elevación, concentración de contaminantes, entre otros [22].

1.4.3.1. Kriging

Kriging es una “técnica de estimación local del mejor estimador lineal imparcial para los valores desconocidos de las variables espaciales y temporales” [23]. Se expresa mediante la siguiente ecuación:

$$Z^*_K = \sum_{i=1}^{n} \lambda_i Z_i$$

donde Z^*_K es la estimación por kriging, λ_i es un peso para Z_i y Z_i es una variable. Se incluye el peso como garantía de eliminación del sesgo en el estimador y para que la varianza de dicha estimación sea mínima [24].

Este método es comúnmente utilizado en las geociencias porque se apoya en el conocimiento del comportamiento de la variable en el espacio [25] y se puede llevar a cabo en un contexto estacionario o no estacionario. Si la media es conocida, se opta por kriging simple; sin embargo, esto ocurre en pocas ocasiones. En cambio, cuando la media es desconocida, se elige kriging ordinario y es el caso de los estudios centrados en el calentamiento global [26].

1.4.4. Índice de Contaminación del Aire (ICA)

Es un valor adimensional que permite monitorear la calidad del aire que miden las estaciones de monitoreo y, por ende, conocer la evolución del estado del aire; además de ser la base para emitir recomendaciones sanitarias a la población en general y especialmente la sensible [27].

El Municipio del DMQ ha diseñado su propio índice de calidad del aire, adaptado a sus necesidades y entorno, y ha sido denominado Índice Quiteño de la Calidad del Aire (IQCA). Sus datos fuente provienen de la Red Metropolitana de Monitoreo Atmosférico (REMMAQ) de la Secretaría de Ambiente, compuesta de estaciones remotas localizadas en el área urbana de la ciudad y los valles aledaños, con la capacidad de medir continuamente los contaminantes más importantes asociados a la calidad del aire: material particulado fino ($PM_{2.5}$), óxidos de nitrógeno expresados como dióxido de nitrógeno (NO_2), dióxido de azufre (SO_2), monóxido de carbono (CO) y oxidantes fotoquímicos expresados como ozono (O_3) [28].

El IQCA se encuentra en una escala numérica entre 0 y 500; mientras más alto es el valor quiere decir que es mayor el nivel de contaminación atmosférica y viceversa. A su vez, se ha dividido en varias categorías para expresar su peligrosidad, como lo muestra la Tabla 1.

Tabla 1 - Límites numéricos de cada categoría del IQCA (µg/m3) [28]

Rango	Categoría	CO^{a}	$O_3{}^{b}$	$NO_2{}^{c}$	$SO_2{}^{d}$	$PM_{2.5}{}^{e}$	$PM_{10}{}^{f}$
0–50	Nivel deseable u óptimo	0–5000	0–50	0–100	0–62.5	0–25	0–50
51–100	Nivel aceptable o bueno	5001–10000	51–100	101–200	63.5–125	26–50	51–100
101–200	Nivel de precaución	10001–15000	101–200	201–1000	126–200	51–150	101–250
201–300	Nivel de alerta	15001–30000	201–400	1001–2000	201–1000	151–250	251–400
301–400	Nivel de alarma	30001–40000	401–600	2001–3000	1001–1800	251–350	401–500
401–500	Nivel de emergencia	>40000	>600	>3000	>1800	>350	>500

Notas: a, concentración máxima de promedio en 8 horas; b, concentración máxima de promedio de 8 horas; c, concentración máxima en 1 hora; d, concentración promedio en 24 horas; e, concentración promedio en 24 horas; f, concentración promedio en 24 horas

La Tabla 2 describe cualitativamente cada uno de los rangos mostrados en relación a la salud pública, y al igual que la anterior, posee un código de colores que facilita la comprensión de su significado.

Tabla 2 - Rangos, significados y colores de las categorías del IQCA [28]

Rangos	Condición desde el punto de vista de la salud	Color de identificación
0– 50	Óptima.	Blanco
50– 100	Buena.	Verde
100 –200	No saludable para individuos extremadamente sensibles (enfermos crónicos y convalecientes).	Gris
200 –300	No saludable para individuos sensibles (enfermos).	Amarillo
300 –400	No saludable para la mayoría de la población y peligrosa para individuos sensibles.	Naranja
400 –500	Peligrosa para toda la población.	Rojo

1.4.5. Trabajos relacionados

En las últimas décadas se ha investigado ampliamente los sistemas distribuidos y desarrollado aplicaciones en torno a esta temática. Una de ellas es el ámbito geográfico, como se muestra en [29], cuyo objetivo es la predicción de clases de calidad del aire. Los autores generaron mapas de riesgo mediante el método de Ponderación de Distancia Inversa (IDW, por sus siglas en inglés) de una ciudad para las siguientes 24 horas, a

través de un clúster Hadoop con el framework Spark, mostrando una velocidad aceptable en el procesamiento de grandes datos espaciales y logrando un incremento de la misma utilizando la característica de escalabilidad horizontal de Hadoop. En [30], se propuso mejorar métodos de interpolación espacio-temporal aplicado a datos de contaminación de aire por $PM_{2.5}$ recolectados de 955 estaciones de monitoreo, usando de igual forma Apache Spark; obteniendo un aumento en velocidad y precisión. Este contaminante también fue tratado en [31], en donde se propuso una arquitectura de predicción instantánea de $PM_{2.5}$ basada en Spark para recopilación de datos en tiempo real y predicción de concentración del contaminante a través de una combinación de tres algoritmos de machine learning (regresión lineal, bosque aleatorio, árbol de decisión de aumento de gradiente); los resultados experimentales mostraron que su modelo de predicción tiene el mejor rendimiento (R^2>0,96). También se reportan estudios de diseño de sistemas de interpolación espacial paralela para mejorar la rapidez y precisión en la ejecución de este tipo de algoritmos, que computacionalmente suelen ser costosos; en donde se logró una mejora del tiempo de ejecución de hasta ×100 con una pérdida de precisión mínima (RMSE relativo del 3%) al paralelizar la carga de los conjuntos de datos probados localmente [32]. Además de los estudios de interpolación, se han dado pasos en la identificación de puntos críticos estadísticamente significativos en datos espacio-temporales de gran volumen [33], utilizando el estadístico Gi^* Getis-Ord sobre Spark, el cual corresponde a un análisis de estadística espacial; y aunque los resultados no se muestran a manera de mapas geográficos, se presentó una línea base y dos variantes de una solución optimizada para el problema y se demostró el desempeño de los algoritmos propuestos a través de una evaluación experimental. Incluso se han presentado frameworks que integran el procesamiento masivo y paralelizado con el manejo de datos con características geográficas como GeoSparkViz [34], mismo que facilita al científico de datos un sistema holístico para gestionar y visualizar datos espaciales, además de proponer un método de partición de datos en mosaicos de mapas que logra un equilibrio de carga en el trabajo de visualización de mapas entre todos los nodos del clúster; experimentos muestran que con este framework fue posible generar un mapa de calor de alta resolución de 1700 millones de objetos OpenStreetMap y 1300 millones de viajes en taxi de la ciudad de Nueva York en un tiempo aproximado de 4 y 5 minutos, respectivamente, y sobre un clúster de cuatro nodos.

Los trabajos mencionados anteriormente son una muestra del interés que existe por estudiar el procesamiento de datos geográficos con arquitecturas de Big Data por la cantidad y velocidad con la que se generan, específicamente en el estudio de la calidad

del aire, dada su importancia para la salud pública; sin embargo, la visualización espacial de esta información a través de mapas no es un factor común, lo cual ayudaría en la toma de decisiones gubernamentales. Por otro lado, en la revisión de literatura no se encontraron premisas de trabajos realizados en Ecuador, siendo valedera y novedosa su implementación.

2. METODOLOGÍA

La metodología empleada para el desarrollo de esta investigación es la denominada Action-Research (AR, en español “Investigación-Acción”), desarrollada por Kurt Lewin, la cual plantea una forma de entender los problemas (investigación) y examinarlos a través de la práctica (acción) [35]. Según Oates [36], tiene las siguientes características: (i) se concentra en cuestiones prácticas, (ii) se resume en un ciclo iterativo de planificar-actuar-reflexionar, (iii) se enfoca en el cambio, (iv) debe existir colaboración constante con el/los investigador/es, (v) utiliza múltiples métodos de generación de datos, y (vi) existe relación entre los resultados de la acción y los resultados de la investigación. La Figura 4 muestra las etapas que componen la metodología aplicada, las que se estructuran a manera de ciclo, garantizando así una continua retroalimentación, el desarrollo del conocimiento científico, la solución al problema y los beneficios inherentes hacia los participantes [37].

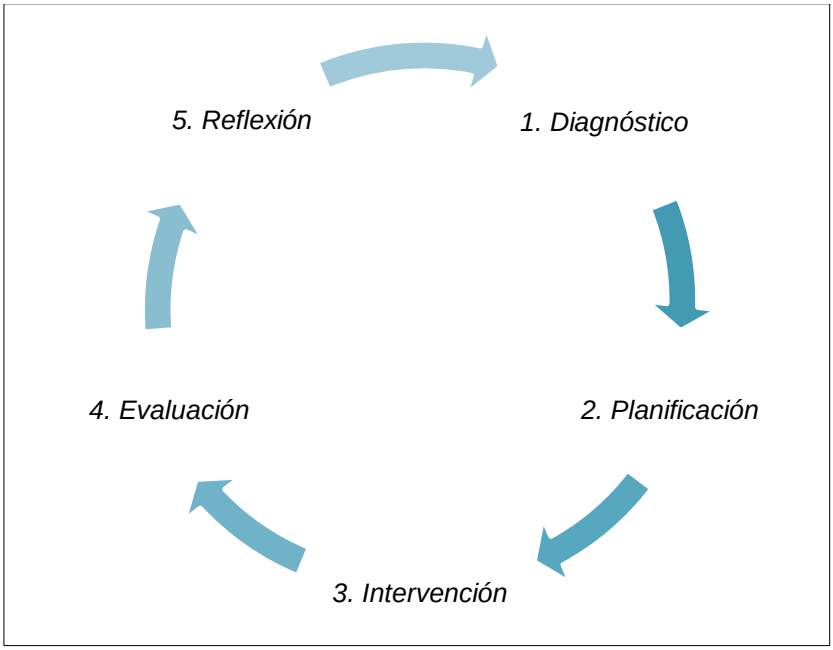


Figura 4 - Metodología AR. Basado en [36].

Cada etapa se compone de las siguientes acciones:

1. **Diagnóstico:** se identifica la naturaleza del problema, incluyendo los factores interrelacionados, además del desarrollo de una base conceptual sobre la situación y cómo podría cambiarse.
2. **Planificación:** se determinan acciones para mitigar el problema o solucionarlo.
3. **Intervención:** se pone en marcha las acciones en el área de aplicación de acuerdo a lo planificado.
4. **Evaluación:** se establece si las acciones se ejecutaron y aliviaron el/los problema/s.

5. **Reflexión:** se decide si se ha alcanzado el objetivo en términos de resultados prácticos y nuevos conocimientos, y si es necesaria la aplicación de un nuevo ciclo AR.

En las siguientes subsecciones se describe con mayor detalle lo realizado en cada una de las fases mencionadas.

2.1. Diagnóstico

Los Art. 42 y 55 del Código Orgánico de Organización Territorial, Autonomía y Descentralización (COOTAD) establecen que los Gobiernos Autónomos Descentralizados (GAD) provinciales y municipales tienen como competencia exclusiva la formulación de sus correspondientes PDOT de manera articulada con la planificación nacional, regional, cantonal y parroquial [38]. En tal virtud, el Municipio del Distrito Metropolitano de Quito, ha desarrollado su plan de ordenamiento para el período 2015-2025, dividido en los tres capítulos establecidos el Art. 42 del Código de Planificación y Finanzas Públicas: diagnóstico, propuesta y modelo de gestión.

Para la elaboración del primer capítulo, los GADs "deberán observar, por lo menos, contenidos que describan las inequidades y desequilibrios socio-territoriales, potencialidades y oportunidades de su territorio, la situación deficitaria, los proyectos existentes en el territorio, las relaciones del territorio con los circunvecinos (…) y, finalmente, el modelo territorial actual" [39]. Es así que el Municipio de Quito ha plasmado su diagnóstico de manera estratégica, dividiéndolo en ejes [6]:

- Eje Ambiental
- Eje Económico
- Eje Social
- Eje Territorial
- Áreas Históricas
- Eje de la Movilidad

Haciendo énfasis en el eje Ambiental, ya que se relaciona con la temática expuesta en este trabajo de investigación, se constató que el contenido evalúa la situación actual del cantón de forma descriptiva y gráfica en cuanto a la huella ecológica, de carbono, hídrica, de residuos sólidos y biodiversidad. En el apartado de huella de carbono, se realiza un análisis en términos de emisiones de Gases de Efecto Invernadero (GEI) que "son

liberados a la atmósfera debido a las actividades de una organización o ámbito geográfico determinado y que consecuentemente contribuyen al calentamiento global" [6]. Se evidenció la existencia de gráficos planos como barras, pasteles y líneas, además de algunos mapas geográficos que complementan el estudio. Sin embargo, no muestran información de verdadera relevancia que permita a las autoridades conocer y entender mejor el medio físico y natural que los rodea y, por consiguiente, planificar acciones y optimizar la toma de decisiones relacionadas a políticas públicas a favor de la población. Si bien se puede apreciar la evolución de la concentración del contaminante (Figura 5), daría mayor información conocer cómo está dispersándose espacialmente en la ciudad y ver la misma evolución a lo largo del tiempo o cuáles son las zonas de mayor y menor contaminación. El mapa presentado (Figura 6) resulta difícil de interpretar, puesto que se encuentra cargado y únicamente muestra la localización de las estaciones que recolectan datos, sin un mapa base que permita al usuario ubicarse.

En vista de lo expuesto, se plantea la necesidad de ir más allá en el tema de visualización y aprovechar herramientas open-source que gestionen grandes volúmenes de datos de manera eficiente. Es decir, se requiere pasar de observar un gráfico comparativo de las concentraciones de contaminantes en cada estación de monitoreo de aire de manera puntual, a un mapa que muestra una superficie continua simulando la dispersión de dichos contaminantes con su componente geográfico para facilitar la ubicación.

