Diseño de una arquitectura de inteligencia de negocios para apoyar la toma de decisiones en el sector público de Colombia: caso de estudio Antioquia

María Inés López Romero

Universidad EAN

Facultad de Ingeniería

Maestría en Inteligencia de Negocios

Bogotá, Colombia

19/05/2023

Diseño de una arquitectura de inteligencia de negocios para apoyar la toma de decisiones en el sector público de Colombia: caso de estudio Antioquia

María Inés López Romero

Trabajo de grado presentado como requisito para optar al título de:

Magister en Inteligencia de Negocios

Director (a):

Carolina María Luque Zabala

Modalidad:

Monografía

Universidad EAN

Facultad de Ingeniería

Maestría en Inteligencia de Negocios

Bogotá, Colombia

19/05/2023

Nota de aceptación:

_____

_____

_____

_____

_____

_____

Firma del jurado

_____

Firma del jurado

_____

Firma del director del trabajo de grado

Ciudad, día/mes/año

De ti proceden la riqueza y el honor; tú lo gobiernas todo. En tus manos están la fuerza y el poder, y eres tú quien engrandece y fortalece a todos. Por eso, Dios nuestro, te damos gracias, y a tu glorioso nombre tributamos alabanzas. (1 Crónicas 29:12-13)

Agradecimientos

Quiero expresar mi agradecimiento a la Universidad EAN y a todos sus colaboradores por sus enseñanzas y apoyo incondicional durante el curso de mi maestría. En especial, deseo agradecer a la profesora Carolina María Luque Zabala por haber sido mi directora de tesis y por su paciencia y dedicación en guiarme hacia la culminación de este importante logro académico.

Resumen

El gobierno colombiano ha impulsado el uso de las tecnologías de la información y el uso de datos en el marco de la cuarta revolución industrial. Para ello, ha establecido políticas y estándares que facilitan la disponibilidad y uso de datos abiertos de diversos entes gubernamentales en pro de comprender de una forma integral las dinámicas socioeconómicas en los territorios. En esta monografía se propone el diseño de una arquitectura de inteligencia de negocios para favorecer la articulación de datos de diferentes entidades con el fin de extraer información para examinar la relación entre datos generados por distintas instituciones y así apoyar la toma de decisiones en términos de políticas públicas. El diseño se ilustra con datos del departamento de Antioquia disponibles para el periodo 2019-2. Específicamente, utilizo datos relacionados con la cobertura de servicios públicos por municipio, el puntaje global de las pruebas Saber 11-2 e información geográfica de los municipios del departamento. El carácter georreferenciado de los datos conlleva a un estudio cuantitativo, aplicado e inductivo. Los hallazgos conducen a proponer algunas orientaciones metodológicas para el análisis geoespacial en el marco de una arquitectura de BI. Finalmente, los resultados resaltan la utilidad de la inteligencia de negocios para la integración de datos de diversas entidades para la toma de decisiones en el sector público.

Palabras clave: Arquitectura de Inteligencia de Negocios, toma de decisiones, sector público, analítica geoespacial, sistemas de información geográfica, datos geoespaciales, Antioquia.

Abstract

The Colombian government has promoted the use of information technologies and the use of data within the framework of the fourth industrial revolution. To this end, it has established policies and standards that facilitate the availability and use of open data from various government entities to comprehensively understand the socioeconomic dynamics in the territories. This monograph proposes the design of a business intelligence architecture to favor the articulation of data from different entities to extract information to examine the relationship between data generated by different institutions and thus support decision-making in terms of public policies. The design is illustrated with data available from the department of Antioquia, for the period 2019-2. Specifically, I use data related to the coverage of public services by municipality, the global score of the Saber 11-2 tests, and geographic information of the municipalities of the department. The georeferenced nature of the data leads to a quantitative, applied, and inductive study. The findings lead to propose some methodological guidelines for geospatial analysis within the framework of a BI architecture. Finally, the results highlight the usefulness of business intelligence for the integration of data from various entities for decision-making in the public sector.

Keywords: Business Intelligence Architecture, public sector, decision-making, geospatial analysis, geospatial data, Geographic Information Systems, Antioquia.

Contenido

Lista de Figuras

Lista de Tablas

Pág.

1. Introducción

La inteligencia de negocios (BI por sus siglas en inglés) se define como un conjunto de hardware, software y mejores prácticas que permiten el acceso interactivo a los datos, su gestión y análisis para una adecuada toma de decisiones en las organizaciones (Joyanes, 2019). Vargas y Valladares (2019) manifiestan que, si bien la inteligencia de negocios es mejor conocida en el sector privado, su aplicación en el sector público es útil para generar estrategias que permitan monitorizar y optimizar recursos e indicadores nacionales en pro de mejorar las condiciones (sociales, económicas, entre otras) de la población. Por su parte, Munné (2016) afirma que los gobiernos que hacen uso eficiente de los datos que generan sus entidades territoriales mejoran positivamente la comunicación con los ciudadanos y entienden mejor el clima político, social, cultural y económico de la nación. Este autor también argumenta que la integración de diferentes fuentes de datos y la disposición de datos abiertos favorece la gobernabilidad puesto que facilita la prevención del crimen, la reducción de riesgos de fraude en la administración de las finanzas públicas, entre otros.

En Colombia, el gobierno ha establecido las políticas CONPES 3920 (Big Data) y CONPES 3975 (Transformación Digital e Inteligencia Artificial) para incentivar el uso de los datos en el sector público del país. Estas políticas tienen como objetivo potenciar la generación de valor social y económico en el país a través del uso estratégico de tecnologías digitales en el sector público y el sector privado, para impulsar la productividad y favorecer el bienestar de los ciudadanos (Conpes, 2019). De esta forma, las organizaciones y los ciudadanos disponen de documentos oficiales que sientan las bases para el aprovechamiento de los datos generados por instituciones públicas y privadas, así como para la apropiación y la aplicación de las tecnologías relacionadas con la cuarta revolución industrial (4RI). Investigaciones recientes indican que el sector público de Colombia debe preparase y avanzar en la extracción de conocimiento a partir de los datos que generan y recopilan las diferentes entidades porque estos son activo estratégico de la nación y su uso mediante la implementación de soluciones de BI permite contribuir al plan de desarrollo del gobierno (Varona-Taborda, et al., 2021).

Así mismo y como parte de la misma iniciativa, el gobierno colombiano promueve la publicación de datos abiertos con información de diferentes entes gubernamentales para su aprovechamiento. Su publicación permite la colaboración entre organizaciones públicas y privadas, la academia y la sociedad civil y está enmarcada en los documentos citados

previamente como una estrategia del Estado colombiano para apoyar la transformación digital del país. Sin embargo, actualmente son varias las entidades territoriales que no tienen la posibilidad, de unificar, relacionar y gestionar de forma sistemática sus datos con los que provienen de otras instituciones gubernamentales (Varona-Taborda et al., 2021). No existe una plataforma pública que permita la integración de datos de diversas fuentes del sector público para su análisis en conjunto o un documento oficial que guie como integrar los datos de diversas entidades públicas para asegurar la calidad de los análisis generados.

Partiendo de la premisa de que una arquitectura BI favorece la articulación de datos de diferentes entidades con el fin de extraer información relevante para examinar la relación entre datos generados por distintas instituciones y así apoyar la toma de decisiones en términos de políticas públicas, se plantea diseñar una arquitectura BI de cinco (5) capas inspirada en el trabajo de Ong et al (2011), modificándola con la adición de una capa base de contexto que oriente el diseño de la arquitectura BI.

El diseño de la arquitectura de BI se ilustra con datos del departamento de Antioquia disponibles para el periodo 2019-2. Específicamente, utilizo datos relacionados con la cobertura de servicios públicos por municipio, el puntaje global de las pruebas Saber 11-2 e información geográfica de los municipios del departamento en cuestión. Luego, el objetivo es diseñar una arquitectura de inteligencia de negocios, para apoyar la toma de decisiones en el sector público del departamento de Antioquia. Particularmente, exploro una posible relación entre la cobertura de los servicios públicos domiciliarios, los puntajes globales de las pruebas Saber 11 y las condiciones geográficas del territorio departamental. En este sentido, extraigo y realizo una articulación de datos provenientes del Departamento Administrativo de Planeación pública de Antioquia (DAP), el Instituto Colombiano para la Evaluación de la Educación (Icfes) y el Instituto Geográfico Agustín Codazzi (IGAC).

Actualmente la implementación de metodologías de BI para mejorar la gestión pública en Colombia es una necesidad y un campo de investigación activo (ver Colmenares-Quintero, et al., 2021; Ordóñez y Gonzales, 2021; Varona-Taborda et al., 2021). Se espera, entonces, que el diseño de una arquitectura de BI ilustre la forma cómo se puede incorporar este tipo de metodologías para apoyar la toma de decisiones en términos de políticas públicas en el contexto social y educativo del país y en particular, en el departamento de Antioquia.

El documento se estructura como sigue. La sección 2 presenta objetivo general y objetivos específicos. La sección 3 expone la justificación y la sección 4 presenta el marco teórico, mientras que en la sección 5 se plantea la hipótesis. Por su parte en la sección 6 se explican las variables y fuentes de información. En la sección 7 se exhibe el diseño metodológico y en la sección 8 se muestran los resultados de su aplicación. En la sección 9 se presenta una discusión de los resultados. En la sección 10 se proponen orientaciones metodológicas para el análisis geoespacial en el marco de una arquitectura BI y en la sección 11 se presentan las conclusiones y el trabajo a futuro

2. Objetivos

2.1 Objetivo general

Diseñar una arquitectura de inteligencia de negocios para apoyar la toma de decisiones en el sector público de Colombia tomando como caso de estudio el departamento de Antioquia para el periodo 2019-2.

2.2 Objetivos específicos

- Establecer en la literatura los referentes teóricos sobre el diseño de una arquitectura de inteligencia de negocios y la aplicación de métodos estadísticos adecuados para examinar datos provenientes de entidades públicas.

- Identificar fuentes de información que permitan recuperar datos sobre el departamento de Antioquia para ilustrar procesos de extracción, transformación y carga con datos abiertos provenientes de distintas entidades gubernamentales.

- Configurar un sistema de almacenamiento de datos que consolide datos de diferentes entidades gubernamentales para el departamento de Antioquia.

- Generar reportes e implementar técnicas de visualización, minería de datos y modelamiento a partir de datos abiertos para apoyar la toma de decisiones en el sector público de Colombia tomando como caso particular el departamento de Antioquia.

- Proponer orientaciones metodológicas para el análisis geoespacial en el marco de una arquitectura de BI.

3.Justificación

Varona-Taborda, et al. (2021) señalan que en Colombia aún se hacen esfuerzos por implementar soluciones de BI que faciliten la extracción, procesamiento y análisis de datos para establecer colaboraciones entre entidades públicas. Además, algunos estudios exponen la necesidad de investigar sobre la forma cómo se pueden implementar metodologías de BI para mejorar la gestión pública en el país (Colmenares-Quintero, et al., 2021; Ordóñez y Gonzales, 2021; Varona-Taborda et al., 2021). En este sentido, propongo describir el diseño de una arquitectura de BI e ilustrarla con datos abiertos del departamento de Antioquia, con el propósito de contribuir desde la academia a impulsar la puesta en marcha de metodologías de BI en el sector público de Colombia para facilitar la toma de decisiones gubernamentales.

Una arquitectura de BI beneficia muchos actores. En particular, este tipo de implementaciones facilitan la consolidación, visualización y análisis de bases de datos por parte del personal administrativo con el fin de recuperar conocimiento a partir de los datos de manera eficiente y oportuna, y con la intención de mejorar la toma de decisiones en términos de políticas públicas encaminadas al bienestar de la población.

Esta monografía toma como caso de estudio el departamento de Antioquia. La elección de este territorio se fundamenta en varias razones. En primer lugar, es un departamento para el cual hay disponibles datos abiertos de diferentes entidades públicas para un mismo periodo de tiempo, lo cual permite ilustrar la articulación de datos provenientes de diferentes fuentes de información. En segundo lugar, es el departamento con el mayor número de municipios en el país (125) y con una amplia diversidad de condiciones sociales, económicas, culturales y geográficas en su territorio. Esto último, permite ilustrar el diseño de la arquitectura de BI en escenarios heterogéneos con relación a sus características de cobertura de servicios públicos, resultados en pruebas educativas estandarizadas y condiciones geográficas. Así, tomar este departamento como caso de estudio es una pequeña aproximación de lo que implicaría diseñar una arquitectura de BI para el territorio nacional.

Finalmente, esta monografía tiene como fundamento los aprendizajes adquiridos en la Maestría de Inteligencia de Negocios de la Universidad EAN. Además, se circunscribe en los temas de investigación del grupo ONTARE, bajo la línea de las tecnologías de la información y comunicación.

4. Marco Teórico

4.1 Inteligencia de Negocios

Schmitt, M. (2023) señala que la inteligencia de negocios (BI, por sus siglas en inglés) es un campo interdisciplinario que busca convertir datos sin procesar en información valiosa para las instituciones, a través del engranaje entre la analítica de datos, los sistemas de información y la gestión del conocimiento. En este sentido, la BI ofrece componentes teóricos, prácticos y tecnológicos útiles para optimizar recursos y tomar decisiones de forma eficiente dentro de las organizaciones a partir de conocimiento extraído de los datos que se generan producto de sus operaciones. Así, las tecnologías de BI permiten la captura, integración, almacenamiento y aplicación de herramientas de analítica de datos para tomar decisiones fundamentadas en evidencia empírica (Loshin,2013; Ong et al, 2011; Kalelkar et al, 2014).

4.2 Arquitectura de Inteligencia de Negocios

Una arquitectura BI es una herramienta que estructura la captura, integración, almacenamiento, análisis de datos, presentación y uso de información en un sistema BI. Según Kalelkar (2014) un diseño inadecuado de una arquitectura de BI conlleva a ineficiencias en el sistema BI e impacta negativamente a la organización que toma decisiones basadas en los reportes que se generan siguiendo la arquitectura.

Van der Lans (2018) y Sherman (2018) describen varias arquitecturas BI. Por ejemplo, afirman que existen arquitecturas en donde las fuentes de datos se conectan directamente a las plataformas de reporte, como Power BI o Tableau, o hay otros diseños en donde las fuentes de datos alimentan a la bodega de datos (DWH por sus siglas en inglés) y de allí se desprenden varios almacenes de datos específicos (denominados datamarts). Hay arquitecturas que se conectan también desde la DWH a sistemas de analítica en línea. Estos autores también describen arquitecturas que no tienen una DWH, sino una serie de datamarts interconectados entre sí y desde los cuales se pueden generar reportes combinados.

La arquitectura BI que se propone en esta monografía se basa en el modelo de cinco (5) capas de Ong et al. (2011). Según Joyanes (2019), la arquitectura tradicional de inteligencia de negocios propuesta por Ong et al. (2011) considera diversos aspectos importantes, tales como la calidad y el valor de los datos, mediante la aplicación de una metodología que garantiza el flujo

de información del sistema, a través de un proceso transparente. Esta arquitectura se estructura en cinco capas que incluyen la capa de fuentes de datos, el de proceso de extracción, transformación y carga (ETL por sus siglas en inglés), el almacen de datos (DWH, Datamart), los metadatos y la capa de usuario final, destinada a la visualización y análisis de resultados. De esta forma, la arquitectura de BI de Ong et al (2011) está diseñada para asegurar la disponibilidad de los datos necesarios para la toma de decisiones de negocio, así como para facilitar su acceso y análisis por parte de los usuarios finales.

Para el diseño de la arquitectura BI propongo agregar una capa de contexto inicial que direccione la arquitectura de BI en una línea o problemática específica y así haya una articulación de todos los componentes de la arquitectura con la naturaleza del contexto y de los datos. Así, la arquitectura propuesta se compone de seis capas: contexto, fuente de datos, procesos ETL, almacenamiento, usuario final y metadatos (información sobre los datos y herramientas usadas en el sistema de BI). (Ver Figura 1).

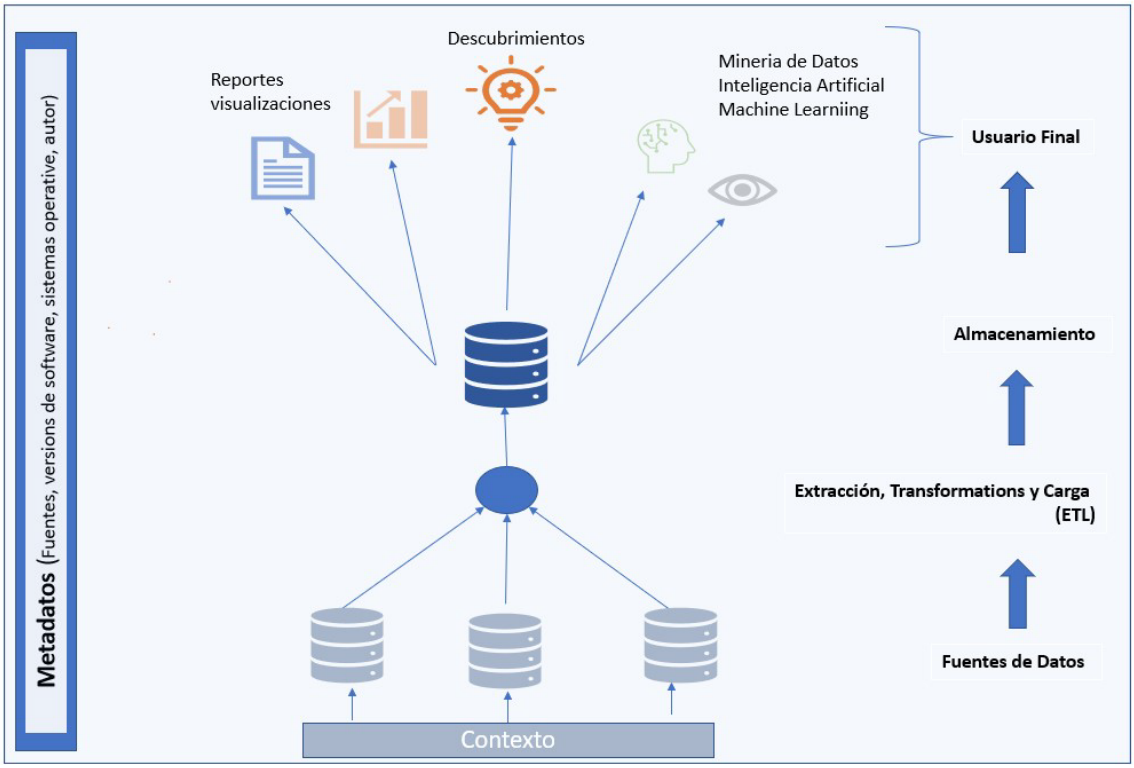


Figura 1. Arquitectura de BI de seis capas basado en Ong et al (2011)

Elaboración propia

4.2.1 Capa de contexto

Esta capa se refiere al objetivo específico que se quiere lograr con la arquitectura de BI. Por ejemplo, en muchas organizaciones, la arquitectura BI se diseña para lograr monitorear los índices clave de desempeño en diferentes materias como lo es el de ventas, costos, seguridad en la operación, gasto público entre otros. En el caso de esta monografía, el contexto lo proporciona la articulación de datos del sector público de Antioquia en pro de analizar cómo se relacionan los resultados de las pruebas estandarizadas con otros factores del territorio. Esta capa orienta la arquitectura, ya que permite definir qué datos se necesitan y están disponibles, que tipo de sistemas se necesitarán para su captura, su procesamiento y almacenamiento, así como que tipo de reportes y visualizaciones serán útiles para comunicar los resultados a las entidades que diseñan las políticas en los planes de desarrollo de los municipios, departamentos o a nivel nacional. La capa de contexto se agrega al modelo de cinco capas, al asociarse con el marco del CRISP-DM (del inglés Cross-Industry Standard Process for Data Mining), específico para la minería de datos, el cual se explica más adelante (Sección 4.2.5).

4.2.2 Capa fuente de datos

Según Fugini et al (2018) los datos son una fuente para explorar oportunidades dentro de un marco de toma de decisiones. Los datos permiten generar información y conocimiento, lo que apoya a que las organizaciones puedan aprender a navegar situaciones nuevas, a controlar variaciones y a regularse para lograr sus objetivos. La capa fuente de datos indica en donde se encuentran los datos a utilizar en el sistema de inteligencia de negocios. Las fuentes de datos pueden ser diversas. Por ejemplo, bases de datos SQL, hojas de cálculo, documentos de texto, páginas web, redes sociales, videos, imágenes, entre otros. También es importante anotar que estas pueden ser públicas o privadas, alojadas en servidores locales o en la nube. En el caso de los datos abiertos, estos son dispuestos en páginas web institucionales, con formatos interoperables, bajo licencias abiertas y de acceso público (Cervera, 2022).

En esta monografía se utilizan datos abiertos, estructurados y georreferenciados. Los datos estructurados tienen la característica principal de estar organizados en filas y columnas. Por su parte, los datos georreferenciados están asociados a coordenadas geográficas (Singleton et al, 2016).

Por otro lado, los datos de tipo georreferenciados pueden clasificarse en datos puntuales o de áreas (también denominados polígonos) y se almacenan en sistemas de información

geográfica (Singleton et al, 2016). Por ejemplo, los municipios de Antioquia en el mapa de la Figura 2 son polígonos a los cuales se les ha añadido diferentes propiedades que permiten visualizar la distribución espacial de las variables de interés, en este caso la cobertura de alcantarillado en los municipios de Antioquia en el 2019.

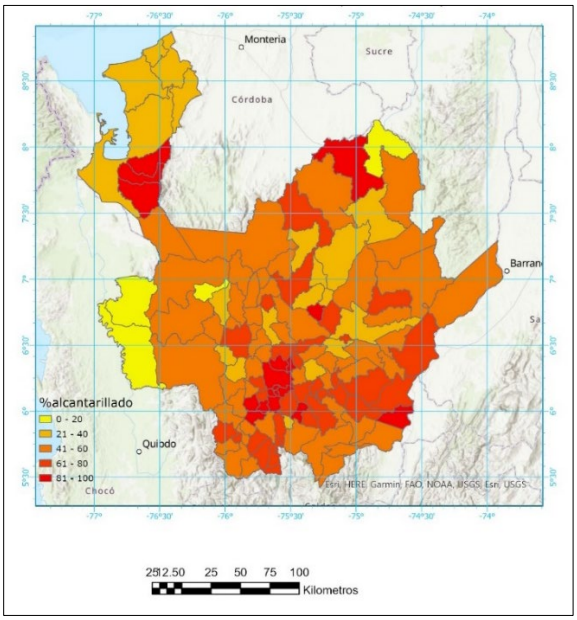


Figura 2. Porcentaje de Cobertura de Servicio de Alcantarillado en el Departamento de Antioquia por municipio.

Elaboración propia

Los datos georreferenciados permiten explorar ciertas características a la hora de aplicar analítica espacial, como la autocorrelación espacial (Fotheringham,2000). Estas características modifican la aplicación de técnicas de analítica y estadística tradicionales que se utilizan con datos no espaciales, con el propósito de incorporar características propias de los datos geoespaciales.

4.2.3 Capa extracción, transformación y carga

Esta capa denominada ETL por sus siglas en inglés, se refiere a la integración de los datos en el repositorio final. Sharda (2018) anota que la integración de los datos se realiza de acuerdo con el modelo físico del DWH o *datamart* y precisa tres fases. La primera es la extracción de los datos, la cual según Ong et al (2011) se refiere al proceso de identificar y agrupar los datos relevantes provenientes de diversas fuentes, los cuales son enviados a un repositorio temporal, en donde se harán los procesos de transformación antes de ser cargados al repositorio final.