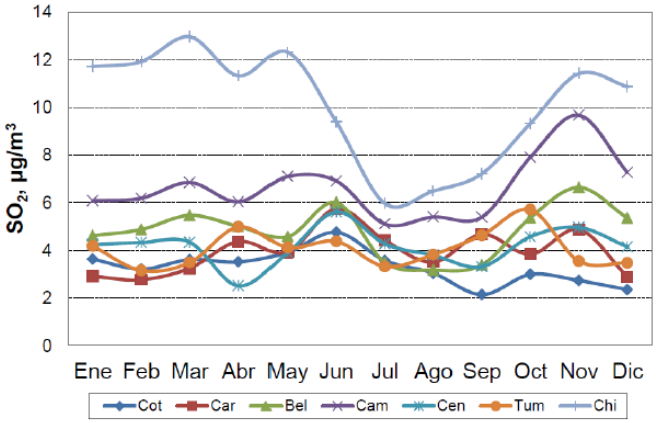


Figura 5 - Ejemplo de gráfico de líneas presentado en el PDOT [6].

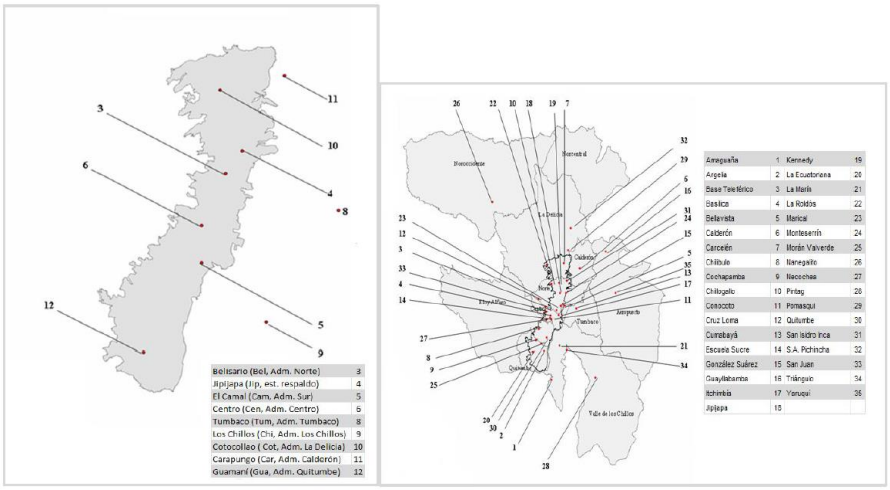


Figura 6 - Ejemplo de mapa geográfico presentado en el PDOT [6].

2.2. Planificación

Una vez realizado el diagnóstico, se ha planificado abordar la solución al problema en base al esquema de la Figura 7. Dentro de la fase de planificación, se ha contemplado actividades de preparación y tratamiento de los datos, básicamente.

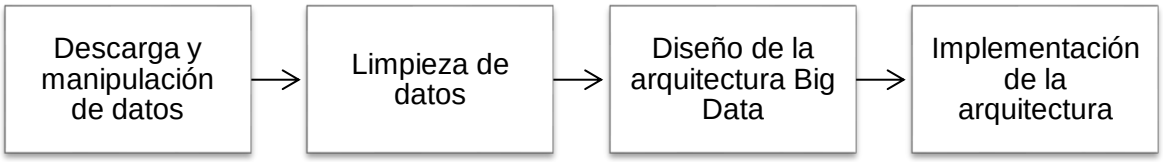


Figura 7 - Planificación de la solución al problema.

2.2.1. Descarga y manipulación de datos

Los datos empleados se encuentran liberados en la página web de la Secretaría de Ambiente del Municipio de Quito[1] en formato *.csv* y son generados por la Red de Monitoreo Atmosférico de la ciudad. El detalle de los archivos descargados se observa en la Tabla 3.

Tabla 3 - Datos descargados de la Red de Monitoreo Atmosférico

Tipo	Elemento	Unidad de medida
Contaminantes	Dióxido de azufre (SO_2)	microgramo por metro cúbico ($\mu g/m^3$)
	Material particulado 2.5 ($PM_{2.5}$)	microgramo por metro cúbico ($\mu g/m^3$)

[1] http://www.quitoambiente.gob.ec/index.php/politicas-y-planeacion-ambiental/red-de-monitoreo

	Material particulado 10 (PM_{10})	microgramo por metro cúbico ($\mu g/m^3$)
	Ozono (O_3)	microgramo por metro cúbico ($\mu g/m^3$)
	Dióxido de nitrógeno (NO_2)	microgramo por metro cúbico ($\mu g/m^3$)
	Monóxido de carbono (CO)	miligramo por metro cúbico (mg/m^3)
Variables climáticas	Temperatura	grados centígrados (°C)
	Humedad	porcentaje (%)
	Precipitación	milímetro (mm)
	Radiación solar	vatio por metro cuadrado (W/m^2)
	Presión barométrica	milibar (mb)
	Radiación ultravioleta	índice ultravioleta (IUV)
	Dirección del viento	grados de azimut (°)

A continuación, se muestra el extracto de uno de los archivos .csv en la Figura 8. Como se puede visualizar, los datos descargados no poseen el componente espacial (coordenadas) necesario para la visualización geográfica; por lo tanto, se procedió a georreferenciar las mismas de manera manual utilizando como base las direcciones de las estaciones encontradas en la página web de la Secretaría del Ambiente[2].

	A	B	C	D	E	F	G	H	I	J	K	L
1		BELISARIO	CARAPUNGO	CENTRO	COTOCOLLAC	ELCAMAL	GUAMANI	LOSCHILLOS	TUMBACO	CONDADO	TURUBAMBA	CHILLOGALLO
2	FECHA \ UNIDAD	ug/m3	ug/m3	ug/m3	ug/m3		ug/m3	ug/m3	ug/m3	ug/m3	ug/m3	ug/m3
3	1/1/2004 0:00			121.37		108.73					65.78	
4	1/1/2004 1:00			82.82		57.96						
5	1/1/2004 2:00			37.27		23.88						
6	1/1/2004 3:00			27.22		16.43						
7	1/1/2004 4:00			16.72		9.51					9.74	
8	1/1/2004 5:00			12.74		11.32						
9	1/1/2004 6:00			15.92		6.57						
10	1/1/2004 7:00			14.07		5.71					10.53	
11	1/1/2004 8:00			12.91		7.65					9.49	
12	1/1/2004 9:00			10.73		5.13						
13	1/1/2004 10:00			8.98		5.54						
14	1/1/2004 11:00			10.78		6.23					7.76	
15	1/1/2004 12:00			9.61		6.25					6.96	

Figura 8 - Extracto de datos relacionados a NO_2.

La Figura 9 muestra la geolocalización de cada estación a lo largo de la ciudad de Quito a través de un mapa geográfico, para mejor comprensión del área de estudio y de los datos disponibles.

[2] http://www.quitoambiente.gob.ec/index.php/politicas-y-planeacion-ambiental/red-de-monitoreo

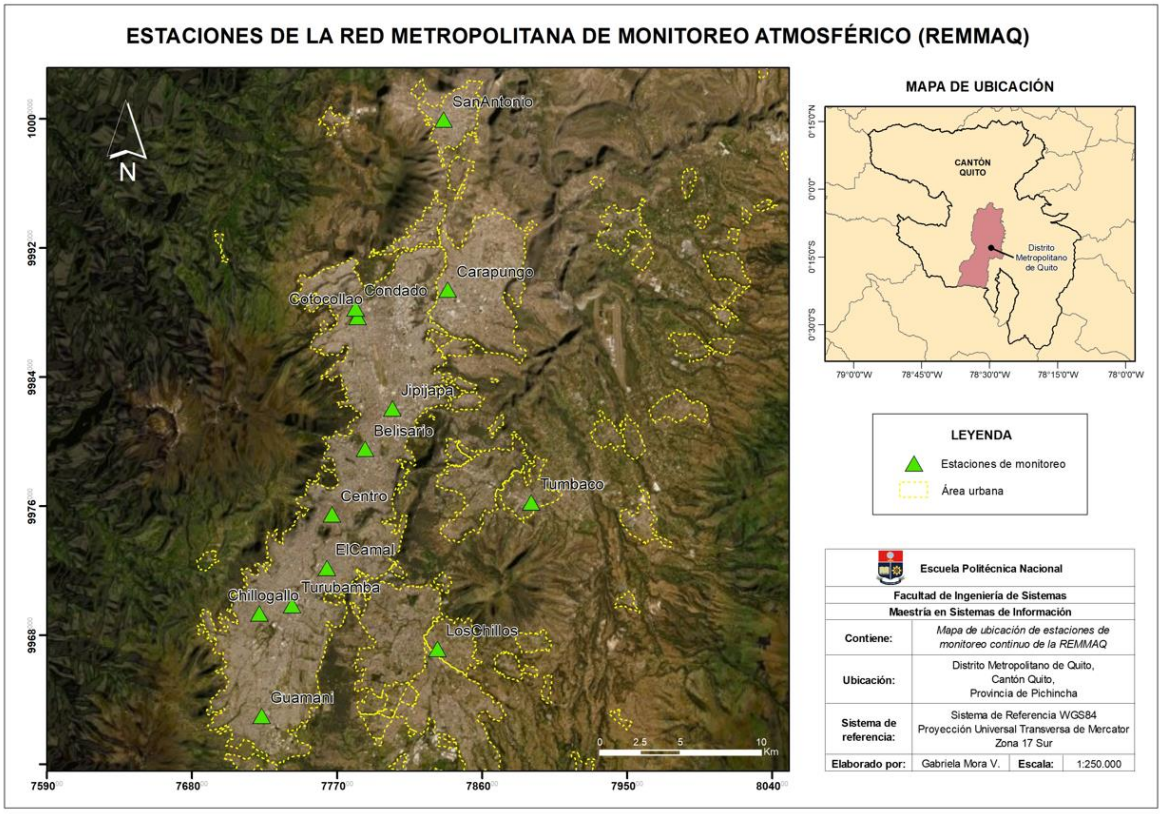


Figura 9 - Mapa de ubicación de las estaciones pertenecientes a la REMMAQ.

2.2.2. Limpieza de datos

Se realizó una revisión rápida de los datos para conocer su naturaleza, lo cual llevó a realizar los siguientes pasos de manera general para limpiarlos:

- *Conversión de tipo de dato* (a 'timestamp' para marcas de tiempo y 'float' para numéricos)
- *Reducción de datos:* De acuerdo a la Tabla 4, los datos se adecuaron a promedios por cada 8 horas o diarios según el contaminante, para el posterior cálculo del IQCA, conociendo que los datos brutos se encuentran por cada hora.

Tabla 4 - Expresiones matemáticas para el cálculo del IQCA [28]

Contaminante	Expresiones matemáticas para cada rango de concentración			
CO, concentración máxima de promedio de 8 horas, mg/m^3	0 < Ci ≤ 10	10 < Ci ≤ 15	15 < Ci ≤ 30	30 < Ci
	IQCA = 10Ci	IQCA = 20Ci – 100.00	IQCA = 6.67Ci + 100.00	IQCA = 10Ci
O_3, concentración máxima de promedios de 8 horas, $\mu g/m^3$	0 < Ci ≤ 100	100 < Ci ≤ 200	200 < Ci ≤ 600	600 < Ci
	IQCA = Ci	IQCA = Ci	IQCA = 0.5Ci + 100.00	IQCA = 0.5Ci + 100.00
NO_2, concentración máxima en 1 hora, $\mu g/m^3$	0 < Ci ≤ 200	200 < Ci ≤ 1 000	1 000 < Ci ≤ 3 000	3 000 < Ci
	IQCA = 0.50Ci	IQCA = 0.125Ci + 75.00	IQCA = 0.1Ci + 100	IQCA = 0.1Ci + 100
SO_2, promedio en 24 horas, $\mu g/m^3$	0 < Ci ≤ 62.5	62.5 < Ci ≤ 125	125 < Ci ≤ 200	200 < Ci
	IQCA = 0.8Ci	IQCA = 1.333Ci - 66.667	IQCA = 0.125Ci + 175.00	IQCA = 0.125Ci + 175.00
$PM_{2.5}$, promedio en 24 horas, $\mu g/m^3$	0 < Ci ≤ 50	50 < Ci ≤ 250	250 < Ci	
	IQCA = 2.00Ci	IQCA = Ci + 50	IQCA = Ci + 50.00	
PM_{10}, promedio en 24 horas, $\mu g/m^3$	0 < Ci ≤ 100	100 < Ci ≤ 250	250 < Ci ≤ 400	400 < Ci
	IQCA = Ci	IQCA = 0.6667Ci+33.333	IQCA = 0.6667Ci + 33.33	IQCA = Ci - 100

Ci: Concentración de un determinado contaminante.

- *Elección de rango de trabajo:* Se utilizaron visualizaciones de datos para analizar la evolución y tendencias en función del tiempo y disponibilidad por contaminante y variable climática (Figura 10, por ejemplo). Se evaluó que exista un período de traslape en aquellos que son necesarios para el cálculo del IQCA (CO, O_3, NO_2, SO_2, $PM_{2.5}$, PM_{10}), eligiendo así el rango de 2005-01-01 a 2020-12-31 en las estaciones de Belisario, Carapungo, Centro y Cotocollao para contaminantes; mientras que Belisario, Carapungo, Cotocollao, El Camal, Los Chillos y Tumbaco se emplearon para el análisis de variables climáticas. Se optó por eliminar el contaminante PM_{10}, debido a que no posee una cantidad de datos significativa para el estudio.
- *Relleno de datos:* Se comprobó que los datos faltantes por cada contaminante no superen el 25% del total, porcentaje de relleno óptimo recomendado en [40]. Se utilizó el método de interpolación (*interpolate* en Python) para completar las series temporales.
- *Suavizamiento de datos:* Se eliminaron los outliers de cada grupo de datos a través del método de la media móvil (*rolling* en Python). La Figura 11 muestra una comparativa entre los datos interpolados (color azul) y los suavizados (color naranja).