La segunda fase, la transformación consiste en la limpieza de los datos de errores, datos incompletos, ajuste de formatos. También incluye agregación de datos, normalización, codificación y en general se refiere a la conversión de los datos según reglas lógicas que aseguren la consistencia y la calidad de los datos traídos de diversas fuentes (Kalelkar et al, 2014). La tercera fase consiste en cargar los datos transformados al repositorio final.

4.2.4 Capa de almacenamiento

Loshin (2013) se refiere al *DWH* como el repositorio de los datos provenientes de diversas fuentes de datos que han sido integrados en la capa ETL. Por su parte los *datamarts,* similar en estructura al *DWH,* son repositorios de subconjuntos de datos, que también pueden provenir de diferentes fuentes, pero que están dirigidos a apoyar a un tópico en específico dentro de la organización.

Según Ong et al (2011) los *DWH* y *datamarts* son construidos sobre modelos multidimensionales de datos que consisten en una o varias tablas de hechos (datos numéricos) y varias tablas de dimensiones (datos que caracterizan a los datos numéricos). Por ejemplo, en esta monografía la tabla de hechos estaría constituida por datos estadísticos que caracterizan a los entes territoriales y las tablas de dimensiones por los entes territoriales (municipios, departamentos, subregiones, regiones, entre otros) y por fechas (años, meses, días entre otros).

Estos repositorios de datos presentan las siguientes características: no volátiles, incrementan en el tiempo, están integrados y son específicos (Inmon en Van der Lans, 2012). No volátil se refiere a que los datos que entran al almacén de datos no se modifican, ni se borran. Estos datos se incrementan a medida que los datos se hacen disponibles y entran al repositorio. Integrados significa que provienen de diversas fuentes de información. Por último, los repositorios son específicos porque han sido diseñados en el marco de un objetivo específico.

En los sistemas de inteligencia de negocios hay varias formas de organizar los datos en los repositorios, siendo los esquemas más usados el de estrella y el de copo de nieve. El primero se caracteriza por tener una tabla de hechos, en la que se guardan los datos que son descritos por atributos y estos atributos son guardados en tablas de dimensiones que no están normalizadas. Por otro lado, el esquema copo de nieve es parecido al esquema estrella, excepto en que las tablas de dimensiones están normalizadas. (Van der Lans, 2012).

La creación de la estructura del repositorio en donde se almacenarán los datos presupone el conocimiento de los requerimientos del proyecto, la creación de un modelo conceptual, un modelo lógico y finalmente, un modelo físico (Sherman, 2018). Ver Figura 3.

Figura 3. Etapas modelamiento de datos

Fuente: Sherman (2018)

Una vez se ha establecido la estructura del DWH o del datamart, se procede a cargar los datos, desde donde el usuario final podrá acceder a los datos.

En esta Monografía el repositorio de datos se hará directamente en una geodatabase del programa ArcGIS, que es un Sistema de Información Geográfica (SIG). Aunque la geodatabase no se organiza con base en los esquemas descritos anteriormente (estrella y copo de nieve), si cuenta con una lógica que permite integrar los datos de diferentes tablas entre sí y puede guardar las principales características de un datamart: no volatidad, integrada y específica.

4.2.5 Capa de usuario final

El usuario final depende del objetivo de la arquitectura BI. En el caso de la presente monografía, el usuario final se referiría a los encargados de diseñar las políticas públicas y la planeación de desarrollo en las entidades territoriales, es decir, funcionarios públicos.

Esta capa consiste en las herramientas que el usuario final tiene para explorar la información, generar reportes y visualizaciones, aplicar analítica y minería de datos para la generación de conocimiento (Sherman, 2018; Ong et al., 2011). En esta monografía se aplicará la visualización de datos y la minería de datos como herramientas de usuario final para proporcionar conocimiento a partir de los datos. Sus resultados se mostrarían en un reporte escrito con visualizaciones y un cuadro de mando interactivo.

Esta capa también incluye la minería de datos. Este es un proceso por medio del cual se descubren patrones y tendencias a partir de un conjunto de datos con el fin de facilitar la toma de decisiones (Larose y Larose, 2014). Uno de los marcos de referencia para la aplicación de la

minería de datos es denominado CRISP-DM el cual presenta un conjunto de procedimientos estándar para manejar el ciclo de vida de un proyecto de minería de datos (Larose y Larose, 2014). Olson (2018) explica los seis pasos que sigue este estándar y hace énfasis en su naturaleza cíclica y no lineal. A continuación, se describen cada una de estas fases.

Fase 1 - Entendimiento del problema a resolver

El entendimiento del problema a resolver permite que el modelo y los resultados del modelo sean consistentes con los conceptos que se manejan en una disciplina o un negocio, igualmente facilita el entendimiento de datos considerados normales o atípicos dentro del contexto específico. Una vez se ha delimitado el problema a resolver se puede generar una estrategia y decidir que herramientas de minería de datos son adecuadas para resolver el problema.

Fase 2 - Entendimiento de los datos

Para el entendimiento de datos se utilizan las herramientas de análisis exploratorio de datos (EDA, por sus siglas en inglés), los cuales incluyen estadística descriptiva, representaciones gráficas y tabulares. En esta etapa se distinguen los tipos de variables (cuantitativas o cualitativas, continuas o discretas, numéricas o categóricas) y tipos de datos (georreferenciados, series temporales, estructurados entre otros).

Fase 3 - Preparación de los datos

En la fase de preparación de datos, se hace limpieza, transformación, conversión de unidades, normalización de datos y se puede realizar modelamiento de datos para integrar varias fuentes, por ejemplo, en un *datamart o DWH.*

Fase 4 - Modelamiento

La fase de modelamiento puede incluir varias técnicas de análisis de datos. Según Olson (2018), es aquí donde se aplican diversos algoritmos según el tipo de datos y necesidades. Los algoritmos disponibles se pueden aplicar con el objetivo de clasificar, agrupar o predecir.

Fase 5 - Evaluación del modelo

Esta fase busca probar la validez del modelo mediante la aplicación de diferentes técnicas dependiendo de los algoritmos aplicados para la generación del modelo.

Fase 6 - Generación de resultados

Larose y Larose (2014) anotan que esta última fase puede tomar la forma de un reporte, un cuadro de mando, un modelo que se aplicará en otros departamentos de una empresa, o un marco de referencia para la toma de una decisión. En esta monografía el resultado se mostrará a manera de reporte con visualizaciones.

Cabe anotar que, en esta monografía, la parte de preparación de datos está imbuida en la capa ETL de la arquitectura, pero no indica que, al necesitarse una modificación o transformación de los datos debido al proceso interactivo del descubrimiento de conocimiento, este no pueda hacerse después de un requerimiento desde la capa de usuario final. En la figura 4 se observa esquemáticamente el proceso para que el usuario final pueda disponer de los datos para su análisis.

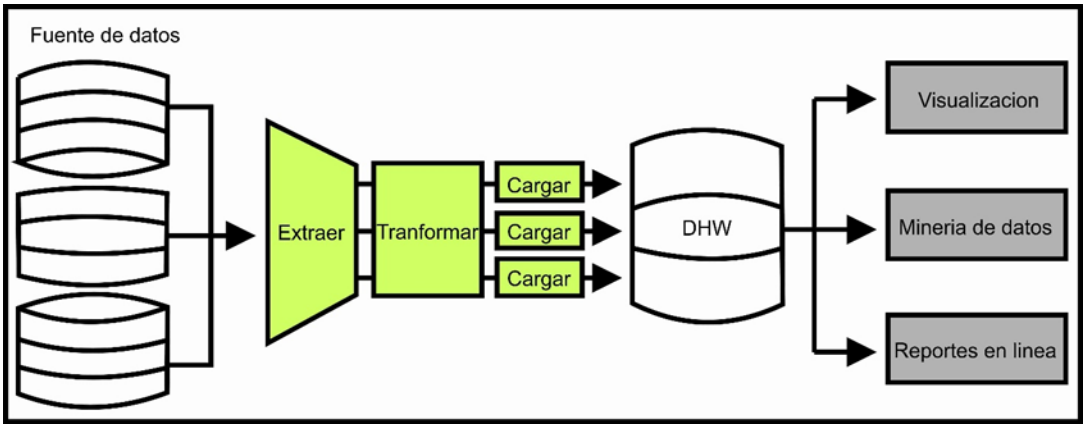


Figura 4. Esquema mostrando el proceso de ETL para llevar los datos al almacenamiento final que pueden ser requeridos por el usuario final- Fuente: Vidal (2014)

4.2.6 Capa de metadatos

Según Ong et al (2011), la capa de metadatos es transversal al resto de las capas. Según estos autores, el repositorio de metadatos se utiliza para almacenar información técnica y de negocios sobre los datos, así como reglas y definiciones de los datos en la organización (Davenport y Harris en Ong et al, 2011). Este repositorio es crítico para la gestión de los datos en un entorno de BI, ya que proporciona información detallada sobre los orígenes de los datos, las transformaciones realizadas, la estructura de los datos, las relaciones entre los mismos, y otros detalles relevantes (Ong et al., 2011).

Estos autores indican que la capa de metadatos garantiza la utilización responsable de los datos almacenados en la arquitectura. Así mismo, permite el mantenimiento y el buen funcionamiento de la solución BI, ya que debe guardar detalles de los diferentes componentes de la arquitectura de BI, al definir los programas utilizados, sus versiones y como están integrados en la arquitectura. Además, la capa de metadatos también proporciona información relevante sobre los informes y los cuadros de mando que se generan en la capa de usuario final.

4.3 Aplicación de Inteligencia de Negocios en el sector público

El proceso de inteligencia de negocios debe transformar datos, históricos y actuales, en conocimiento y permitir el descubrimiento de hechos y patrones que favorezcan la toma de decisiones. Entre tanto Vargas (2019) apunta que, si bien la inteligencia de negocios es mejor conocida en el sector privado, su aplicación en el sector público no es muy común, pero enfatiza que su utilización apoyaría a la toma de decisiones por parte de los funcionarios públicos en cuanto a dineros públicos, salud pública, bienestar social y en general el desarrollo del país teniendo en cuenta la información disponible. Es así como Hartley y Seymour (2011) apuntan que la adopción de sistemas de inteligencia de negocios por parte de las organizaciones públicas debe estar orientadas al mejoramiento de la prestación de servicios a la sociedad.

Por otro lado, en el sector privado la inteligencia de negocios se aplica para aumentar ganancias y reducir costos, al tomar decisiones basadas en el correcto uso de los datos para generar descubrimientos sobre fallas y oportunidades en sus procesos internos y externos (Williams y Williams, 2007). Si bien el objetivo del sector público no es incrementar ganancias, si lo es manejar eficientemente dineros públicos y reducir costos, además de incrementar la calidad de la prestación de servicios a la sociedad. Por su parte Elbashir et al (2022) afirman que, bajo

los paradigmas de la nueva administración pública, las organizaciones necesitan recurrir a la inteligencia de negocios para poder reportar las métricas que indican la eficiencia e impacto de sus servicios al público.

Sin embargo, la implementación de sistemas de inteligencia de negocios es una tarea compleja en el sector público, ya que se necesita un cambio cultural de las organizaciones, así como de personal con el conocimiento en las áreas técnicas como ingenieros de datos, programadores, científicos de datos y gerentes de tecnología que lideren estos programas en conjunto con los expertos en las materias de la administración pública (Hartley y Seymour, 2011; Abai et al, 2019; Elbashir et al, 2022).

En Colombia, el gobierno le ha apostado a la transformación digital desde la década de los 2000, cuando lanzó la iniciativa del gobierno electrónico. En los documentos CONPES 3975 (Inteligencia Artificial y Transformación Digital) y 3920 (Big Data) se han trazado las políticas que sostienen el uso y la explotación de los datos en un marco ético, para el desarrollo del país. En el CONPES 3920 se resalta la necesidad de ir más allá de la digitalización de los datos y su almacenamiento, y tomar el reto de cómo convertir los datos en insumo para la toma de decisiones que transformen al país para mejorar la calidad de vida de los colombianos y el desarrollo económico sostenible de sus regiones.

En Colombia, existen soluciones de BI que integran información para la toma de decisiones en campos específicos de entidades gubernamentales. Por ejemplo, SISPRO es un sistema del Ministerio de Salud y Protección Social que brinda “Información oportuna, suficiente y estandarizada para la toma de decisiones del Sector Salud y Protección Social, centrada en el Ciudadano. El SISPRO está conformado por bases de datos y sistemas de información del sector sobre oferta y demanda de servicios de salud, calidad de los servicios, aseguramiento, financiamiento y promoción social”. El DNP también ha puesto a la disposición de los ciudadanos varias aplicaciones, que se alimentan de las bases de datos oficiales para dar seguimiento a varios temas de interés como lo son los gastos públicos y el avance del plan plurianual de inversiones, por ejemplo, la Plataforma Integrada de Inversión Pública – PIIP. Por su parte el Ministerio de Hacienda y Crédito Público, ha dispuesto en la página https://www.pte.gov.co/ visores con datos relevantes al presupuesto general de la nación y como se distribuye por sector.

Por otra parte, la publicación de los datos abiertos se encuentra en diferentes repositorios que le permiten a cualquier ciudadano acceder a datos oficiales. En estas plataformas oficiales se encuentran herramientas para conectar los datos a bases de datos, también se encuentran geo-visores, cuadros de mando y reportes elaborados con herramientas BI, como lo son Power BI o Tableau. Sin embargo, estas herramientas no incluyen una arquitectura que les permita a los usuarios cruzar información de diversas fuentes y así facilitar la colaboración entre diferentes entes gubernamentales, pues solo permiten el aprovechamiento de la información de forma aislada.

En las siguientes secciones se abordarán, brevemente, nociones importantes que sustentarán el diseño de la arquitectura BI planteada en esta monografía. En la sección 4.3.1 se hace una breve reseña de las pruebas saber 11, en la sección 4.3.2 se hace una breve reseña de los servicios públicos domiciliarios en Colombia, en la sección 4.3.3 se denotan los estudios relacionados con los servicios públicos y los resultados de las pruebas Saber 11, mientras que en la sección 4.3.4 se indica cual es la relación entre los servicios públicos, como factor socioeconómico a analizar, y el rendimiento académico.

4.3.1 Pruebas Saber 11

Castro y Ruiz (2019) apuntan que los sistemas de evaluación de la educación colombiana miden el desempeño estudiantil frente a los estándares establecidos por los currículos. Más allá de la calidad de la educación, las pruebas evalúan la eficacia de las instituciones de impartir los conocimientos y habilidades que el gobierno ha establecido como las necesarias para forjar una población competente que soporte el desarrollo del país. Las Pruebas Saber 11, son las encargadas de medir estos aspectos en la educación secundaria en Colombia.

Las pruebas saber 11 se dividen en 5 asignaturas. Estas son matemáticas, ciencias naturales, sociales y ciudadanas, lectura crítica e inglés, las cuales se ponderan aproximadamente por un factor de 1.15 las 4 primeras y por un factor de 0.39 la última, es decir la materia de inglés tiene menos peso para el cálculo del puntaje global. El puntaje global se mide en una escala de 0 a 500. Este puntaje es una de las variables de interés en el presente trabajo. Con los resultados también se presentan otras variables, de las cuales se destaca el municipio de residencia del estudiante, por ser la otra variable de interés dentro de esta fuente de datos.

Según el ministerio de educación nacional (MEN) y el ICFES los objetivos de las pruebas Saber 11 se resumen en: medir las competencias de los estudiantes al acabar el onceavo grado de educación secundaria, orientar al estudiante mediante autoevaluación sobre sus habilidades y vocación Informar a las instituciones educativas y profesorado sobre su gestión en comparación con la media nacional, proporcionar información a las universidades y escuelas técnicas sobre las competencias de los aspirantes a cupos, generar datos para entender la calidad de la educación en Colombia, las competencias de los estudiantes, y formular políticas para su mejora.

La metodología que se utilizó para presentar las pruebas Saber 11 en el segundo semestre del año 2019 fue de manera presencial. Además de los cuadernillos con las pruebas y la hoja de respuestas, también se adjuntó un cuestionario, que no influyó en el puntaje final, para entender la situación socioeconómica y cultural del estudiante, con fines de investigación y mejora de la educación en Colombia. Este cuestionario indaga la composición del núcleo familiar, el estrato socioeconómico, las características de la vivienda, conexión a internet y a televisión por cable, así como las maneras de esparcimiento de las familias de los estudiantes. El estudio de Collazos et al (2021) utilizó este cuestionario para relacionar aspectos socioeconómicos con el logro académico en las pruebas Saber 11.

Las pruebas Saber 11 son de carácter nacional y se presentan dos veces al año, uno para el calendario A y otra para colegios de calendario B. El calendario A representa a los colegios cuyo año escolar va de febrero a noviembre y el calendario B a los colegios que tienen el año escolar entre septiembre a junio. Esta monografía utiliza los datos de las pruebas Saber 11 en los municipios de Antioquia para el calendario A.

Los resultados de las pruebas Saber 11 son publicados con la lista de estudiantes identificados por un código para respetar su privacidad, es decir están anonimizados para proteger a los individuos de revelar datos personales y relacionados con su rendimiento académico. Esto en concordancia con las recomendaciones del Conpes 3920 en donde se hace énfasis a la protección del derecho de habeas data, la privacidad e intimidad, para que los datos puedan ser activos en una economía del conocimiento que beneficien a la sociedad sin afectar a los individuos, uno de los riesgos a mitigar en la era de la inteligencia artificial y la explotación de macrodatos (Conpes, 2018).

4.3.2 Cobertura de servicios públicos en Antioquia

Los servicios domiciliarios básicos definidos en la Constitución Política de Colombia son: agua, electricidad, alcantarillado, aseo y gas. La ley 2108 de 2021, declara el servicio de internet como un servicio público esencial. Russo en Matias (2015) afirma que “los servicios domiciliarios son servicios públicos esenciales y derechos fundamentales, entendiéndose por tales aquellos sin los cuales no puede existir el hombre en su triple dimensión de ser físico (vivo), ser síquico (sensorial y racional) y ser social (conjunto de relaciones sociales)” a lo que Matías (2015) agrega que “son bienes insustituibles y su prestación es una actividad económica, que debe buscar la satisfacción de las necesidades esenciales de la población, en beneficio del mejoramiento de su calidad de vida y de la materialización de sus derechos fundamentales”.

Según el Anuario Estadístico de Antioquia, en el año 2019, el departamento contaba con una cobertura promedio de 56% de alcantarillado, 74.86% de acueducto, 98.33% de electricidad, un 65.29% de recolección de basuras, un 34.03% de gas domiciliario y un 99.18% de internet. Vale la pena anotar que, en Colombia, cada municipio asigna presupuesto a invertir en infraestructura de servicios públicos anualmente y a su vez éste depende de la riqueza del municipio, de la nación y de la voluntad política (Burgos, 2022).

El total de viviendas servidas por los servicios públicos en el departamento de Antioquia en el 2019 suman 2´096.981, que van desde el municipio de Abriaquí con 896 viviendas hasta Medellín que cuenta con 830.515 viviendas.

4.3.3 Estudios enfocados en pruebas Saber 11 y cobertura de servicios públicos

El Icfes ha reconocido que hay variables socioeconómicas que tienen un impacto en la educación de los estudiantes y ha incluido un cuestionario socioeconómico en las pruebas para ser respondida por los estudiantes, con preguntas que permitan entender la relación entre estas variables y el rendimiento en las pruebas Saber 11. Collazos et al (2021) realizaron un estudio cuantitativo para entender la relación entre los resultados de las pruebas Saber 11 y las características socioeconómicas de los estudiantes. Estos autores encontraron una correlación positiva entre el estrato socioeconómico y el rendimiento académico utilizando los resultados de las pruebas Saber 11 del 2014 al 2019.

Por su parte, Junca (2019) confirma haber encontrado múltiples trabajos que aseveran la existencia de la correlación entre factores socioeconómicos, medidos con base en promedio de ingreso de los hogares y clasificación socioeconómica del plantel educativo, y el desempeño de los estudiantes en las pruebas Saber 11. Por otro lado, Martínez y Turriago (2015) confirman esta relación mediante el uso del NBI (Índice de necesidades básicas insatisfechas) de los municipios colombianos y su comparación con el rendimiento en las pruebas saber 11 entre los años 2005 y 2012.

Aunque la revisión de la literatura ha mostrado que no existen estudios específicos de la relación de la cobertura de los servicios públicos por municipio y los resultados de la prueba Saber 11, se puede enmarcar en el análisis de la relación entre las características socioeconómicas de los municipios y el rendimiento en las pruebas Saber 11.

4.3.4 Relación de los servicios públicos y el rendimiento académico

Hay factores ambientales en los que los servicios públicos domiciliarios podrían mitigar su impacto en procesos cognitivos. Por ejemplo, Lundgren et al (2012) y Antonnen et al (2009) anotan que ciertas condiciones climáticas causan una disminución en las habilidades cognitivas y de destreza física. Las condiciones climáticas extremas se pueden atenuar por medio de servicios de electricidad, gas domiciliario y acueducto. Por otro lado, la falta de acueducto, alcantarillado y recolección de basuras puede aumentar el riesgo de enfermedades relacionadas con parásitos, virus y hongos, lo cuales pueden causar problemas de memoria, atención y aprendizaje en el largo plazo, ya que causan o empeoran desnutrición y anemia, o porque afectan el cerebro directamente (Ezeama et al, 2006: Ezeama et al, 2008; Ezeama et al, 2019; Idrovo, 2012). Además, temas como no contar con electricidad o acueducto o alcantarillado o recolección de basuras, suponen un esfuerzo adicional para estudiar (Burleson y Thoron, 2014). También existen estudios, en los que se encuentra relación entre polución y la disminución de los procesos cognitivos. La recolección de basuras y la correcta ubicación de los botaderos ayudaría a mitigar este problema, que se agudizaría en zonas en donde hay muchas fuentes de material particulado originado por actividades humanas industriales.

Finalmente, El acceso a la información a través de internet también afecta al rendimiento de los estudiantes. Aquellos que no cuentan con este servicio muestran un rendimiento académico inferior frente a quienes lo tienen disponible (Collazos et al, 2021; Diaz et al, 2021).

4.4 Métodos de analítica de datos

Según Van Barneveld en en Sedkaoui (2018) la analítica de datos es el proceso por medio del cual se implementan métodos de descripción, modelamiento, análisis, e interpretación de datos. Su desarrollo requiere del uso de la estadística, algoritmos computacionales y sistemas de información. Según Sedkaoui (2018), la analítica se clasifica en descriptiva, predictiva y prescriptiva según su objetivo.

La analítica descriptiva se apoya en la estadística descriptiva y herramientas de visualización para entender los datos disponibles en un almacén de datos (Groebner et al, 2018). Existen varios tipos de gráficos utilizados en la analítica descriptiva como los son los de dispersión, de barras, de líneas, diagramas de torta, mapas de árbol, histogramas. Los sistemas de información geográfica apoyan la visualización de variables espaciales con mapas temáticos, por ejemplo, mapas de cuantiles, mapas de calor, entre otros. Por medio de la analítica descriptiva se entiende la situación pasada o actual de las variables en estudio y posibilita la formulación de metodologías para aplicar técnicas de analítica predictiva.

La analítica predictiva utiliza algoritmos con el fin de entender una situación para tomar decisiones que ayuden a optimizar resultados en una organización. Groebner et al (2018) también afirman que los modelos predictivos son utilizados para anticipar el comportamiento de las variables dependientes dados cambios en las independientes. Hay varias técnicas de minería de datos y algoritmos que se pueden aplicar según el objetivo del modelo. Por ejemplo, técnicas de clasificación, regresión, detección de anomalías, reglas de asociación, entre otros. Algunos sistemas de información geográfica incorporan algoritmos para aplicar diversas técnicas de minería de datos con datos geoespaciales, tales como Geoda versión 1.20.0.22 o ArcGIS Pro versión 2.9.