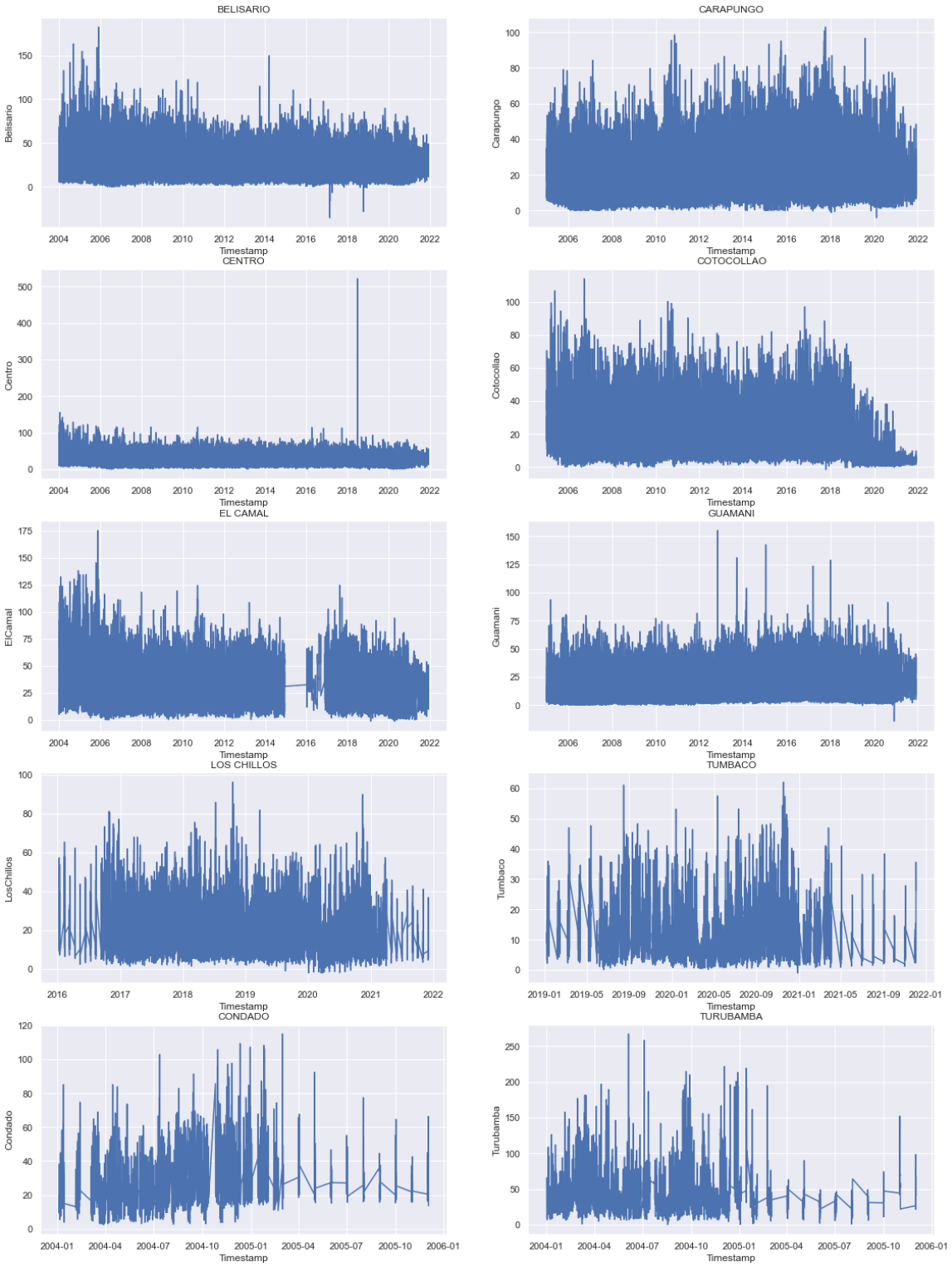


Figura 10 - Medición de NO_2 por estación desde el año 2004.

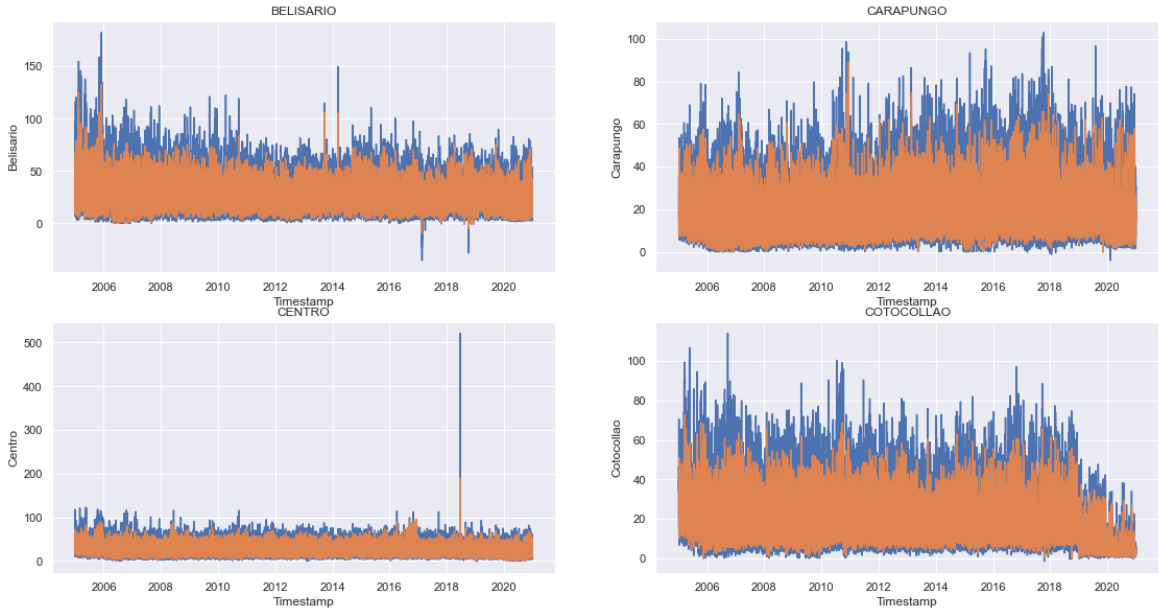


Figura 11 - Comparativa entre datos interpolados y suavizados para NO_2.

Finalmente, se exportaron los datos suavizados en formato .csv para ser almacenados y utilizados (Figura 12).

	A	B	C	D	E
1	Timestamp	Belisario	Carapungo	Centro	Cotocollao
2	1/1/2005 6:00	30.6614286	15.25	52.8128571	22.21
3	1/1/2005 7:00	30.1228571	15.25	48.3785714	22.21
4	1/1/2005 8:00	29.44	15.25	45.4085714	22.21
5	1/1/2005 9:00	27.3828571	15.25	41.69	22.21
6	1/1/2005 10:00	24.7871429	15.25	39.8271429	22.21
7	1/1/2005 11:00	22.2385714	15.25	36.7328571	22.21
8	1/1/2005 12:00	20.3428571	15.25	32.4842857	22.21
9	1/1/2005 13:00	18.6	15.25	28.0485714	22.21
10	1/1/2005 14:00	15.21	15.25	22.3742857	22.21
11	1/1/2005 15:00	12.87	15.25	18.1057143	22.21
12	1/1/2005 16:00	10.58	15.25	16.4285714	22.21
13	1/1/2005 17:00	8.79	15.25	15.1071429	22.21

Figura 12 - Extracto de datos listos para ser procesados (NO_2).

2.2.3. Diseño de la arquitectura

Una vez limpios y estructurados los datos, se procedió a bosquejar la arquitectura que permitiría resolver el problema, la misma que, en su versión definitiva, se describe en la Figura 13.

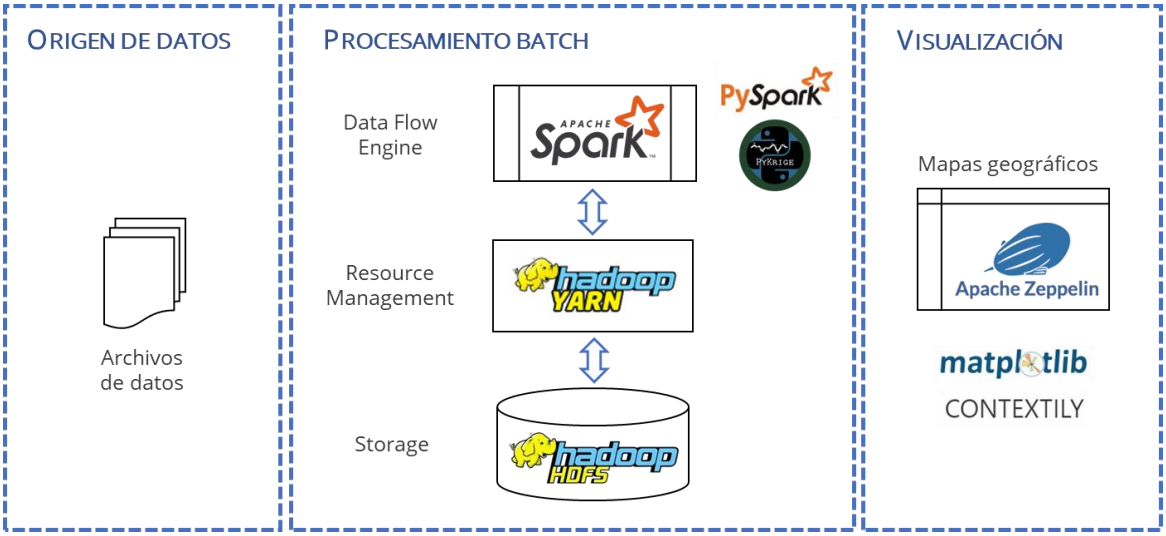


Figura 13 - Arquitectura propuesta.

- *Origen de datos:* Proviene de archivos planos (.csv), resultantes de la fase anterior de limpieza.
- *Procesamiento batch (o por lotes):* permite procesar grandes lotes de trabajos en lotes más pequeños de forma simultánea, en orden secuencial y sin detenerse. Se compone de Hadoop HDFS para almacenar el origen de datos y resultados; Hadoop Yarn, quien se encarga de administrar recursos a cada nodo del clúster; y Apache Spark, que es el motor de procesamiento, utilizando PySpark y PyKrige, principalmente.
- *Visualización:* A través de la interfaz gráfica provista por Apache Zeppelin, se pueden observar mapas geográficos con información interpolada de IQCA y variables climáticas, empleando librerías como MatPlotLib y Contextily.

2.2.4. Implementación de la arquitectura

Después del planteamiento de la arquitectura, arrancó la fase de su implementación, la cual se aborda a continuación.

2.3. Intervención

Esta fase comprende las actividades que se ejecutaron para la construcción del clúster y obtención de resultados, las cuales se pueden apreciar en la Figura 14.

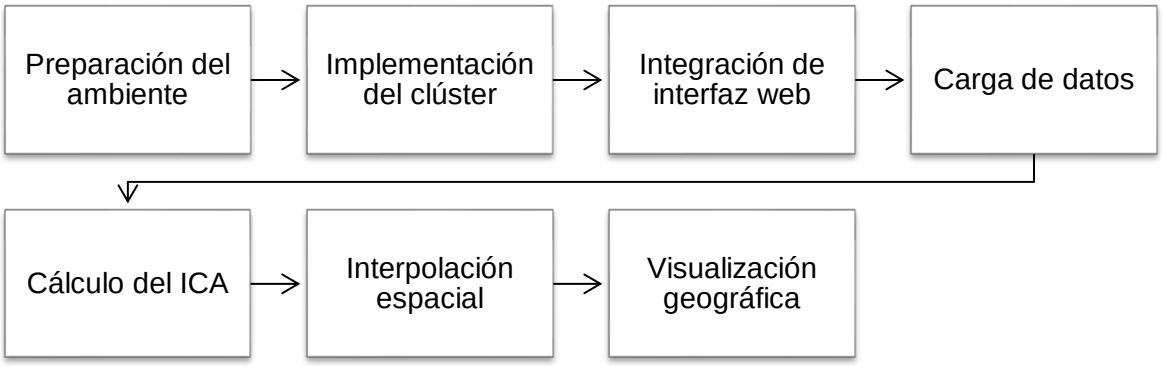


Figura 14 - Pasos en la implementación de la arquitectura planteada.

2.3.1. Preparación del ambiente

La arquitectura propuesta se montó sobre un clúster de computadoras virtuales en Oracle VM VirtualBox, compuesto por un nodo máster y 2 workers, cuya estructura se observa en la Figura 15.

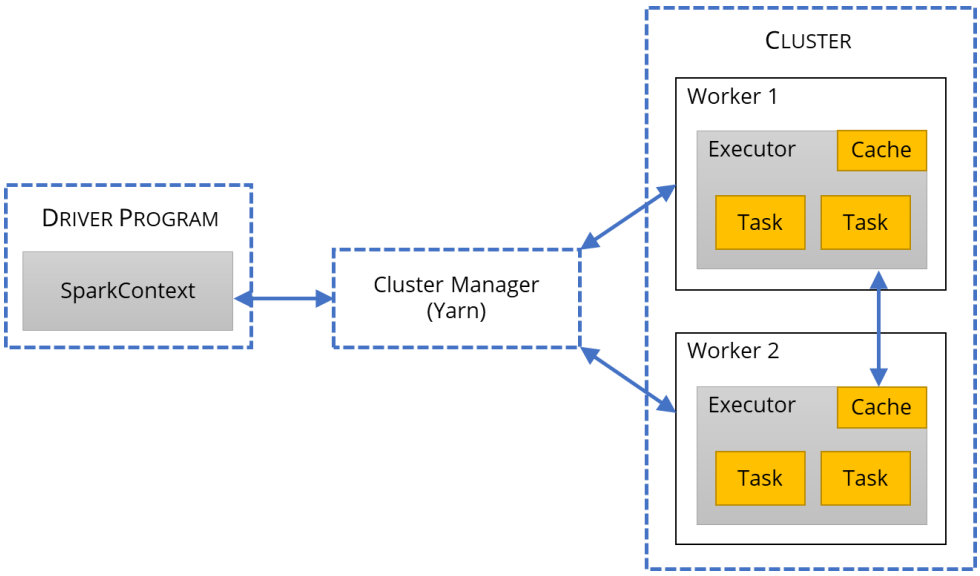


Figura 15 - Clúster construido, basado en [17].

Cada máquina del clúster fue configurada con las características de hardware y software que se describen en la Tabla 5.