Por su parte la analítica prescriptiva utiliza técnicas de modelado matemático y análisis de datos para recomendar acciones específicas que pueden llevarse a cabo para alcanzar un resultado deseado. Los enfoques basados en modelos pueden utilizar software de optimización para probar muchos escenarios diferentes hasta que se encuentre uno que cumpla mejor con los criterios de optimización (Greasley,2019). Es decir, a partir de los modelos encontrados a partir de la analítica descriptiva y predictiva, se busca aplicar una acción que lleve a un resultado óptimo.

4.4.1 Modelos de regresión

Los modelos de regresión permiten explorar relaciones entre una variable dependiente y una o más variables independientes (predictores) con el propósito de explicar cómo influyen los predictores en el comportamiento de la variable dependiente. En el caso de los datos geoespaciales (los cuales representan características de las entidades territoriales) se utilizan modelos de auto regresión espacial simultanea (SAR), con los cuales se busca entender la relación entre una variable dependiente y una o más independientes teniendo en cuenta la estructura espacial de los datos (Acevedo, 2012).

Matriz de pesos

La estructura espacial de datos geoespaciales se puede determinar usando el concepto de contigüidad y de distancia. Acevedo (2012) explica que el método de contigüidad se define, como el número de polígonos que se encuentran adyacentes a un polígono que se usa como referencia. En tanto, la distancia es una métrica que permite cuantificar la cercanía de los centroides de los polígonos entre sí. Las distancias permiten construir una matriz de pesos, que resume la estructura espacial de los datos. En la Figura 5, se ilustra la matriz teniendo en cuenta el número de polígonos adyacentes y a partir de las distancias entre los centroides. En la parte a y c de la figura la matriz se construye a partir de la adyacencia. Si los polígonos comparten bordes se marca como 1, en caso contario 0. En la parte b y d de la figura la matriz de pesos se construye con las distancias en los centroides.

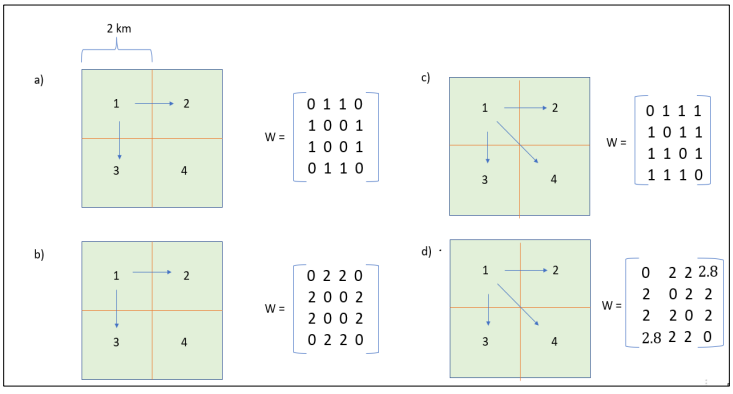


Figura 5. Ilustración de la construcción de la matriz de pesos.

Fuente: Elaboración Propia

Autocorrelación espacial

Cuando se comparan variables en un espacio geográfico, es importante entender que tanto afecta la componente espacial a ambas, para saber si los coeficientes encontrados son debido a una genuina correlación entre ellos o si ambos están afectados de la misma manera por la componente espacial. La autocorrelación espacial, es decir, la relación de los datos de una variable entre sí a causa de la variación espacial se determina utilizando estadísticos como el índice de Moran.

El índice global de Moran mide la autocorrelación espacial global y su valor varía entre 1 y -1. La interpretación del índice global de Moran es la siguiente: valores positivos cercanos a 1 indican autocorrelación espacial positiva, valores negativos cercanos a -1 indican autocorrelación espacial negativa y valores cercanos a cero indican falta de autocorrelación espacial. (Ordóñez Galán en Castillo et al 2015). La expresión matemática del Índice de Moran global es la siguiente (Rogerson en Castillo et al, 2015)

$$I=\frac{n\Sigma_i\Sigma_j w_{ij}(x_i-\bar{x})(x_j-\bar{x})}{(\Sigma_i\Sigma_j w_{ij})\Sigma(x_i-\bar{x})^2}$$

En donde n es el número de muestras, x_i es el valor de la variable en un área determinada y x_j el valor de la variable en otra área (donde $i \neq j$), $\bar{x}$ es la media de la variable y w_{ij} es un peso aplicado a la comparación entre la localización i y la localización j (Castillo et al, 2015).

Por otra parte, el multiplicador de Lagrange (LM, por sus siglas en inglés) es una prueba estadística que se utiliza para evaluar la autocorrelación espacial específicamente de los residuos de un modelo de regresión espacial (Matthews, 2006; Fotheringham, 2000; Acevedo, 2012). La fórmula es:

La fórmula del LM es:

$$LM = n \times R^2$$

Donde n es el tamaño de la muestra y R^2 es el coeficiente de determinación de la regresión auxiliar que se utiliza para modelar la relación entre los residuos del modelo y los valores vecinos.

Ecuación modelo SAR

Según Matthew (2018) si se determina autocorrelación espacial substancial, es decir que la ubicación espacial influye en el comportamiento de la variable objeto de análisis, se aplica el modelo de auto regresión espacial por rezago dado por la fórmula:

$$y - \rho W y = \beta_0 + \beta_1 x + \varepsilon$$

En donde y representa la variable dependiente, x representa a la variable independiente, Wy representa una matriz de pesos y ρ representa su coeficiente de autocorrelación (Fotheringham et al, 2000). Acevedo (2012) también anota que la estructura de correlación espacial determina el parámetro ρ que luego explica la intensidad de la correlación de la respuesta.

Anselin (2005) incorpora el método de máxima verosimilitud (ML por sus siglas en inglés) en el software Geoda, para ajustar el modelo SAR. ML es un método de estimación estadístico que permite encontrar los de los parámetros del modelo que maximizan la probabilidad de obtener los datos observados (James et al, 2017).

Geoda arroja varios parámetros al realizar una regresión espacial tipo SAR, entre ellas un pseudo-R^2 , el cuál es válido para analizar la dirección y fuerza explicativa del modelo. Anselin (2005) menciona que el valor del pseudo-R^2 debe ser analizado en conjunto con la significancia de los coeficientes, para así confirmar que la variable dependiente no solo cambia por influencia de la estructura espacial, sino que también cambia en función de las otras variables independientes elegidas en el modelo.

5. Hipótesis

El diseño de una arquitectura de BI favorece la articulación de datos de diferentes entidades gubernamentales con el fin de extraer información relevante para examinar la relación entre los datos que generan distintas entidades y así apoyar la toma de decisiones en términos de políticas públicas.

6. Fuentes de datos y variables

6.1 Datos abiertos en Colombia

En Colombia el Ministerio de Tecnologías de la Información y las Comunicaciones (MinTIC) es el encargado de delinear las políticas y mejores prácticas para la publicación de los datos abiertos oficiales en el portal https://www.datos.gov.co/ (Cervera, 2022; Arboleda y Anaya, 2017). El CONPES 3920 del 2018 es el documento que establece la política para el uso de los datos abiertos en Colombia.

Por su parte Cervera (2022) afirma que el uso adecuado de los datos abiertos apoya al mejoramiento de temas estratégicos, como la educación, respecto a su “transparencia y la promoción articulada del trabajo articulado entre universidad, empresa y Estado”. Esta autora también afirma que desde el 2019 ha habido un aumento significante en la publicación de los datos abiertos en el país y así mismo su uso, y los temas más consultados fueron relacionados con la salud y protección social, educación y función pública (Figura 6).

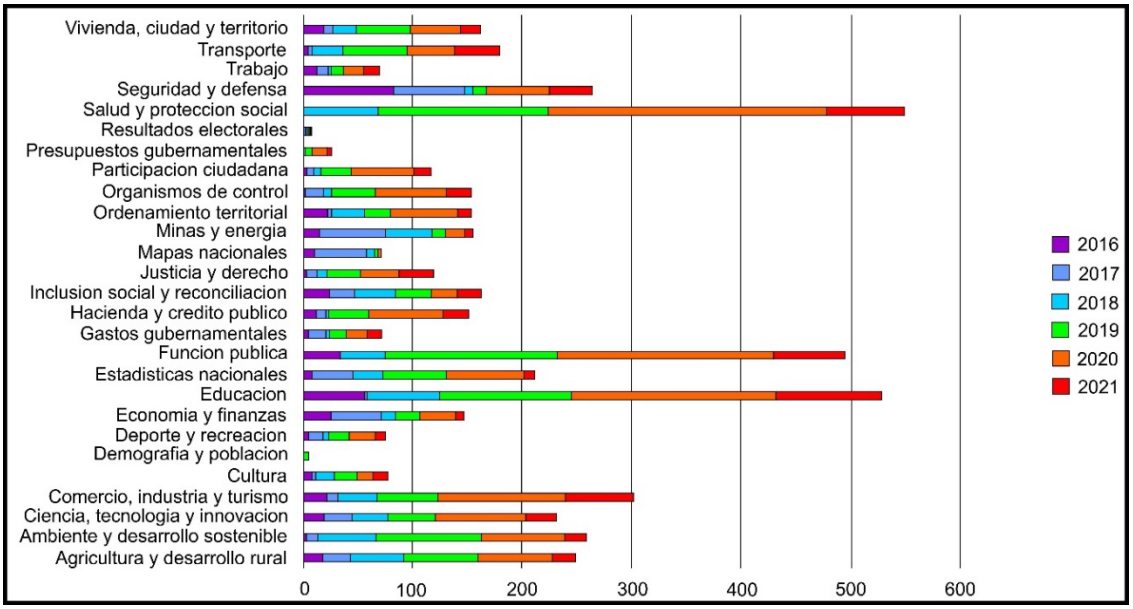


Figura 6.Número de publicaciones basadas en los datos abiertos del portal datos.gov.co. Fuente: Cervera (2022)

Según la Guía de Datos Abiertos de Colombia, publicado por el MinTIC, los beneficios que trae la publicación de los datos abiertos se pueden resumir en transparencia y control social, mejoramiento o creación de productos, servicios y negocios innovadores, pronóstico y prevención de fenómenos y generación de conocimiento.

El MinTIC en el documento Hoja de Ruta de los Datos Abiertos Estratégicos para el Estado Colombiano (2021), afirma que, debido a la gran cantidad de datos producidos por entidades públicas y privadas, se debe priorizar y determinar los datos que son estratégicos para que se asegure su disponibilidad y actualización en el portal oficial. La priorización se basó en el Plan de Desarrollo Nacional (2018-2022), temas priorizados por organizaciones internacionales como la OCDE y la ONU y el Programa Interamericano Anti-Corrupción (PIDA). En esta hoja de ruta se priorizaron 16 categorías de información y entre ellas se encuentran el de educación y vivienda, ciudad y territorio, temas relevantes en esta monografía. El uso de los datos abiertos publicados por las entidades gubernamentales, aseguran su confiabilidad y calidad, ya que existen estándares para su preparación, limpieza, agregación y publicación.

6.2 Fuentes de datos

La arquitectura BI que se diseñará en este trabajo, para ejemplificar su funcionamiento a servicio de la toma de decisiones en el sector público, se basará en los datos resultantes de las

pruebas Saber 11-2 para los 125 municipios de Antioquia (Colombia), la cobertura de los servicios públicos de los mismos municipios y su geolocalización.

Los datos de las pruebas Saber-11 se obtuvieron de la plataforma de datos abiertos. El archivo contiene los resultados a nivel nacional, así que se descargaron en formato csv y se filtraron para solo almacenar los datos referentes a los estudiantes de Antioquia. Este archivo contiene los resultados anonimizados por estudiante, más sus respuestas al cuestionario socioeconómico. La Figura 7 muestra una imagen del formato de este archivo. 34,042 estudiantes presentaron las pruebas Saber 11-2 en el departamento de Antioquia.