Tabla 5 - Configuración del clúster

Característica	Descripción
Sistema Operativo (SO)	CentOS 8 Linux
RAM (Gb)	4
Disco duro (Gb)	50

CPU Cores	1
Versión Hadoop	3.1.2
Versión Spark	2.2.0
Versión Zeppelin	0.9.0

2.3.2. Implementación del clúster

La construcción del clúster incluyó:

- Definición de configuración de hardware y software virtual
- Descarga e instalación de Sistema Operativo
- Configuración de IPs, SSH y hosts para conexión entre máquinas
- Descarga e instalación de Hadoop
- Configuración de variables de entorno
- Configuración del nodo máster (Java-Home, NameNode, HDFS, Yarn)
- Configuración de nodos esclavos
- Formateo y ejecución de HDFS y Yarn
- Descarga e instalación de Spark
- Integración Spark-Yarn
- Configuración de cluster mode

2.3.3. Integración de interfaz web

Dado que las máquinas virtuales instaladas no proveen de una interfaz gráfica, se integró Apache Zeppelin al clúster, el cual es un notebook web al estilo Jupyter. Esta fase incluyó los siguientes pasos:

- Descarga e instalación de Zeppelin
- Configuración de variables de entorno
- Integración Yarn-Zeppelin
- Configuración del intérprete Spark
- Ejecución de Zeppelin

2.3.4. Carga de datos

En primer lugar, los datos limpios deben ser almacenados en HDFS:

```
# Después del comando, se coloca el nombre del archivo y la ruta
# donde se guardará:
hdfs dfs -put interp_co_8h.csv /user/root/tesis
```

Posteriormente, se procedió a conectar desde Zeppelin con HDFS y así cargar en memoria las tablas:

```
# Con la ayuda de Spark, se lee el archivo:
df = spark.read.format('csv')\      # extensión del archivo fuente
      .option('header','true')\     # si el archivo tiene encabezado
      .option('delimiter','|')\     # define que el delimitador es '|'
      .load('hdfs://master-node:9000/user/root/tesis/interp_co_8h.csv')
                        # define la ruta donde se encuentra el archivo
```

2.3.5. Cálculo del IQCA

Recurriendo a la Tabla 6, se generaron UDFs (User Defined Functions) para calcular el índice dependiendo del contaminante, debido a que en Spark no es posible aplicar un cálculo línea por línea de un dataframe. Por ejemplo, para CO, se evalúan las concentraciones medidas en 4 rangos y se aplica una ecuación, como se muestra a continuación:

Tabla 6 - Extracto de cálculo de IQCA para CO

Contaminante	Expresiones matemáticas para cada rango de concentración			
CO, concentración máxima de promedio de 8 horas, mg/m^3	0 < Ci ≤ 10	10 < Ci ≤ 15	15 < Ci ≤ 30	30 < Ci
	IQCA = 10Ci	IQCA = 20Ci – 100.00	IQCA = 6.67Ci + 100.00	IQCA = 10Ci

En base a la Tabla 6, la función quedó definida de la siguiente manera:

```
# Se define la función con el parámetro 'value':
def ica_co(value):
    if (value>0) & (value<=10):      # Primer rango: de 0 a 10
        return value*10
    elif (value>10) & (value<=15):  # Segundo rango: de 10 a 15
        return (value*20)-100
    elif (value>15) & (value<=30):  # Tercer rango: de 15 a 30
        return (value*6.67)+100
    elif value>30:                   # Cuarto rango: mayor a 30
        return value*10
```

Es así, que la función se llama:

```
# 'lambda' permite aplicar la función línea a línea. Es necesario definir
# el tipo de dato que tendrá el resultado:
ica_co_udf = udf(lambda x: ica_co(x), FloatType())
```

Los resultados obtenidos se observan en la Tabla 7.

Tabla 7 - Resultados del cálculo de IQCA para CO

Timestamp	estacion	co	ica_co
2005-01-06 00:00:00	Belisario	2.78125	27.8125
2005-01-06 00:00:00	Carapungo	2.0719643	20.719643
2005-01-06 00:00:00	Centro	2.6095917	26.095917
2005-01-06 00:00:00	Cotocollao	1.9767857	19.767857
2005-01-06 08:00:00	Belisario	2.9064286	29.064285
2005-01-06 08:00:00	Carapungo	2.1205356	21.205357
2005-01-06 08:00:00	Centro	2.6008418	26.008417
2005-01-06 08:00:00	Cotocollao	1.9709184	19.709185

Después de calcular el IQCA para todos los contaminantes, se realizó un promedio anual (función *resample* en Python), debido a la cantidad de datos que se manejaba. Como paso final en este punto, el índice general se estableció como el valor máximo de los contaminantes. Además, se creó una nueva columna para determinar a qué rango pertenece el índice de acuerdo al valor que tomaron, como lo indican la Tabla 1 y Tabla 2.

2.3.6. Interpolación espacial

Para obtener el mapa de interpolación, fue necesario generar una grilla de latitud y longitud de 100x100, en donde cada elemento de la matriz contendrá un valor z de IQCA o variable climática:

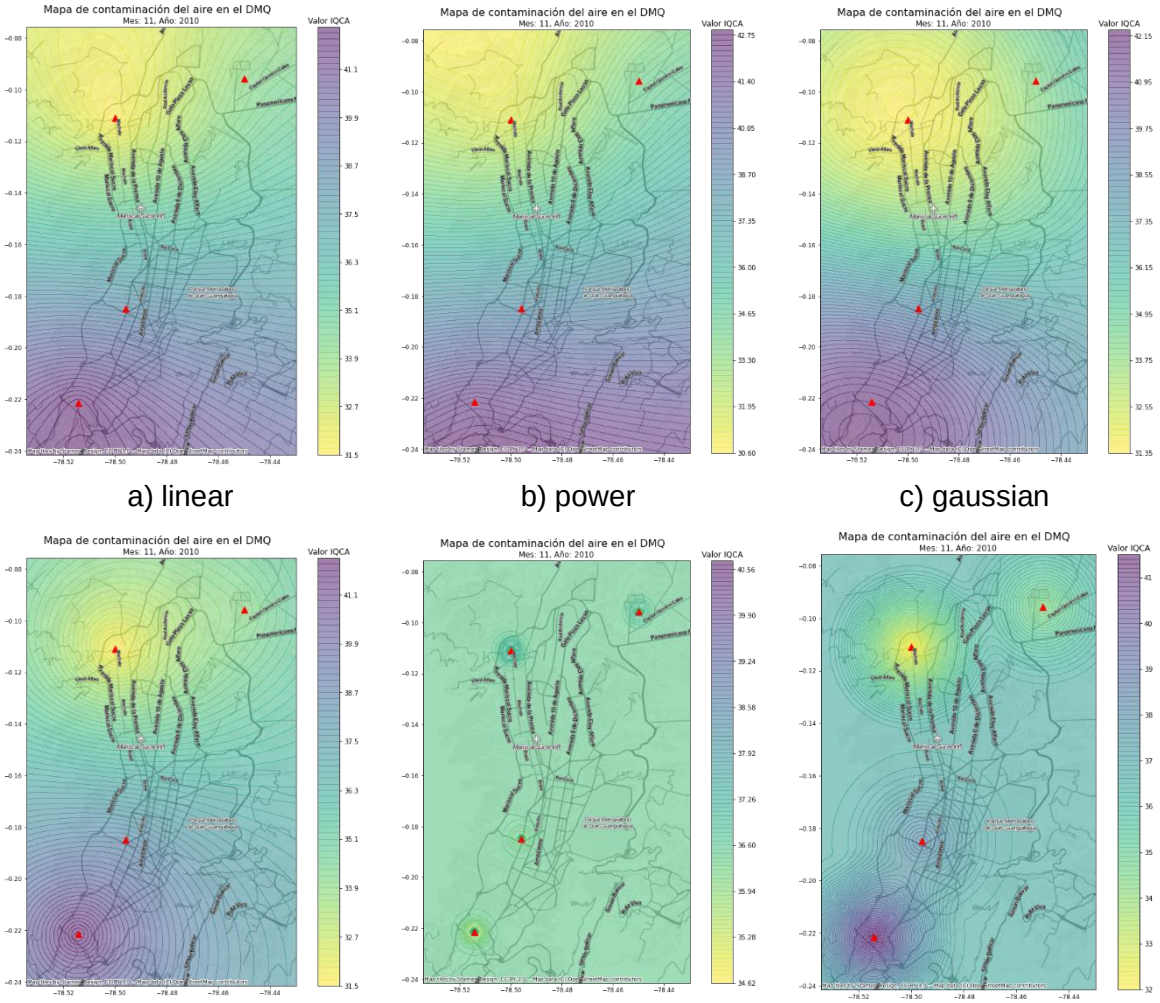
```
# 'linspace' permite crear una grilla, se requiere definir límites
# y tamaño
grid_lon = np.linspace(xmin, xmax, 100)
grid_lat = np.linspace(ymin, ymax, 100)
```

Luego se aplicó el método de kriging ordinario, ya que se desconoce la media [26], con lo que se pudo trasladar a manera de dispersión los valores de z puntuales por toda la grilla:

```
# función que establece los parámetros del modelo Kriging:
OK = OrdinaryKriging(x, y, z,        # define variables x, y, z
      variogram_model='linear',      # define el modelo del variograma
      verbose=False,                 # deshabilita resultado de texto
      enable_plotting=False,         # deshabilita trazado del variograma
      coordinates_type='geographic')  # define tipo de coord. geográficas
```

```
# función que ejecuta el modelo sobre la grilla creada anteriormente:
z1, ss1 = OK.execute('grid', grid_lon, grid_lat)
```

Cabe aclarar que, según la variable, se alternaron los modelos de variograma, con el fin de encontrar la mejor visualización de resultados. Para ello, se generaron mapas por cada variograma y se compararon para elegir el más adecuado. La Figura 16 muestra estas diferencias para el IQCA en el mes 11 del año 2010, en donde los modelos más adecuados son a), b) y c).



a) linear b) power c) gaussian

d) exponential e) hole-effect f) spherical

Figura 16 - Comparativa de modelos de variograma.

La Tabla 8 muestra un resumen de los modelos de variograma empleados por año de análisis y variable.

Tabla 8 - Modelos de variograma empleados por variable y por año

Año	IQCA	Precipitación	Presión	Dirección del viento	Humedad
2005	gaussian	linear	linear	exponential	linear
2006	linear	linear	linear	spherical	spherical
2007	gaussian	linear	linear	linear	exponential
2008	linear	linear	spherical	linear	linear
2009	spherical	linear	gaussian	linear	hole-effect
2010	spherical	linear	gaussian	linear	gaussian
2011	spherical	linear	gaussian	linear	linear
2012	spherical	linear	spherical	linear	linear
2013	spherical	linear	spherical	linear	linear
2014	power	linear	spherical	spherical	linear
2015	linear	linear	spherical	exponential	linear
2016	linear	linear	gaussian	gaussian	linear
2017	spherical	linear	exponential	linear	linear
2018	linear	linear	spherical	linear	linear
2019	linear	linear	spherical	linear	linear
2020	spherical	linear	spherical	linear	linear

2.3.7. Visualización geográfica

Gracias a las librerías MatPlotLib y Contextily, se consiguió representar las variables en estudio con su componente geográfico:

```
# 'meshgrid' devuelve una matriz de coordenadas a partir de vectores de
coordenadas:
xintrp, yintrp = np.meshgrid(grid_lon, grid_lat)
fig, ax = plt.subplots(figsize=(12,12)) # se configura el tamaño del plot

# 'contour' traza curvas de nivel, recibe como parámetros la matriz de
# coordenadas, z (IQCA o variable climática), la paleta de colores y la
# transparencia de la línea
contour = plt.contourf(xintrp,
      yintrp, z1, len(z1),
      cmap='RdYlBu_r', alpha=0.5)

# 'add_basemap' añade un mapa base o de fondo, se debe definir el sistema
# de coordenadas que se ajuste con el usado en los datos que se está
# manejando y la fuente del mapa base
cx.add_basemap(ax,
      crs='EPSG:4326',
      source=cx.providers.Stamen.TonerLite)
```

```
# se agrega un basemap adicional que contiene etiquetas o labels de
# lugares principales para un mejor reconocimiento de la zona
cx.add_basemap(ax,
      crs='EPSG:4326',
      source=cx.providers.Stamen.TonerLabels)
```

2.4. Evaluación

Para evaluar el rendimiento de la arquitectura propuesta, se ejecutó el mismo algoritmo de interpolación espacial empleado en el presente trabajo (Kriging) dentro de un Sistema de Información Geográfica (SIG), el cual tiene automatizado este proceso, utilizando las mismas configuraciones que en el sistema diseñado. El software elegido fue QGIS versión 3.22.7. Esta comparativa se puede revisar en la sección de discusión del siguiente capítulo.

2.5. Reflexión

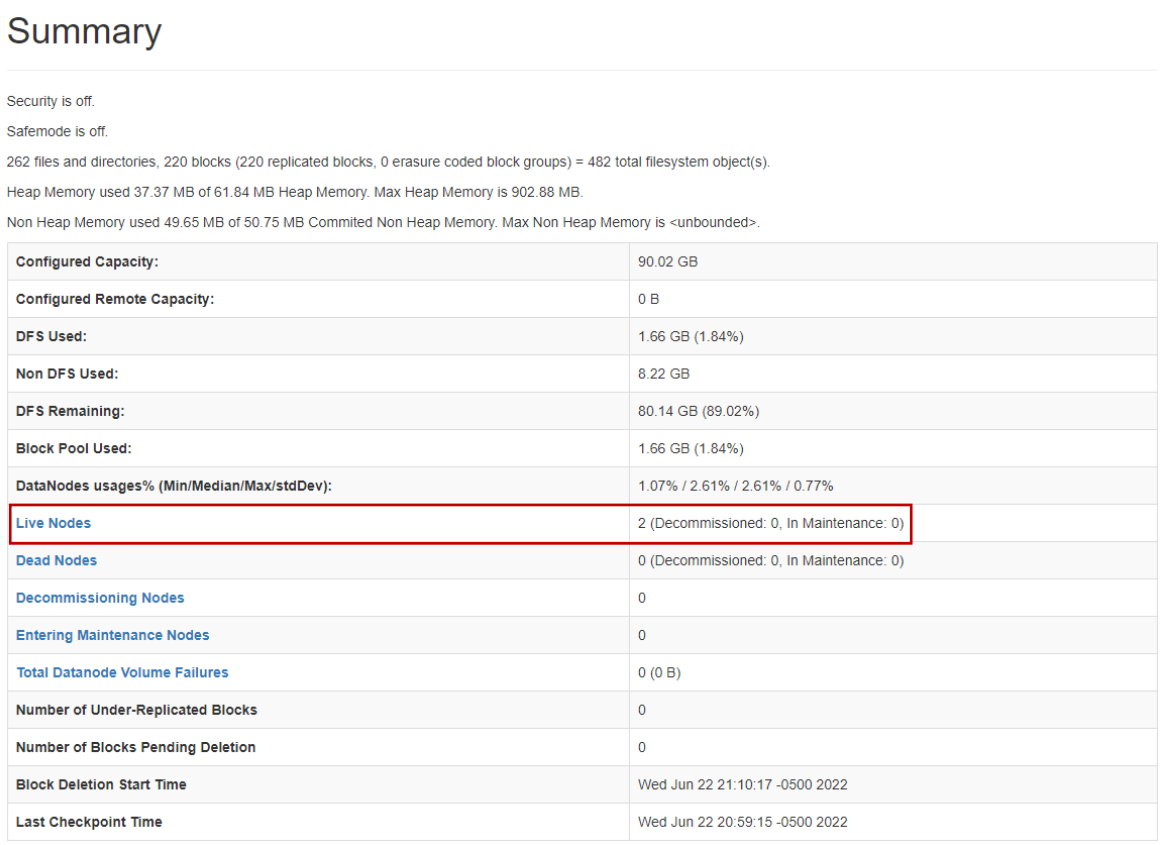
Los resultados obtenidos en la presente investigación permitieron evaluar geográficamente la situación histórica del DMQ respecto a la problemática de contaminación del aire, demostrando que este tipo de información contribuye a una mejor toma de decisiones a nivel gubernamental.