F1 ESTU_MCPIO_RESIDE

	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	ESTU_MCPIO_RESIDE	ESTU_COD_I	COLE_GENER	COLE_NATUR	COLE_CARACTER	COLE_AREA_	COLE_JORNA	ESTU_DEPTO	ESTU_COD_D	PUNT_LECTU	PUNT_MATEI	PUNT_C_NAT	PUNT_SOCIA	PUNT_INGLES	PUNT_GLOBAL
2	MEDELLÍN	05001	MIXTO	OFICIAL	TÉCNICO/ACADÉMICO	URBANO	NOCHE	ANTIOQUIA	5	54	49	40	33	44	220
3	MEDELLÍN	05001	MIXTO	OFICIAL	TÉCNICO/ACADÉMICO	URBANO	UNICA	ANTIOQUIA	5	45	50	53	52	50	250
4	RIONEGRO	05615	MIXTO	NO OFICIAL	ACADÉMICO	URBANO	SABATINA	ANTIOQUIA	5	77	61	61	61	75	329
5	BELLO	05088	MIXTO	OFICIAL	TÉCNICO	URBANO	NOCHE	ANTIOQUIA	5	56	61	46	46	53	262
6	SAN JUAN DE URABÁ	05659	MIXTO	OFICIAL	ACADÉMICO	RURAL	MAÑANA	ANTIOQUIA	5	54	54	42	52	48	252
7	MEDELLÍN	05001	MIXTO	NO OFICIAL	ACADÉMICO	URBANO	MAÑANA	ANTIOQUIA	5	39	26	36	35	31	169
8	MEDELLÍN	05001	MIXTO	OFICIAL	TÉCNICO/ACADÉMICO	URBANO	SABATINA	ANTIOQUIA	5	48	42	39	31	27	195
9	MEDELLÍN	05001	MIXTO	OFICIAL	TÉCNICO/ACADÉMICO	URBANO	MAÑANA	ANTIOQUIA	5	59	39	38	41	39	219
10	SOPETRÁN	05761	MIXTO	OFICIAL	ACADÉMICO	URBANO	SABATINA	ANTIOQUIA	5	37	48	34	38	31	193
11	MEDELLÍN	05001	MIXTO	OFICIAL		URBANO	MAÑANA	ANTIOQUIA	5	52	54	58	47	57	265
12	CHIGORODÓ	05172	MIXTO	OFICIAL	TÉCNICO/ACADÉMICO	URBANO	SABATINA	ANTIOQUIA	5	46	34	41	49	37	210
13	BELLO	05088	MIXTO	NO OFICIAL	ACADÉMICO	URBANO	MAÑANA	ANTIOQUIA	5	26	26	32	39	32	154
14	APARTADÓ	05045	MIXTO	OFICIAL	ACADÉMICO	RURAL	MAÑANA	ANTIOQUIA	5	49	37	45	50	38	223
15	EL BAGRE	05250	MIXTO	OFICIAL	ACADÉMICO	RURAL	TARDE	ANTIOQUIA	5	49	50	48	40	42	232

Figura 7. Visualización de archivo en csv de los resultados de las pruebas Saber 11-2 2019

Este archivo se filtrará para tomar solo los datos referentes a Antioquia, y los resultados de las pruebas se agregarán calculando el promedio de los puntajes globales por municipio. A este archivo se le agregará una columna con el número de estudiantes por municipio. Este archivo se exportará a formato csv.

Los datos referentes a la cobertura de los servicios públicos provienen de datos abiertos disponibles en el Anuario Estadístico de Antioquia del año 2019. Esta publicación está disponible en el sitio web de la Gobernación de Antioquia y es recopilado por el Departamento Administrativo de Planeación (DAP) del departamento y su objetivo es difundir datos estadísticos de Antioquia de diversa índole, entre ellos datos geográficos, demográficos, socioeconómicos, administrativos, ciencia y tecnología entre otros. Es importante anotar que esta encuesta fue hecha para la totalidad de los hogares de los municipios.

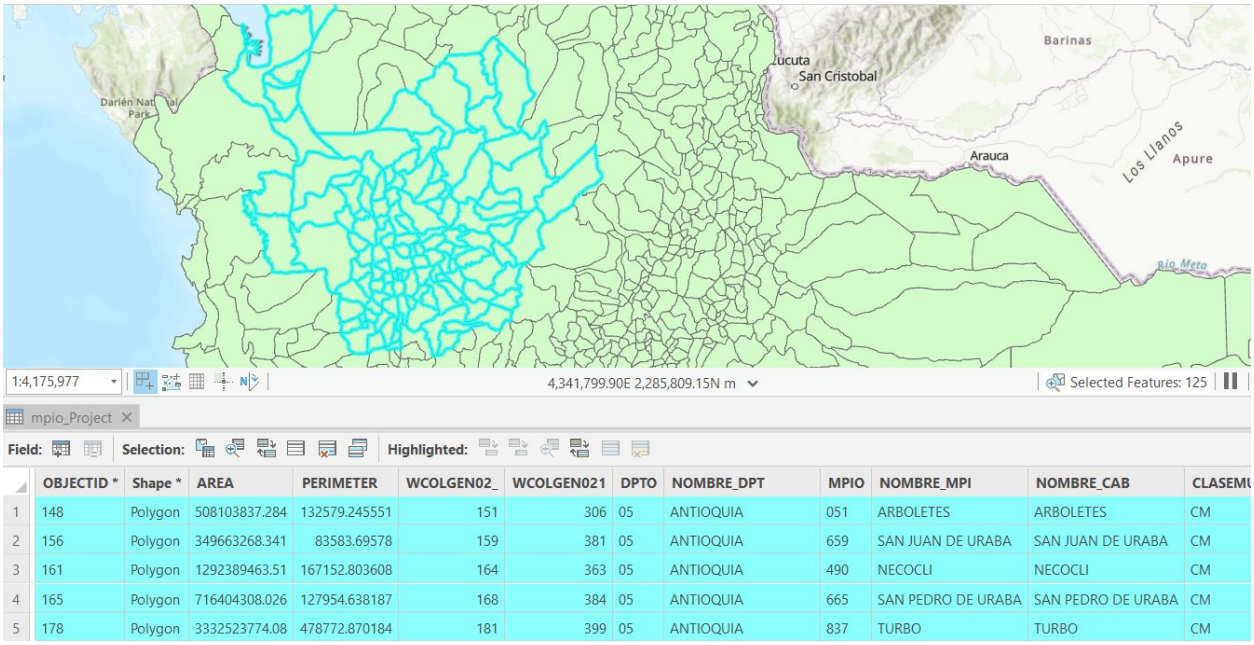
Los datos fueron descargados en archivos en formato Excel. Se descargaron siete (7) archivos, cada uno referente a un servicio público, a saber: agua, alcantarillado, acueducto, electricidad, gas, recolección de basuras e internet. La Figura 8 muestra un ejemplo del formato de los archivos descargados.

Código DANE del Municipio	Municipios y Subregiones	Viviendas ocupadas estimmadas 2019			Viviendas con acceso al servicio de Agua potable			Cobertura de Agua potable%		
		Total	Cabecera	Resto	Total	Cabecera	Resto	Total	Cabecera	Resto
05	**Total Departamento**	2,096,981	1,588,226	508,755	1,751,679	1,546,951	204,728	**83.53**	**97.40**	**40.24**
05	**Resto del Departamento (1)**	786,299	431,860	354,439	500,234	408,136	92,098	**63.62**	**94.51**	**25.98**
SR01	**Valle de Aburrá**	1,310,682	1,156,366	154,316	1,251,445	1,138,816	112,630	**95.48**	**98.48**	**72.99**
05001	Medellín	830,515	729,824	100,691	817,740	720,774	96,965	98.46	98.76	96.30
05079	Barbosa	18,782	7,661	11,121	8,087	7,318	769	43.06	95.52	6.92
05088	Bello	171,542	164,539	7,003	159,564	157,283	2,281	93.02	95.59	32.57
05129	Caldas	25,593	20,556	5,037	22,141	20,437	1,705	86.51	99.42	33.84
05212	Copacabana	25,392	20,603	4,789	24,343	20,426	3,917	95.87	99.14	81.80
05266	Envigado	80,707	77,245	3,462	78,964	77,152	1,812	97.84	99.88	52.34
05308	Girardota	16,154	9,153	7,001	9,606	9,131	475	59.47	99.76	6.79
05360	Itagüí	88,259	80,191	8,068	82,103	79,854	2,248	93.02	99.58	27.87
05380	La Estrella	23,291	19,732	3,559	19,863	19,614	250	85.28	99.40	7.01
05631	Sabaneta	30,447	26,862	3,585	29,033	26,827	2,206	95.36	99.87	61.54
SR02	**Bajo Cauca**	71,952	45,167	26,785	40,558	40,558	0	**56.37**	**89.80**	**0.00**

Figura 8. Visualización formato archivos con la información de los servicios públicos fuente: Anuario Estadístico de Antioquia, 2019

En cada uno de los archivos, se tiene un recuento del total de viviendas del municipio, un recuento de las que cuentan con el servicio y la cobertura se calcula como un porcentaje donde 100% indica que todas las viviendas cuentan con el servicio. Para este estudio estos siete archivos se unificarán dejando las columnas comunes de código de municipio, nombre de municipio, total viviendas y los porcentajes de cada servicio público: alcantarillado, acueducto, electricidad, gas, recolección de basuras e internet y se exportará como un archivo csv.

Por último, se descargó un archivo *.zip desde la plataforma del Instituto Geográfico Agustín Codazzi con la información de los polígonos que contienen al área de los municipios colombianos. El archivo zip contiene cinco archivos, los cuales conforman la información desplegada en un software de sistemas de información geográfica al elegir un archivo denominado shapefile con formato shp. En la Figura 9 se observan los datos visibles tanto en una tabla como en un mapa, en este caso con una selección de los municipios antioqueños. Esta selección se exportó y se generó la capa con la geometría de los municipios de Antioquia.



	OBJECTID *	Shape *	AREA	PERIMETER	WCOLGEN02_	WCOLGEN021	DPTO	NOMBRE_DPT	MPIO	NOMBRE_MPI	NOMBRE_CAB	CLASEMU
1	148	Polygon	508103837.284	132579.245551	151	306	05	ANTIOQUIA	051	ARBOLETES	ARBOLETES	CM
2	156	Polygon	349663268.341	83583.69578	159	381	05	ANTIOQUIA	659	SAN JUAN DE URABA	SAN JUAN DE URABA	CM
3	161	Polygon	1292389463.51	167152.803608	164	363	05	ANTIOQUIA	490	NECOCLI	NECOCLI	CM
4	165	Polygon	716404308.026	127954.638187	168	384	05	ANTIOQUIA	665	SAN PEDRO DE URABA	SAN PEDRO DE URABA	CM
5	178	Polygon	3332523774.08	478772.870184	181	399	05	ANTIOQUIA	837	TURBO	TURBO	CM

Figura 9. Visualización en ArcGIS Pro de archivo shp con la información de la localización de los municipios de Antioquia.

6.3 Variables

Se han escogido las siguientes variables para analizar cómo la cobertura de los servicios públicos afecta al resultado promedio de las pruebas Saber 11 en los municipios de Antioquia. La variable dependiente es el promedio del puntaje global de las pruebas por municipio y las variables explicativas a explorar son los servicios públicos, que se agregarán en una sola variable mediante la suma de la cobertura de electricidad, alcantarillado, acueducto, gas, recolección de basuras e internet. Si algún municipio no tiene cobertura de ninguno de estos servicios el valor de la variable será cero, mientras si un municipio tiene cobertura de 100% en todos estos servicios, el valor de la variable será de 600. Esta suma supone que los servicios públicos pueden complementarse o sustituirse en cuanto a la mitigación de riesgos que afecten los procesos cognitivos de los estudiantes, asociados con el clima y la salubridad.

La integración de los datos de las diversas fuentes de información se facilita debido al código identificador de los municipios según el DANE, el cual es un estándar nacional.

La información de la variable del número de estudiantes que presentaron la prueba por municipio en el 2019 se utilizará para analizar si alcanzan a una muestra adecuada estadísticamente hablando. Mientras que la variable, total de viviendas, que informa sobre las

viviendas servidas por municipio por los servicios públicos se utilizará para dar contexto a los porcentajes de cobertura por municipio. La Tabla 1 muestra el resumen de las variables a utilizar en esta monografía

Tabla 1. Descripción de variables

Variable	Escala	Medida	Origen Archivo	Tipo de archivo
Código Municipio Dane	Nominal		Cobertura Servicios Públicos y Pruebas Saber 11-2 2019, capa geográfica IGAC	Estructurado (csv) y archivo georreferenciado (shp)
Municipios de Colombia (Residencia)	Nominal, georreferenciado		Cobertura Servicios Públicos y Pruebas Saber 11-2 2019, capa geográfica IGAC	Estructurado (csv) y archivo georreferenciado (shp)
Cobertura Electricidad	Cuantitativa Continua	0-100%	Cobertura Servicios Públicos	Estructurado (csv)
Cobertura Acueducto	Cuantitativa Continua	0-100%	Cobertura Servicios Públicos	Estructurado (csv)
Cobertura Alcantarillado	Cuantitativa Continua	0-100%	Cobertura Servicios Públicos	Estructurado (csv)
Cobertura Gas Domiciliario	Cuantitativa Continua	0-100%	Cobertura Servicios Públicos	Estructurado (csv)
Cobertura Recolección de Basuras	Cuantitativa Continua	0-100%	Cobertura Servicios Públicos	Estructurado (csv)
Cobertura Internet en el hogar	Cuantitativa Continua	0-100%	Cobertura Servicios Públicos	Estructurado (csv)
Suma cobertura servicios públicos (SPB)	Cuantitativa Continua	0-600	calculado	Estructurado (csv)
Promedio de puntaje global por municipio (Pglobal)	Cuantitativa discreta	0-500	Pruebas Saber 11-2 2019	Estructurado (csv)
Total de estudiantes por municipio	Cuantitativa discreta		Agregado	Estructurado (csv)
Total de viviendas por municipio	Cuantitativa discreta		Cobertura Servicios Públicos	Estructurado (csv)

Fuente: Elaboración propia

7. Metodología

La metodología general aplicada en este trabajo para el diseño de la arquitectura BI se resume en el esquema ilustrado en la Figura 10.

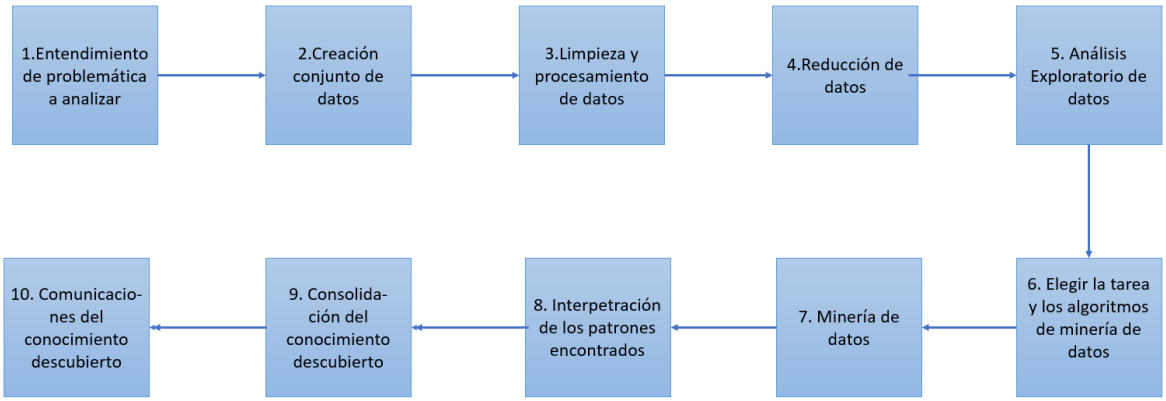


Figura 10. Resumen de la metodología seguida en esta monografía. Modificado de García et al (2018)

El gobierno colombiano ha puesto a disposición datos abiertos que pueden ser utilizados para la investigación y el desarrollo del conocimiento a partir de datos de la nación. Es por esto por lo que se procede a realizar una investigación de tipo aplicada, en donde se tomaran los conceptos existentes para realizar un modelo que integre datos educativos, de prestación de servicios públicos y de condiciones geográficas y de esta manera ejemplificar la aplicación de un diseño de arquitectura BI a un caso práctico.

Así, también, es una investigación de tipo cuantitativo en donde se persigue establecer la relación entre el nivel de cobertura de los servicios básicos en los municipios antioqueños y el desempeño en las pruebas Saber 11 de los estudiantes de esos municipios mediante los datos, herramientas y algoritmos disponibles para este fin en el marco de la arquitectura BI diseñada. La muestra de esta investigación se acotará en el tiempo (año 2019) y en espacio (municipios del departamento de Antioquia), por lo que este estudio es de carácter transversal. Es una investigación de tipo inductivo, ya que se utilizará un conjunto de datos para un departamento particular, para concluir acerca de los aspectos a considerar a la hora de implementar una arquitectura de BI en el sector público de Colombia.

La recopilación de los datos estuvo a cargo de entidades estatales bajo normas que aseguran su calidad e integridad, como lo son el Icfes, el DAP y el IGAC. En resumen, se tienen 125 municipios de Antioquia, con datos referentes a porcentajes de la cobertura de servicios públicos domiciliarios, al puntaje global obtenido en las pruebas Saber 11-2019 agregados como promedios por municipio y a su estructura espacial.

La integración de datos abiertos en Colombia en un sistema de inteligencia de negocios referentes a entidades territoriales se logra debido a que los departamentos y municipios tienen un código único establecido por el DANE. Los conjuntos de datos a utilizar tienen diferentes granularidades, sin embargo, para efectos de este trabajo el diseño de las tablas incluirá los datos agregados a nivel municipal. Por otro lado, los archivos *shapefile (shp)* publicados por el IGAC*,* con la estructura espacial de los datos también hacen referencia a los códigos del DANE, lo que permitirá integrar los datos con los archivos csv y aplicar estadística espacial al análisis de los datos.

7.1 Diseño arquitectura BI

Se diseña una arquitectura BI (ver Figura 11) en la que el contexto es la base para orientar el desarrollo de esta. Así, la capa de datos está constituida por archivos con formato csv con los datos de las pruebas Saber 11 del segundo periodo de 2019, el porcentaje de cobertura de los servicios públicos domiciliarios en los municipios de Antioquia y el archivo con formato shp con la estructura espacial. Los archivos csv pasan al proceso de ETL, mientras que el archivo shp va directamente a la capa de almacenamiento y contiene la geometría y localización de los entes territoriales y es el centro de una arquitectura BI que contemple la estructura espacial de los datos.

La capa de ETL se realiza en Excel con *Power Query* (Office 365), una herramienta que cuenta con una interfaz gráfica de usuario que permite realizar una serie de operaciones de transformación de datos, como limpiar y filtrar datos, fusionar y combinar tablas, agregar columnas calculadas, eliminar duplicados entre otros. Con esta herramienta se integraron seis (6) archivos con la información de los servicios públicos de los 125 municipios de Antioquia. Se eligieron las columnas relevantes a cargar a la geodatabase, como lo son región, municipio, código DANE del municipio, total de viviendas, porcentajes de cobertura de alcantarillado, acueducto, electricidad, gas, recolección de basuras e internet. Por otro lado, del archivo con los

datos del Icfes se filtraron los datos para el departamento de Antioquia y los resultados que vienen por estudiante, se agregaron como promedios por municipio. Igualmente se eligieron las columnas a cargar, las cuales son municipio, código DANE del municipio, número de estudiantes por municipio y puntaje global promedio por municipio.

Como transformación importante, se tiene la suma de los servicios básicos públicos domiciliarios en una nueva variable SPB, cuya escala varia de 0 a 600. Esta transformación se realizó ya que los beneficios de los servicios públicos básicos domiciliarios, no solo se complementan, sino que también pueden llegar a sustituirse, como por ejemplo el gas y la electricidad para la regulación de la temperatura ambiente, o la presencia de estos dos mitigan la falta de agua potable, al poder hervirse o filtrarse con electrodomésticos o gasodomésticos.

De esta capa saldrá un único archivo csv, con los siguientes campos: código región, código municipio, nombre región, nombre municipio, total de estudiantes, total viviendas, porcentajes de electricidad, alcantarillado, acueducto, gas, recolección de basuras, Internet, SPB (Suma de las coberturas de los servicios públicos), y Pglobal (promedio del Puntaje global pruebas Saber 11-2 2019). Es importante señalar que el puntaje global es el promedio de los resultados de los estudiantes residentes del mismo municipio sin tener en cuenta sí presentaron las pruebas en el municipio en donde residen o no.

La capa de almacenamiento se genera en una geodatabase del programa ArcGIS Pro, la cual es una base de datos de objetos geográficos que permite la gestión, edición, análisis y publicación de información geográfica en diferentes formatos y tipos de datos. El shp del IGAC con la información geográfica de los municipios de Antioquia se carga directamente a la geodatabase, después de descargarse de https://www.colombiaenmapas.gov.co/, mientras que el archivo con formato csv resultante del proceso ETL de la capa anterior se importa como una tabla al sistema. La geodatabase tiene varias herramientas para asociar las tablas con el shapefile para generar nuevos shapefiles enriquecidos con los datos de los servicios públicos, del Icfes y de cualquier otra fuente que pueda relacionarse a la nomenclatura del DANE de los municipios.

La capa de usuario final está conformada por otro sistema de información geográfica denominado Geoda, el cual cuenta con herramientas para el análisis geoespacial de los datos, tales como la regresión espacial autorregresiva (SAR) y la agrupación. Este sistema de

información geográfica también se utilizará para la exploración de datos (EDA) en conjunto con Tableau. Para la visualización se generarán mapas en Tableau y Geoda, así como histogramas, diagramas de caja, diagramas de líneas y estadística descriptiva. Con base en los hallazgos se diseñarán reportes y recomendaciones que orienten a las personas en el diseño y aplicación de políticas públicas.

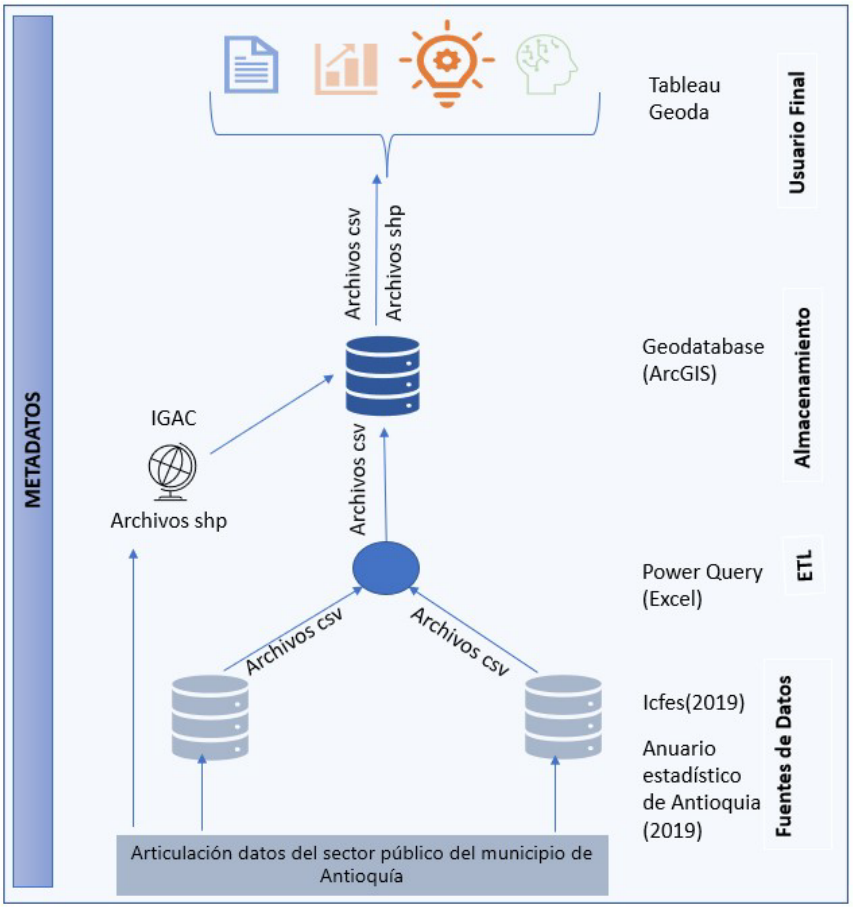


Figura 11. Arquitectura BI general para este proyecto.