3. RESULTADOS Y DISCUSIÓN

3.1. Resultados

3.1.1. Clúster implementado

A través de la interfaz web del HDFS, se puede comprobar el estado general del clúster y con ello verificar su buen funcionamiento. La Figura 17 muestra un resumen del mismo, en el que hay que destacar que se encuentran activos dos nodos, los que corresponden a los esclavos o workers.



Summary

Security is off.

Safemode is off.

262 files and directories, 220 blocks (220 replicated blocks, 0 erasure coded block groups) = 482 total filesystem object(s).

Heap Memory used 37.37 MB of 61.84 MB Heap Memory. Max Heap Memory is 902.88 MB.

Non Heap Memory used 49.65 MB of 50.75 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	90.02 GB
Configured Remote Capacity:	0 B
DFS Used:	1.66 GB (1.84%)
Non DFS Used:	8.22 GB
DFS Remaining:	80.14 GB (89.02%)
Block Pool Used:	1.66 GB (1.84%)
DataNodes usages% (Min/Median/Max/stdDev):	1.07% / 2.61% / 2.61% / 0.77%
Live Nodes	2 (Decommissioned: 0, In Maintenance: 0)
Dead Nodes	0 (Decommissioned: 0, In Maintenance: 0)
Decommissioning Nodes	0
Entering Maintenance Nodes	0
Total Datanode Volume Failures	0 (0 B)
Number of Under-Replicated Blocks	0
Number of Blocks Pending Deletion	0
Block Deletion Start Time	Wed Jun 22 21:10:17 -0500 2022
Last Checkpoint Time	Wed Jun 22 20:59:15 -0500 2022

Figura 17 - Resumen del estado general del clúster (HDFS).

La Figura 18 amplía la información sobre la salud de los nodos, observando que tienen un estado activo.

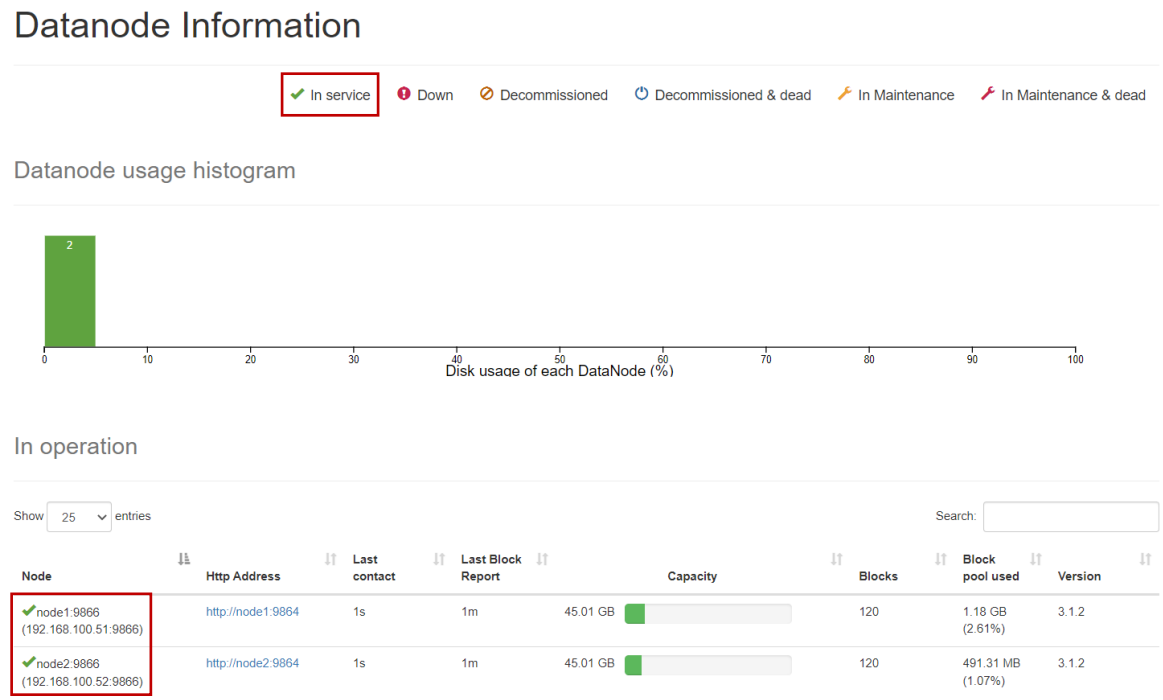


Figura 18 - Estado de los workers (2).

De la misma forma, Hadoop a través de su interfaz web muestra el estado del clúster (Figura 19), además de que permite el monitoreo de los trabajos o tareas que se le envían a procesar.

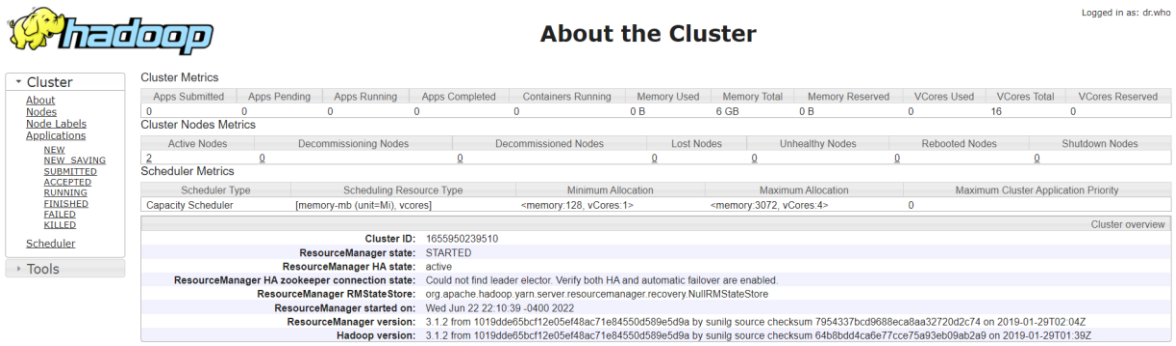


Figura 19 - Resumen del estado general del clúster (Hadoop).

Una vez levantado HDFS y Yarn, se puede verificar los procesos que se encuentran en ejecución dentro del nodo máster (Figura 20) mediante el comando **jps** (Java Virtual Machine Process Status Tool), observando que el NameNode o nodo principal se encuentra activo, junto a Yarn (ResourceManager).

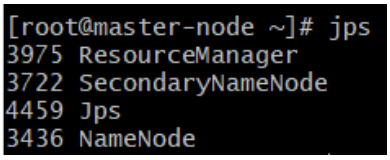


Figura 20 - Procesos en ejecución después de levantar HDFS y Yarn.

Después de levantar Zeppelin, se verifica nuevamente los procesos activos con jps, donde se percibe que el mismo está en ejecución (Figura 21).

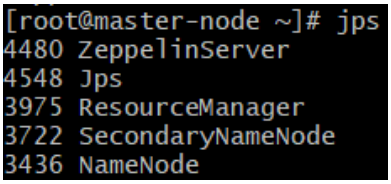


Figura 21 - Procesos en ejecución después de levantar Zeppelin.

Finalmente, se accede a la interfaz web de Zeppelin. El botón verde en la esquina superior derecha indica el estado activo, como lo muestra la Figura 22.

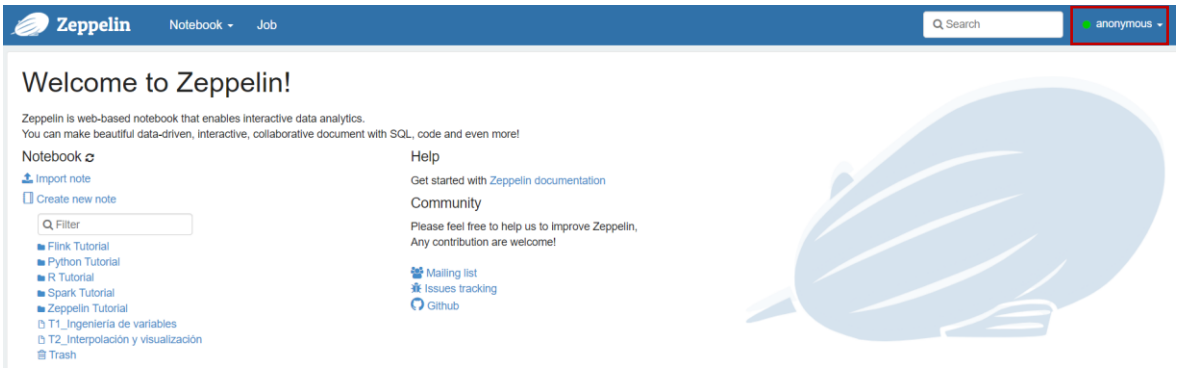


Figura 22 - Página de inicio de Zeppelin.

3.1.2. Mapas geográficos

Los mapas obtenidos se componen de tres partes:

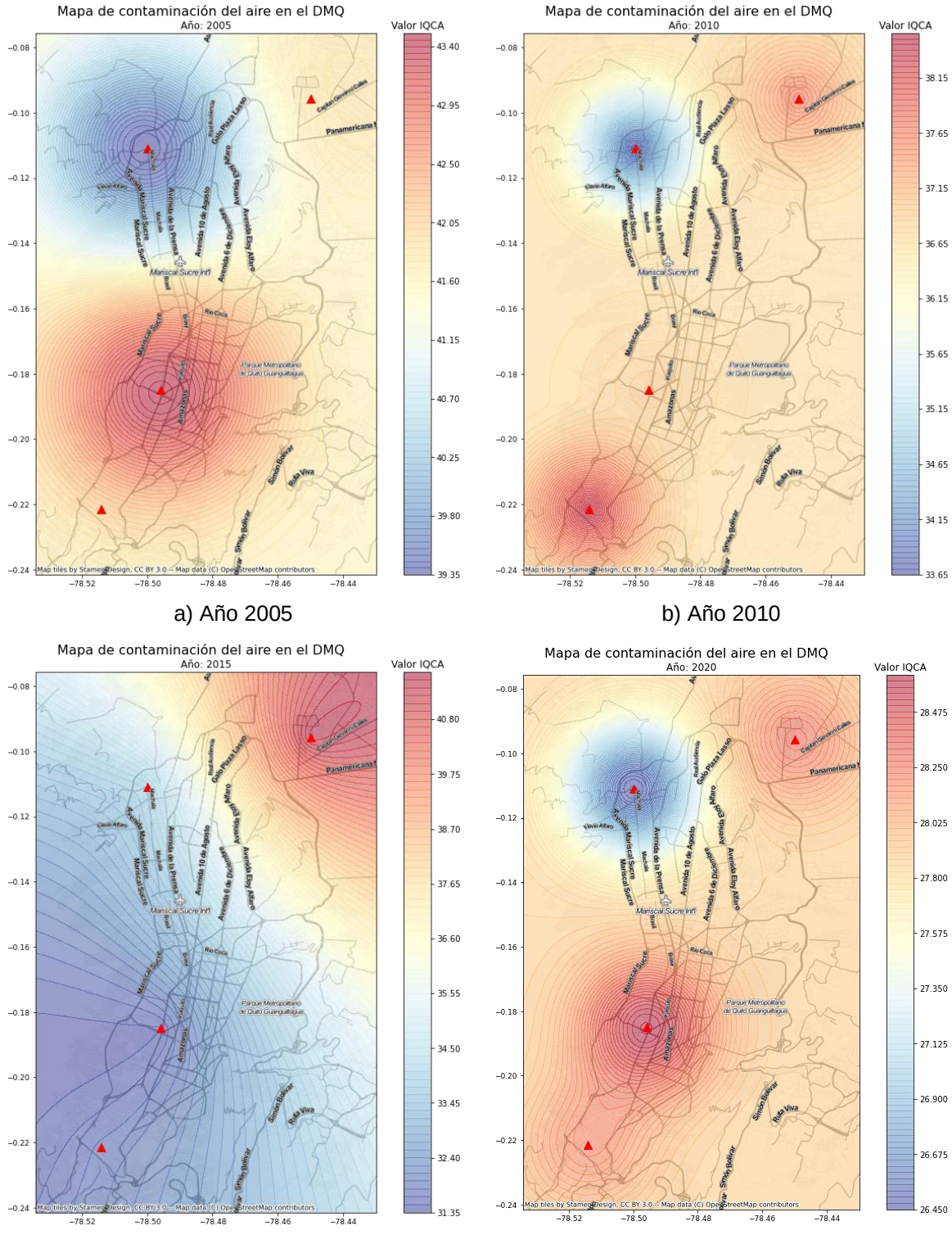
- Encabezado: incluye el título que describe el propósito del mapa, y el subtítulo con el año al que corresponden los datos observados.
- El cuerpo del mapa: es la parte fundamental donde se representa, en este caso, el IQCA o variable climática evaluada de manera espacial, junto con unas líneas sutiles que muestran el cambio de valores, asemejándose a curvas de nivel. Incluye un mapa base, es decir algunas vías principales de la ciudad y labels que permiten orientar rápidamente al usuario; y el símbolo ▲, que hace referencia a las estaciones de monitoreo utilizadas en el estudio.

- Leyenda: se encuentra a la derecha del cuerpo del mapa y corresponde a una escala secuencial de colores que explica los valores que puede tomar el índice o cualquier otra variable. Mientras más alto se ve en la escala, tendrá un valor más alto y viceversa.

Estos mapas se extrajeron por año (desde 2005 hasta 2020) y por cada variable (dirección del viento, humedad, precipitación y presión barométrica). Al ser resultados extensos, se muestran cuatro por variable (cada cinco años) para realizar comparativas.

Cabe mencionar que para el caso de los valores IQCA, las categorías encontradas en todos los años evaluados fueron 1 y 2, que corresponden a *nivel deseable u óptimo* y *nivel aceptable o bueno*, respectivamente; lo que significa que el índice se encontró entre 0 y 100.

En la Figura 23, se muestra la evolución del índice de la calidad del aire (IQCA). Se observa que existe una mayor concentración en la estación Centro; el resto de estaciones muestran valores medios y hacia la estación Cotocollao el valor se ve disminuido. En el año 2010 existió una variante: la calidad del aire empeoró hacia el noreste de la ciudad y mejoró en el Centro, siendo la estación Carapungo la que presentó mayor valor.



a) Año 2005

b) Año 2010

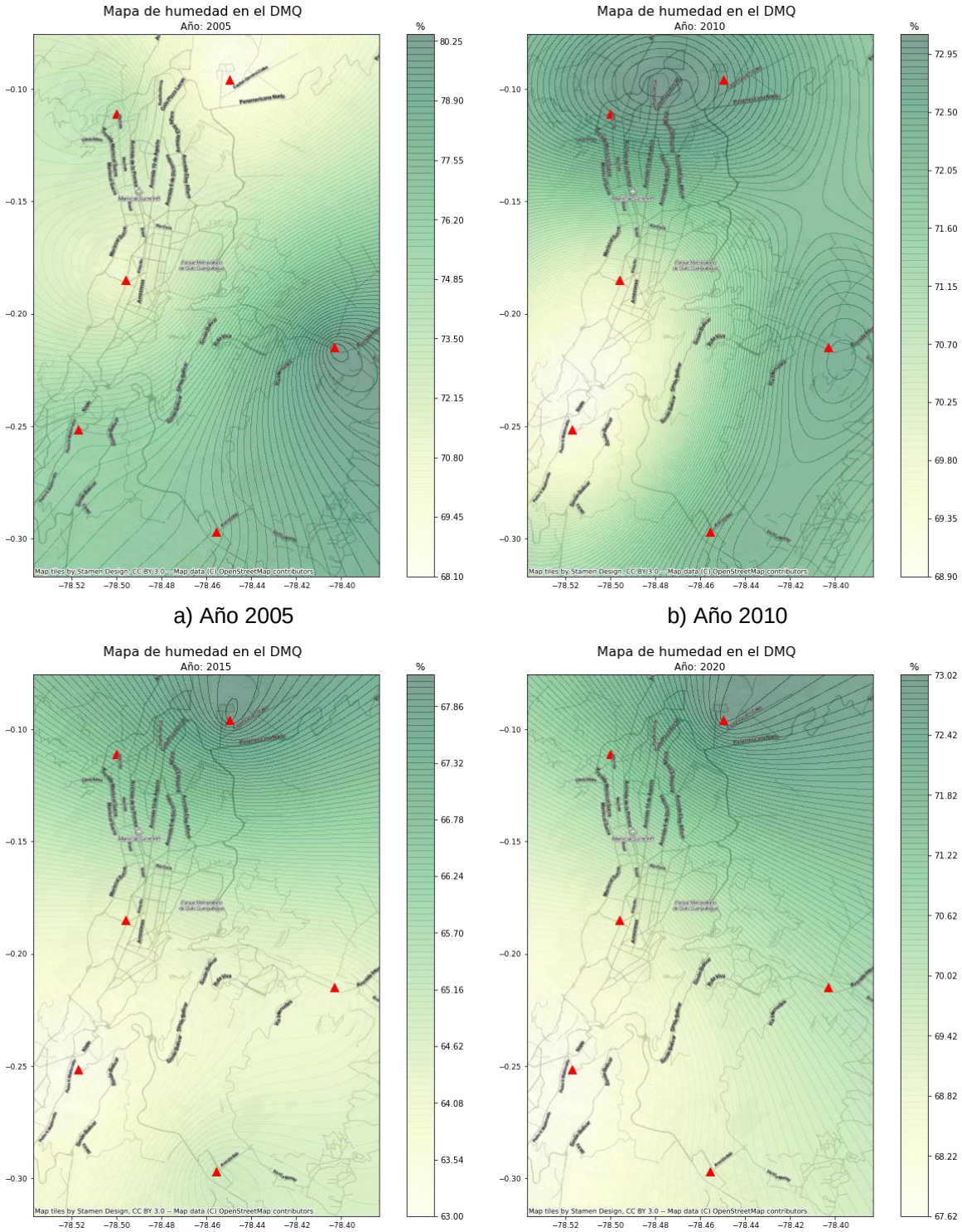
c) Año 2015

d) Año 2020

Figura 23 - Evolución del índice IQCA en el DMQ.

La variable de humedad y su evolución se puede observar en la Figura 24. Se aprecia que desde el 2010, aproximadamente, existe un patrón de comportamiento, puesto que la

humedad tiene valores más altos hacia el norte; a diferencia del año 2005, que la humedad presentó valores más altos hacia el sur.



a) Año 2005

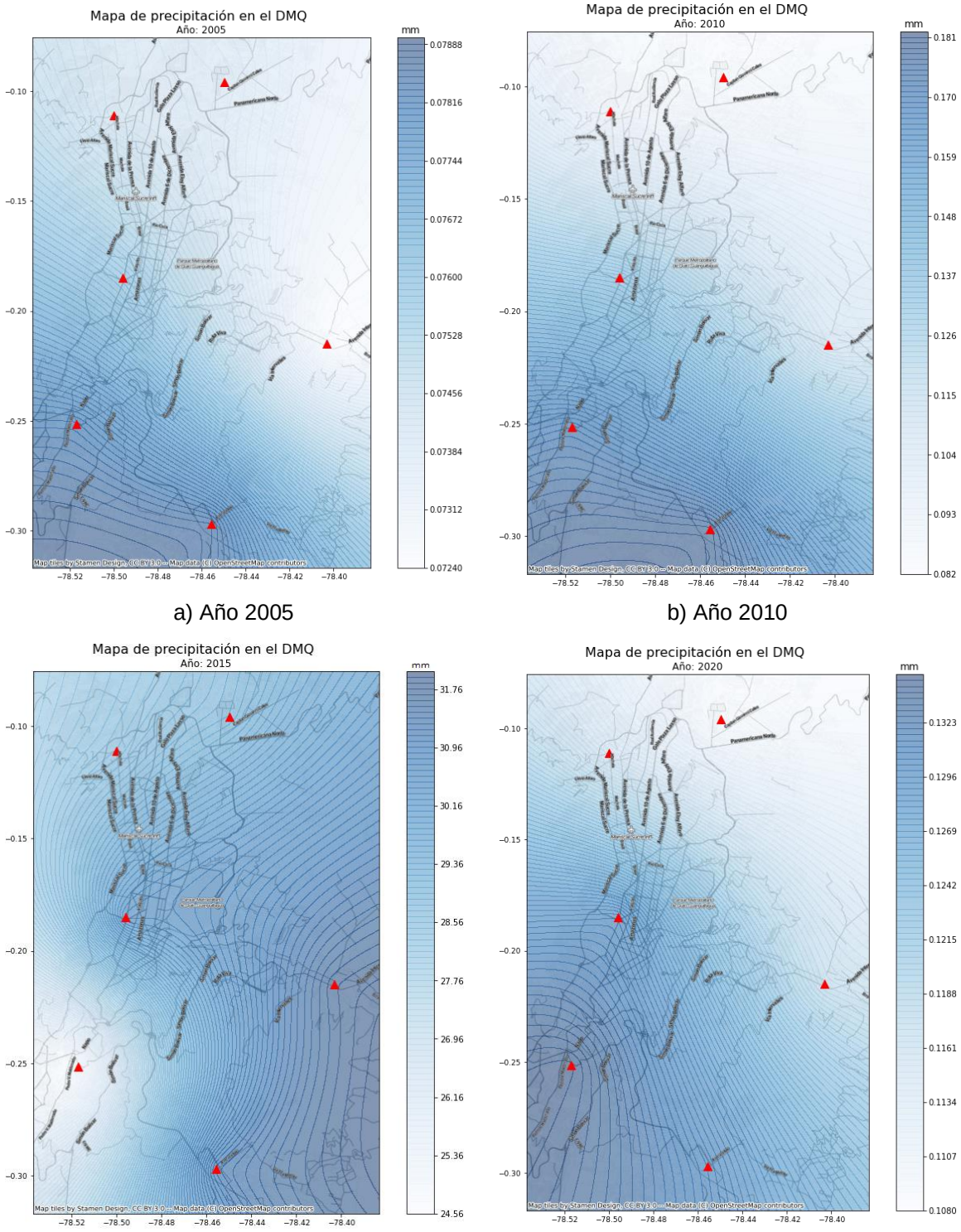
b) Año 2010

c) Año 2015

d) Año 2020

Figura 24 - Evolución de la humedad en el DMQ.

En la Figura 25 se aprecia la variación de la precipitación en el DMQ. Al igual que el IQCA, se observa una tendencia en 2005, 2010 y 2020, donde las lluvias son mayores hacia el centro y sur de la ciudad; mientras que, al parecer, el año 2015 fue un año atípico con promedios altos de lluvias en la mayoría de la ciudad.



a) Año 2005

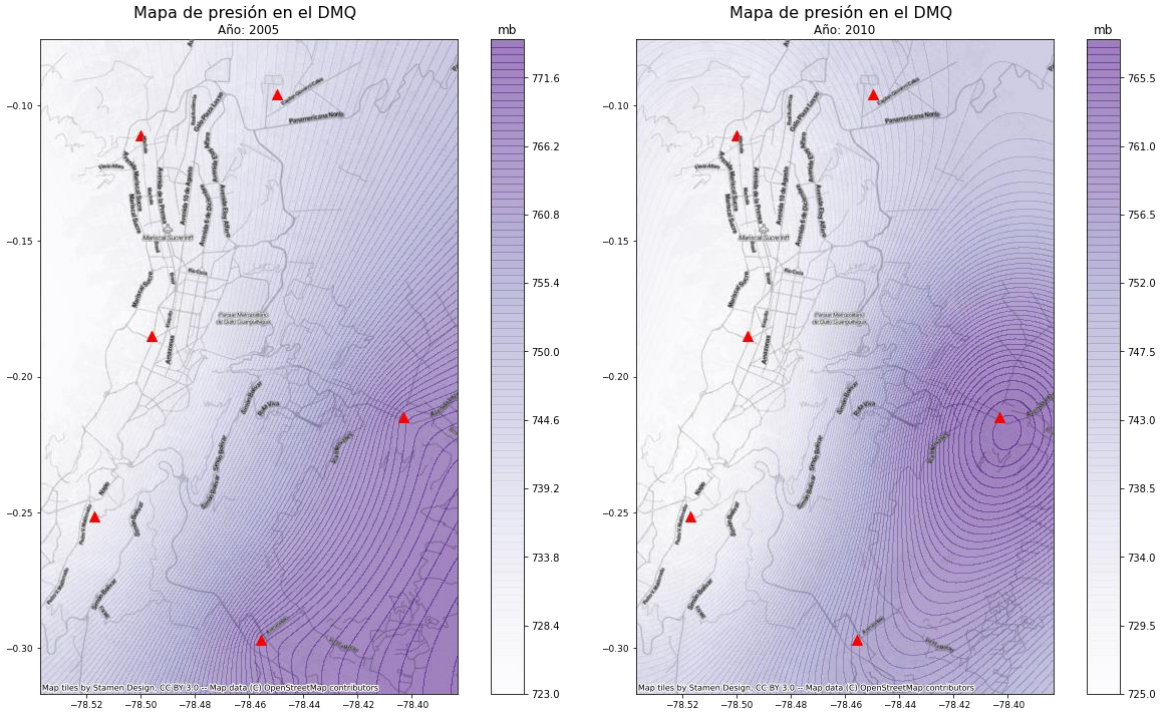
b) Año 2010

c) Año 2015

d) Año 2020

Figura 25 - Evolución de la precipitación en el DMQ.

La Figura 26 muestra los cambios que han existido en cuanto a la presión barométrica.



a) Año 2005

b) Año 2010

c) Año 2015

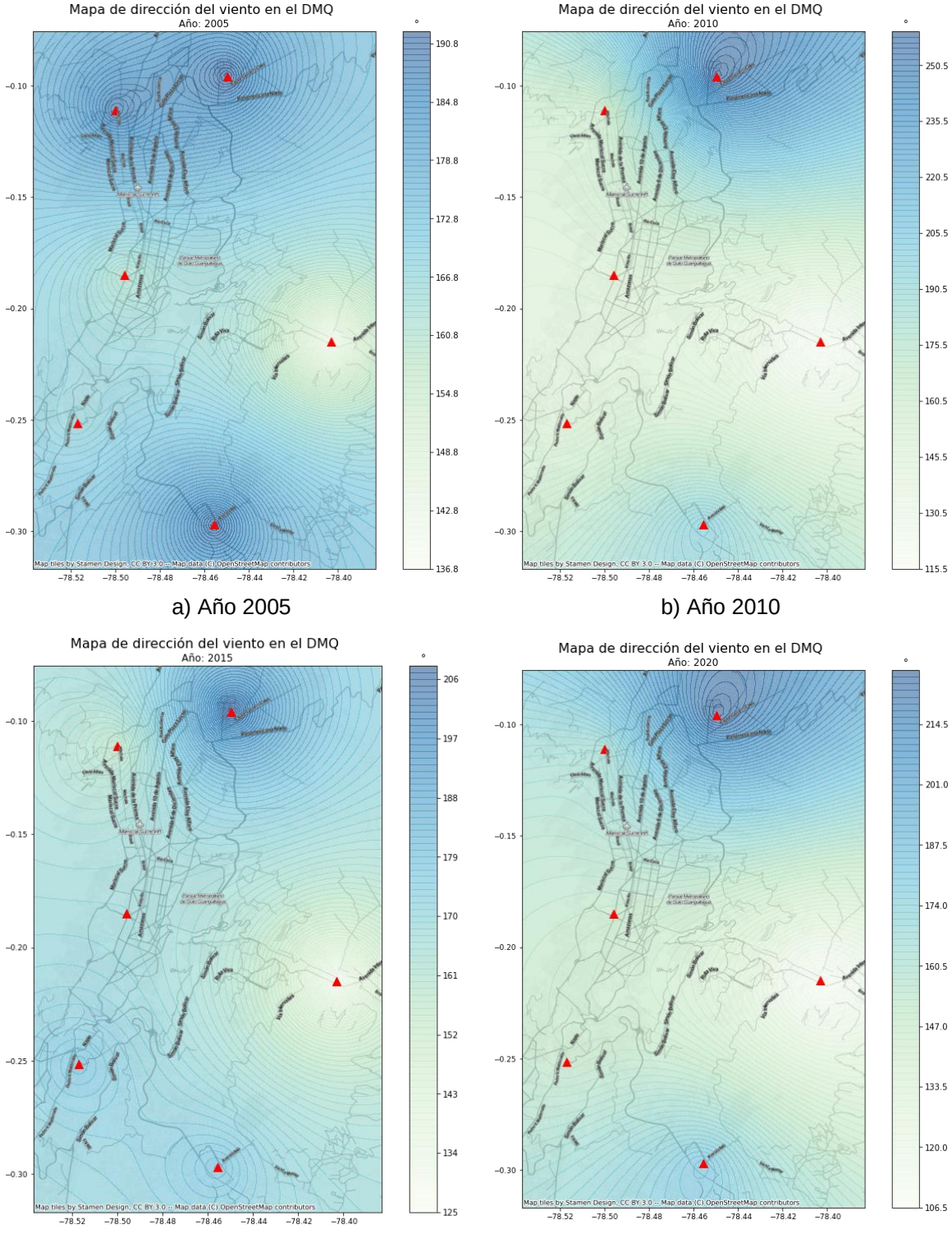
d) Año 2020

Figura 26 - Evolución de presión barométrica en el DMQ.

Entre 2005 y 2010 se observa un patrón: valores más altos hacia el sureste de la ciudad, representados por las estaciones de los valles (Tumbaco y Los Chillos). En los años 2015

y 2020, los mapas muestran un descenso en casi toda el área de estudio, excepto en Tumbaco que presenta un valor más alto.

En cuanto a la dirección del viento, sus cambios se aprecian en la Figura 27.



a) Año 2005

b) Año 2010

c) Año 2015

d) Año 2020

Figura 27 - Evolución de la dirección del viento en el DMQ.

Como se mencionó anteriormente, la dirección del viento se mide en grados. Para comprender los mapas presentados en esta variable, es necesario considerar la Figura 28, la cual muestra gráficamente la dirección en relación a los puntos cardinales.

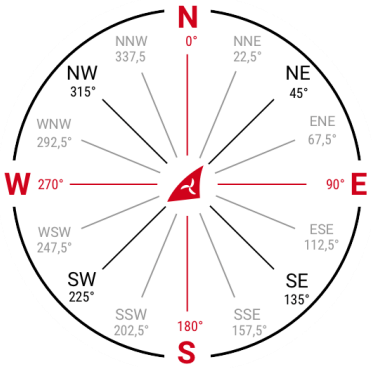


Figura 28 - Dirección del viento en grados [41]

Para el año 2005, la dirección del viento en su mayoría se encuentra hacia el sur; en el 2010, esta dirección cambia hacia el sureste, exceptuando la zona de Carapungo que tiene vientos hacia el suroeste. Los años 2015 y 2020 presentan un ligero patrón de vientos hacia el sureste en la zona de Carapungo, Los Chillos y El Camal; mientras que el resto de la ciudad presenta vientos con dirección sureste.

3.2. Discusión

Para cuantificar recursos naturales o mostrar problemas sociales, en este caso, de contaminación, los mapas geográficos son documentos imprescindibles [42], mismos que resultan ser un complemento ideal al momento de analizar fenómenos de este tipo. Éstos exponen la información visualmente, siendo una ventaja, ya que las imágenes pueden comunicar de una manera más efectiva. Esta razón es la que impulsó el desarrollo de la presente investigación: demostrar que la visualización geográfica vinculada a una arquitectura de procesamiento de datos masivos puede mejorar la toma estratégica de decisiones a nivel gubernamental, puesto que hay un mejor sustento (visual) para elaborar políticas que beneficien a la población en general; a diferencia de observar estas estadísticas de manera textual o con gráficos planos. La aplicación de esta arquitectura es una innovación en este ámbito en el país, lo cual viene a ser necesario por la gran cantidad de datos que se recolectan a través de sensores en campo. Esta nueva forma de procesamiento trae consigo beneficios que se reflejan en los resultados obtenidos: reducción del tiempo de ejecución, herramientas open-source implementadas (sin inversión en software) y la utilidad per sé de los mapas.

Por otro lado, se evaluaron los resultados obtenidos en el índice IQCA. Como se mencionó, los niveles de contaminación del aire encontrados fueron deseables (óptimos y buenos), lo que incurre a inferir que el aire en la ciudad de Quito es bueno. Sin embargo; a través de medios informativos se encontraron noticias del año 2018 como la siguiente:

> *En los días con mayor contaminación, el aire de Quito superó los 100 puntos, lo que se considera un nivel de 'precaución'* [5].

También se mostró una evolución semanal de este índice, visto en la Figura 29.

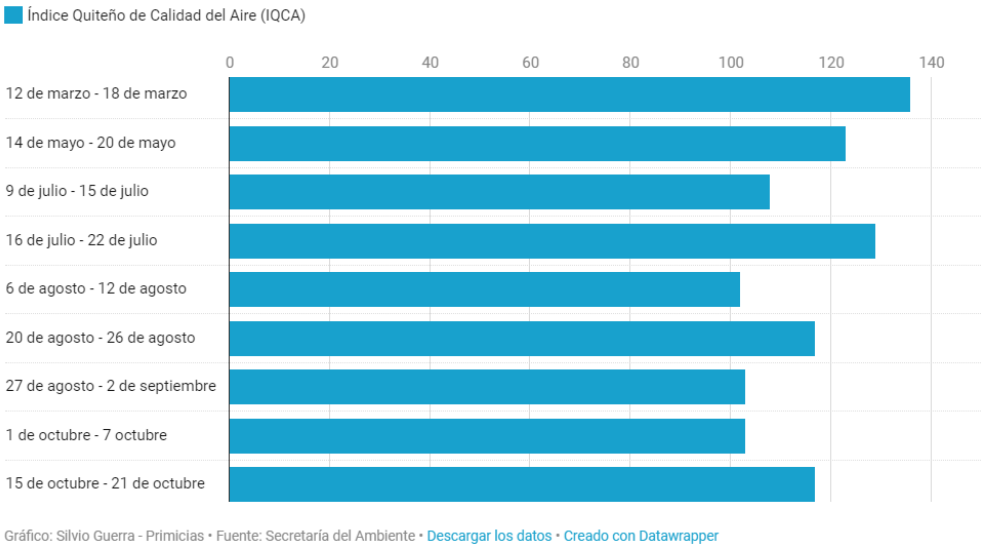
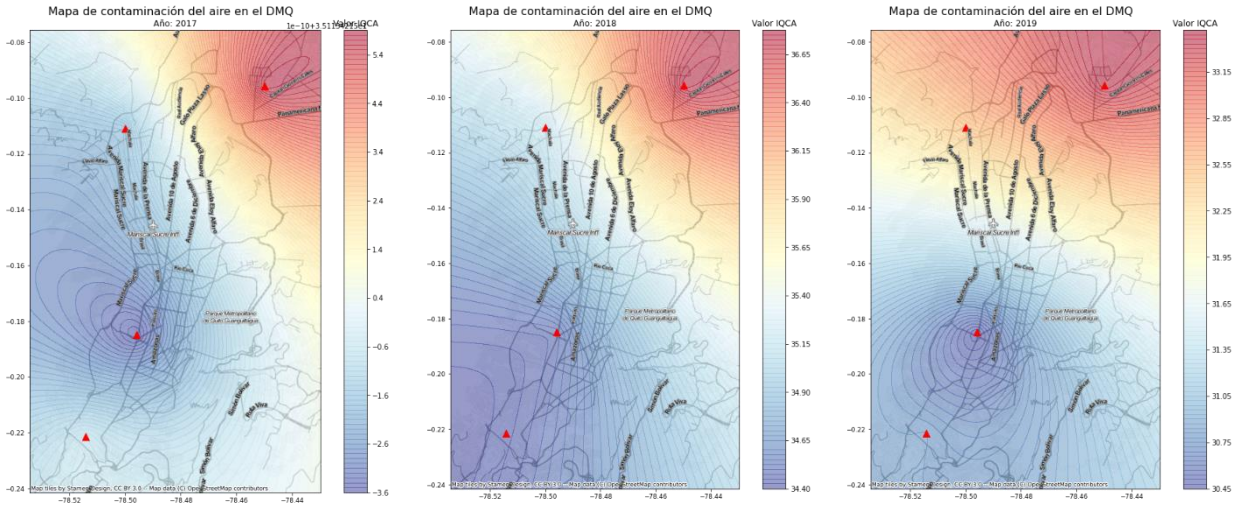


Figura 29 - Evolución de la contaminación del aire en Quito en 2018 [5]**.**

En efecto, se aprecian unas cuantas semanas del año en el que el índice superó los 100 puntos, lo cual podría ser contrario a lo mostrado en la presente investigación. No obstante, hay que considerar que la información mostrada a nivel espacial tuvo que ser reducida a un promedio anual, debido a la cantidad de datos iniciales; y ya que fueron pocas semanas con esta variación, el resultado final no se ve afectado. En la Figura 30, se puede apreciar la evolución del índice un año antes y después del analizado, como evidencia de que se no existen alteraciones significativas y se percibe casi el mismo patrón. En todo caso, a nivel de Municipio de Quito, se podría optar por análisis con períodos de tiempo más cortos, ya que la aplicación que la entidad puede dar al sistema es el monitoreo de valores incrementados y la búsqueda de posibles causas.



a) IQCA año 2017 b) IQCA año 2018 c) IQCA año 2019

Figura 30 - Comparativa IQCA.

En relación a las variables en estudio, no se puede hablar de un mismo patrón a lo largo de los años y esto se debe a los diversos factores que alteran a las mismas. Por ejemplo, el índice de calidad del aire se ve fuertemente afectado por eventos naturales como erupciones volcánicas [43], que hacen que el valor se dispare; o efectos del cambio climático que hacen que llueva más o existan sequías en referencia a años anteriores [44].

Por otra parte, la eficiencia de la arquitectura planteada fue medida a través de una comparativa de tiempos de respuesta con y sin el uso de la misma. La interpolación espacial (Kriging) dentro de QGIS tomó un tiempo de 4.52 segundos (Figura 31); mientras que el sistema propuesto mostró un tiempo de ejecución de 2.0 segundos (Figura 32), obteniendo una reducción del 56%, mejorando así el camino tradicional de interpolación y visualización de mapas y generando una nueva alternativa open-source con optimización de recursos y tiempo.

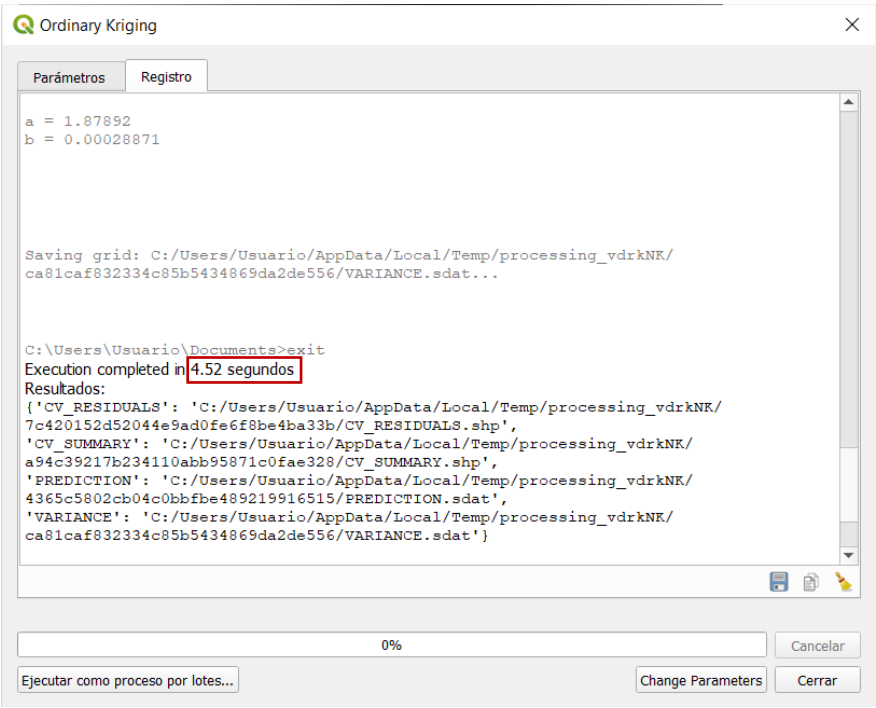


Figura 31 - Tiempo de ejecución de kriging en QGIS.

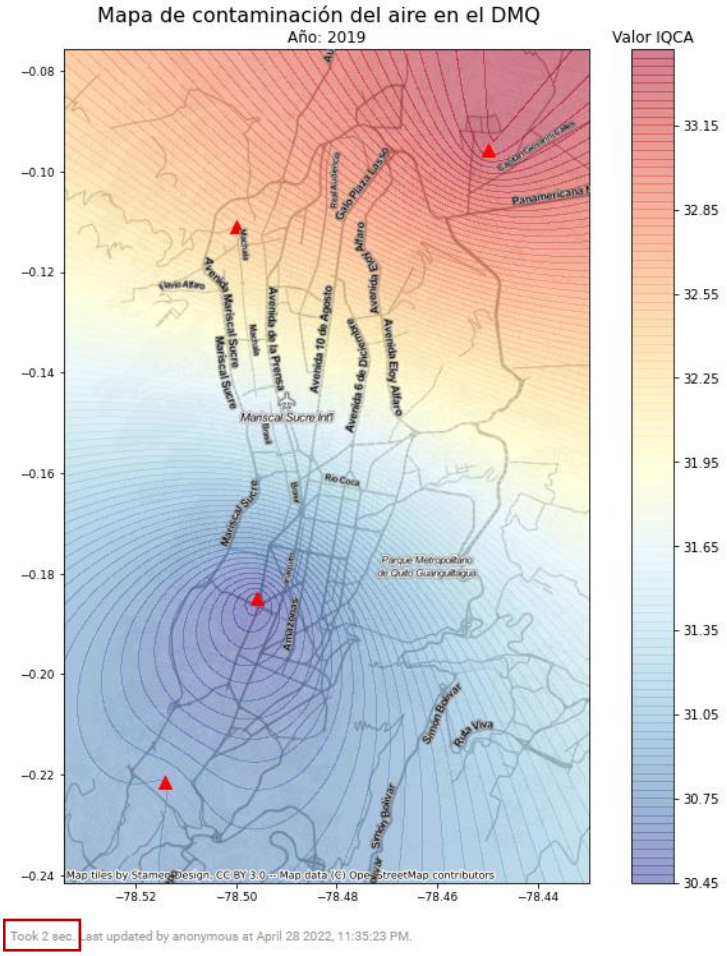


Figura 32 - Tiempo de ejecución de kriging con arquitectura propuesta.

4. CONCLUSIONES Y RECOMENDACIONES

4.1. Conclusiones

- De acuerdo a la literatura revisada, se encontró que la mayoría de investigaciones relacionadas al Big Data e información geográfica se han centrado en evaluar el rendimiento de los algoritmos de predicción y la optimización de tiempo, pero muy pocos han expuesto la importancia de visualizar los datos de manera geográfica, a diferencia del presente estudio que se ha centrado en el manejo y visualización de grandes cantidades de datos geográficos como deberían mostrarse: con mapas; además de demostrar que existe un ahorro de tiempo y recursos en su implementación.
- El artículo en [29] es el que aportó con la mayor cantidad de ideas para esta investigación, ya que expuso la factibilidad de integrar la arquitectura de Big Data con interpolación espacial de datos mostrada a través de mapas.
- Se depuraron, revisaron y evaluaron datos de la página de la Secretaría de Ambiente del Municipio de Quito del Ecuador. Fueron en total 13 archivos con aproximadamente 2 millones de registros desde el año 2004.
- Se calculó el Índice Quiteño de la Calidad del Aire (IQCA) en base a los contaminantes CO, O_3, NO_2, SO_2 y $PM_{2.5}$, el cual permitió medir dicha calidad en todo el Distrito Metropolitano.
- La arquitectura diseñada permite el procesamiento de grandes volúmenes de datos que incluyen el componente geográfico, la cual se basó en un clúster de tres nodos que utiliza el principio de procesamiento paralelo de datos. Es open-source, puesto que utiliza Apache Spark como motor de procesamiento, administrado por Yarn y enlazado a Zeppelin para mostrar los resultados de manera espacial a través de mapas geográficos.
- El IQCA representa, en una escala del 0 al 500, el nivel de contaminación de aire; para el caso de la ciudad de Quito, este índice no sobrepasó en promedio el valor de 100, categorizando a la calidad del aire dentro de los niveles óptimos y buenos. Sin embargo, existen épocas del año, háblese de días o semanas, que este índice puede aumentar debido a exceso de tráfico vehicular, descargas industriales o eventos naturales. Esto fue encontrado en fuentes externas y no pudo ser validado mediante este estudio, debido a que los datos se agregaron anualmente para presentar los mapas.

- A partir de este caso de estudio, se determinó que la arquitectura desarrollada puede emplearse para modelar cualquier variable medioambiental de la que se disponga de una gran cantidad de datos y se requiera visualizar a manera de mapas, permitiendo vincular la analítica de datos con representación geográfica como innovación y contribuyendo así a la toma de decisiones eficientes y acertadas en temas de políticas públicas, en este caso.

4.2. Recomendaciones

- Basados en la arquitectura propuesta, los estudios posteriores que se pueden ejecutar serían algoritmos de predicción de contaminación del aire o cualquier modelamiento de variables medioambientales, cuyos resultados se presenten en mapas.
- A esta arquitectura se le podría añadir una conexión con Apache Hbase o Hive como gestores de bases de datos, los cuales brindarían una mejor organización en caso de que se requiera trabajar con mayor cantidad de datos e incluir predicciones.
- Esta arquitectura recoge datos estáticos, pero se podría incluir el procesamiento de datos en streaming al implementar Apache Kafka o Flink, además de vincular una interfaz dinámica para mostrar un fenómeno en tiempo real.

REFERENCIAS BIBLIOGRÁFICAS

[1] A. Rodríguez and N. Cuvi, “Contaminación del Aire y Justicia Ambiental en Quito, Ecuador,” *Front. J. Soc. Technol. Environ. Sci.*, vol. 8, no. 3, pp. 13–46, 2019, doi: 10.21664/2238-8869.2019v8i3.p13-46.

[2] Organización Mundial de la Salud (OMS), “Nueve de cada diez personas de todo el mundo respiran aire contaminado,” 2018. https://www.who.int/es/news/item/02-05-2018-9-out-of-10-people-worldwide-breathe-polluted-air-but-more-countries-are-taking-action (accessed Jul. 21, 2021).

[3] Organización Mundial de la Salud (OMS), “Calidad del aire ambiente (exterior) y salud,” 2018. https://www.who.int/es/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health (accessed Jul. 21, 2021).

[4] Organización Mundial de la Salud (OMS), “Ambient air pollution: A global assessment of exposure and burden of disease,” 2016. Accessed: Jul. 21, 2021. [Online]. Available: https://apps.who.int/iris/bitstream/handle/10665/250141/9789241511353-eng.pdf?sequence=1&isAllowed=y.

[5] Diario Primicias, “El aire de Quito supera los límites permitidos de contaminación,” 2019.

[6] Municipio del Distrito Metropolitano de Quito, “Plan Metropolitano de Desarrollo y Ordenamiento Territorial,” Quito, 2015. Accessed: Mar. 24, 2022. [Online]. Available: http://sitp.pichincha.gob.ec/repositorio/diseno_paginas/archivos/PDOT cantonal del Distrito Metropolitano de Quito 2015.pdf.

[7] H. Isabel and J. i Caralt, “Uso de analítica para dar soporte a la toma de decisiones docentes,” in *Actas de las XX JENUI*, 2014, vol. 9, no. 11, pp. 83–90, Accessed: Jul. 11, 2021. [Online]. Available: https://core.ac.uk/download/pdf/41791824.pdf.

[8] O. Barzaga, H. Vélez, J. Nevárez, and M. Arroyo, “Gestión de la información y toma de decisiones en organizaciones educativas,” *Rev. Ciencias Soc.*, vol. XXV, no. 2, pp. 120–130, 2019, Accessed: Jul. 13, 2021. [Online]. Available: https://dialnet.unirioja.es/servlet/articulo?codigo=7025997&info=resumen&idioma=ENG.

[9] Y. Rodríguez and M. Pinto, “Modelo de uso de información para la toma de decisiones estratégicas en organizaciones de información,” *Transinformacao*, vol. 30, no. 1, Pontificia Universidade Catolica de Campinas, pp. 51–64, Jan. 01, 2018.

[10] P. Pesado *et al.*, “Ingeniería de Software para Sistemas Distribuidos. Experiencias

en la Gestión, Desarrollo e Implantación de Sistemas," in *XVI Workshop de Investigadores en Ciencias de la Computación*, 2014, pp. 611–616, Accessed: Apr. 29, 2022. [Online]. Available: https://digital.cic.gba.gob.ar/bitstream/handle/11746/2096/11746_2096.pdf-PDFA.pdf?sequence=1&isAllowed=y.

[11] IBM, "What is distributed computing," 2021. https://www.ibm.com/docs/en/txseries/8.2?topic=overview-what-is-distributed-computing (accessed May 02, 2022).

[12] D. C. Schmidt, D. L. Levine, and C. Cleeland, "Architectures and Patterns for Developing High-performance, Real-time ORB Endsystems," *Adv. Comput.*, vol. 48, no. C, pp. 1–118, 1999, doi: 10.1016/S0065-2458(08)60018-2.

[13] R. Mohanán, "What Are Distributed Systems? Architecture Types, Key Components, and Examples," 2022. https://www.toolbox.com/tech/cloud/articles/what-is-distributed-computing/ (accessed May 02, 2022).

[14] Apache Spark, "Spark Overview." https://spark.apache.org/docs/latest/index.html (accessed May 04, 2022).

[15] A. Ghaffar and T. Rahim, "Big Data Analysis: Ap Spark Perspective," *Glob. J. Comput. Sci. Technol.*, vol. XV, no. 1, 2015, Accessed: May 04, 2022. [Online]. Available: https://computerresearch.org/index.php/computer/article/view/1137/1124.

[16] B. Buitrago, "¿Qué hay detrás del procesamiento de Apache Spark?," *iWannaBeDataDriven*, 2020. https://medium.com/iwannabedatadriven/qué-hay-detrás-del-procesamiento-de-apache-spark-ii-474402939ca5 (accessed May 05, 2022).

[17] Apache Spark, "Cluster Mode Overview." https://spark.apache.org/docs/latest/cluster-overview.html (accessed May 04, 2022).

[18] IBM, "What is geospatial data?," 2020. https://www.ibm.com/topics/geospatial-data (accessed May 02, 2022).

[19] GeoSLAM, "Why is Geospatial Information so Important?," 2022. https://geoslam.com/blog/2021/10/05/why-is-geospatial-information-so-important/ (accessed May 02, 2022).

[20] QGIS, "Spatial Analysis (Interpolation)," *Documentation QGIS 2.18*. https://docs.qgis.org/2.18/en/docs/gentle_gis_introduction/spatial_analysis_interpolation.html (accessed May 09, 2022).

[21] L. Mitas and H. Mitasova, "Spatial interpolation," *Spat. interpolation. Geogr. Inf.*

Syst. Princ. Tech. Manag. Appl., vol. 1, no. 2, 1999.

[22] ArcGIS Pro, “Comprender el análisis de interpolación,” *Documentación ArcGIS Pro 2.8*. https://pro.arcgis.com/es/pro-app/2.8/tool-reference/spatial-analyst/understanding-interpolation-analysis.htm (accessed May 09, 2022).

[23] S. Y. Chung, S. Venkatramanan, H. E. Elzain, S. Selvam, and M. V. Prasanna, “Supplement of Missing Data in Groundwater-Level Variations of Peak Type Using Geostatistical Methods,” *GIS Geostatistical Tech. Groundw. Sci.*, pp. 33–41, Jan. 2019, doi: 10.1016/B978-0-12-815413-7.00004-3.

[24] A. Journel and C. Huijbregts, *Mining geostatistics*. 1976.

[25] A. Zucarelli, M. Paris, and J. Macor, “Rainfall variation structure in the province of Santa Fe (Argentina),” *Cad. do Lab. Xeolóxico Laxe. Rev. Xeol. Galega e do Hercínico Penins.*, vol. 41, pp. 59–73, 2019, Accessed: May 09, 2022. [Online]. Available: https://revistas.udc.es/index.php/CADLAXE/article/view/cadlaxe.2019.41.0.5814.

[26] A. Dauphiné, “Models of Basic Structures: Points and Fields,” *Geogr. Model. with Math.*, pp. 163–197, Jan. 2017, doi: 10.1016/B978-1-78548-225-0.50010-5.

[27] Ministerio para la Transición Ecológica y el Reto Demográfico, “Índice de Calidad del Aire.” https://www.miteco.gob.es/es/calidad-y-evaluacion-ambiental/temas/atmosfera-y-calidad-del-aire/calidad-del-aire/ICA.aspx (accessed Jun. 01, 2022).

[28] Secretaría de Ambiente del MDMQ, “Índice Quiteño de La Calidad Del Aire,” 2013. https://es.scribd.com/document/341447085/Indice-Quiteno-de-La-Calidad-Del-Aire (accessed Jun. 02, 2022).

[29] M. Asgari, M. Farnaghi, and Z. Ghaemi, “Predictive mapping of urban air pollution using apache spark on a hadoop cluster,” *ACM Int. Conf. Proceeding Ser.*, pp. 89–93, Sep. 2017, doi: 10.1145/3141128.3141131.

[30] A. Tong *et al.*, “Machine Learning on Spark for the Optimal IDW-based Spatiotemporal Interpolation,” *Int. Conf. GIScience Short Pap. Proc.*, vol. 1, no. 1, 2016, doi: 10.21433/B3114DW721GN.

[31] D. H. Shih, T. H. To, L. S. P. Nguyen, T. W. Wu, and W. T. You, “Design of a Spark Big Data Framework for PM2.5 Air Pollution Forecasting,” *Int. J. Environ. Res. Public Health*, vol. 18, no. 13, p. 7087, Jul. 2021, doi: 10.3390/IJERPH18137087.

[32] A. Asratyan, “Performance Analysis of Distributed Spatial Interpolation for Air Quality Data,” KTH Royal Institute of Technology, 2021.

[33] P. Mehta, C. Windolf, and C. De Agnès Voisard, “Spatio-Temporal Hotspot Computation on Apache Spark (GIS Cup),” 2016, doi: 10.1145/2996913.3004063.

[34] J. Yu, Z. Zhang, and M. Sarwat, “GeoSparkViz: A scalable geospatial data visualization framework in the apache spark ecosystem,” Jul. 2018, doi: 10.1145/3221269.3223040.

[35] A. Taoufikallah, “Mejora de la gestión de un proceso de producción de prototipos del sector automoción basada en la metodología Action-Research,” Escuela Superior de Ingenieros de Sevilla, 2010.

[36] B. Oates, *Researching Information Systems and Computing*. SAGE Publications Ltd., 2006.

[37] M. Martínez, “La investigación-acción en el aula,” *Agenda Académica*, vol. 7, no. 1, pp. 27–39, 2000.

[38] Asamblea Nacional, “Código Orgánico de Organización Territorial, COOTAD,” Quito, 2019. Accessed: May 30, 2022. [Online]. Available: https://www.cpccs.gob.ec/wp-content/uploads/2020/01/cootad.pdf.

[39] Asamblea Nacional, “Código Orgánico de Planificación y Finanzas Públicas,” Quito, 2010. Accessed: May 30, 2022. [Online]. Available: https://www.finanzas.gob.ec/wp-content/uploads/downloads/2012/09/CODIGO_PLANIFICACION_FINAZAS.pdf.

[40] L. Campozano, E. Sanchez, A. Aviles, and E. Samaniego, “Evaluation of infilling methods for time series of daily precipitation and temperature: The case of the Ecuadorian Andes,” *Maskana*, vol. 5, no. 1, 2014, Accessed: Jun. 17, 2022. [Online]. Available: https://publicaciones.ucuenca.edu.ec/ojs/index.php/maskana/article/view/431/371.

[41] Windfinder, “Unidades de la velocidad del viento y direcciones del viento,” 2022. https://es.windfinder.com/wind/windspeed.htm (accessed Aug. 27, 2022).

[42] E. Delgado, “El mapa: importante medio de apoyo para la enseñanza de la historia,” *Rev. Mex. Investig. Educ.*, vol. 7, no. 15, pp. 331–356, 2002, Accessed: Aug. 29, 2022. [Online]. Available: http://www.redalyc.org/articulo.oa?id=14001507.

[43] Agencia de Protección Ambiental de Estados Unidos (EPA), “Volcanes,” 2021. https://espanol.epa.gov/espanol/volcanes (accessed Jun. 24, 2022).

[44] National Geographic, “Cambio climático, sequías e inundaciones,” 2022. https://www.nationalgeographic.es/medio-ambiente/cambio-climatico-sequias-e-inundaciones (accessed Jun. 24, 2022).