Elaboración propia

7.2 Análisis exploratorio de datos

El análisis exploratorio de datos (EDA, por sus siglas en inglés) es una técnica utilizada para analizar y entender los datos antes de aplicar técnicas de modelado. El objetivo del EDA es

resumir y visualizar los datos, identificar patrones y relaciones, detectar valores atípicos y errores en los datos, y formular hipótesis que puedan ser probadas posteriormente. También permite evaluar la calidad de los datos y conceptualizar el modelo a realizar.

Este análisis se realiza mediante la aplicación de técnicas estadísticas y gráficas para explorar los datos, tales como histogramas, diagramas de caja, diagramas de dispersión, gráficos de densidad, matrices de correlación y mapas coropléticos, entre otros. También incluye la realización de análisis descriptivos, como la media, la mediana, la desviación estándar y el rango Inter cuartil, para resumir las características estadísticas de los datos. En el presente trabajo se utilizará Tableau Public 2023.1 para explorar los datos, ya que este tiene acceso directo a la geodatabase de ArcGIS Pro.

7.3 Modelos de auto regresión espacial simultanea (SAR)

Para este paso se utiliza la metodología explicada por Anselín (2005) en su manual del programa Geoda. Se inicia conectando Geoda a la geodatabase de ArcGIS directamente. Una vez elegido el archivo con los datos a analizar, se procede a calcular la matriz de pesos, que en este caso se basó en el número de vecinos y diseño de reina (los vecinos son buscados en todas las direcciones).

Una vez, se tiene la matriz de pesos, se procede a realizar la regresión lineal con diagnóstico para evaluar si existe autocorrelación, utilizando el diagnóstico del Multiplicador de Lagrange (LM) y el índice de Moran. Así mismo, de ser necesario, se procederá a aplicar la correlación espacial autorregresiva SAR. La Figura 11 muestra el proceso para decidir si se debe aplicar la regresión espacial SAR.

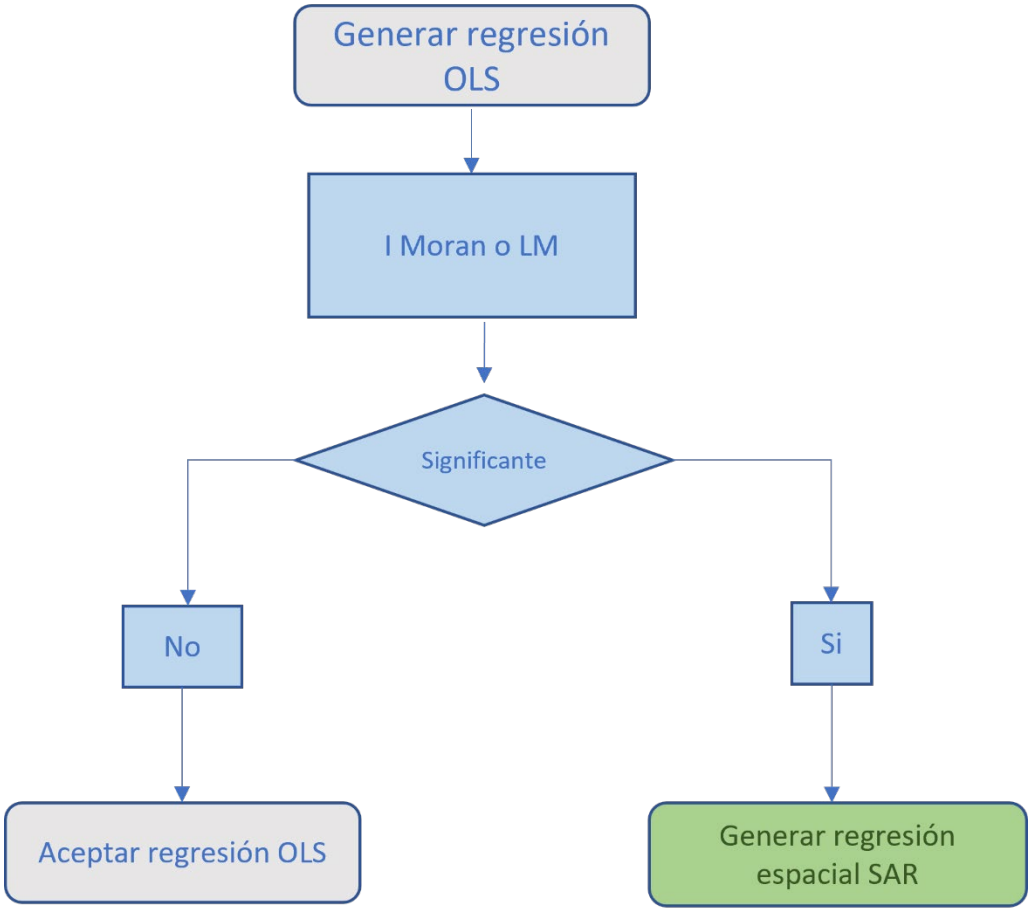


Figura 12. Diagrama de decisión para decidir generar un modelo de autorregresión espacial SAR en Geoda. Fuente: Modificado de Anselin, 2005.

En la Figura 12 se observa que en el flujo de trabajo se da inicio con una regresión lineal simple o multivariada ajustada por el método de mínimos cuadrados (OLS por sus siglas en inglés) con diagnóstico. Este diagnóstico es arrojado por Geoda y se presenta con la evaluación de la significancia de los valores p para el índice de Moran y LM. Así, si p presenta un valor inferior a 0.01 en ambos casos, indica que hay rechazo de la hipótesis nula (no existe autocorrelación espacial) y se elige la alternativa (existe autocorrelación espacial).

Si se encuentra que la hipótesis nula se rechaza, se procede a generar el modelo SAR. Este modelo en Geoda se ajusta por el método de máxima verosimilitud (ML por sus siglas en inglés), que es un método no paramétrico que busca que los resultados en los puntos conocidos sean lo más parecido a los datos originales (James et al, 2017).

7.4 Visualización de resultados

Los resultados se reportarán con el apoyo de mapas, gráficos y tablas que resuman los hallazgos encontrados mediante el análisis de los datos y la analítica geoespacial. Estas visualizaciones se generarán en Geoda y Tableau.

8. Resultados

Los datos descargados para la realización de esta monografía son publicados bajo estándares de calidad y rigor estadístico por ser provenientes de instituciones gubernamentales. No obstante, los datos se revisaron para asegurar que no existiesen datos erróneos, faltantes o en el formato incorrecto. Una vez los datos fueron revisados, se procedió a integrarlos, después de filtrar y agregar los datos como se describió en la metodología, en una geodatabase de ArcGIS Pro.

Tableau Publico y Geodas tienen conectividad a los archivos típicos de una geodatabase (con extensión *.gdb), desde donde se procedió a analizarlos, empezando con la exploración de datos y hasta los modelos de analítica (autorregresión espacial simultanea). Es importante anotar que, dentro de la geodatabase, se encuentran los archivos formato *.shp y *.csv descritos en la metodología.

8.1 Arquitectura de BI

En la Figura 13 se muestra el esquema de la arquitectura resultante aplicada para el desarrollo de la metodología planteada en esta monografía. Es de resaltar que en el esquema de la arquitectura se incluye el factor humano, en donde el analista de datos es quien entiende la información que alimentará los modelos aplicados en el proceso de analítica de datos. También, en la misma capa de usuario final, se encuentra el receptor de los resultados del análisis y quien con su experiencia aplicará los hallazgos en el diseño de las políticas públicas. Cabe anotar que los metadatos están imbuidos en cada capa, al describir el contexto, la procedencia de los datos, sus formatos y año, las operaciones realizadas al cargar los datos a la geodatabase. También se registra la versión de los programas utilizados, así como las técnicas utilizadas para explorar y modelar los datos.

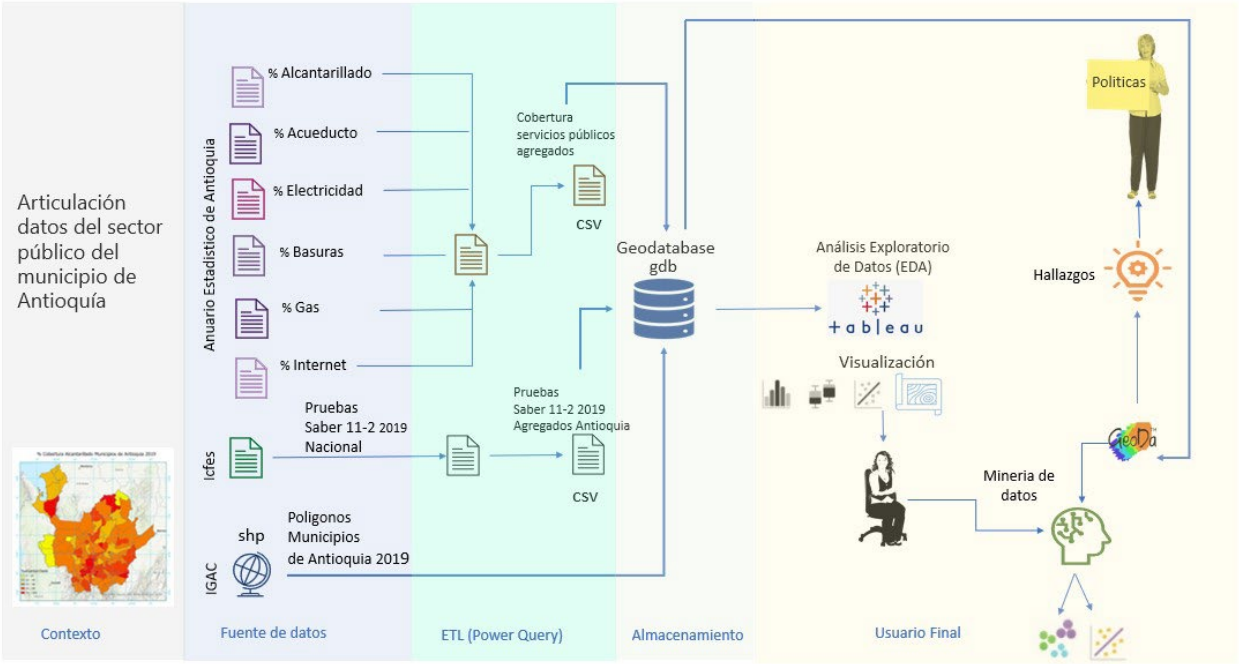


Figura 13. Arquitectura BI aplicada en esta monografía.

Elaboración propia

8.2 Exploración de Datos (EDA)

Al realizar la EDA con Tableau se encontraron patrones útiles a la hora de plantear los posibles modelos explicativos a partir de las variables analizadas de forma interactiva. En la Figura 14, se puede observar la distribución espacial de las variables utilizadas en este proyecto. También se aprecian las subregiones en que se divide el departamento antioqueño, sus municipios y los filtros usados a la hora de explorar los datos. Aunque los datos no muestran una distribución normal perfecta, son aproximadamente normales. En el gráfico de dispersión se sugiere una relación lineal positiva, en donde los puntajes globales aumentan al aumentar la cobertura de los servicios públicos. También se observa que en general las variables decrecen desde el Valle de Aburrá hacia las periferias del departamento. Sin embargo, en la Figura 15, entrando en detalle, se observa que en todas las subregiones existe variabilidad en cuanto a la cobertura de los servicios públicos y los puntajes globales obtenidos en las pruebas Saber 11-2 2019.

Una de las ventajas de utilizar Tableau para realizar la EDA es que se pueden combinar información utilizando texto, colores, líneas, histogramas para resumir información de manera coherente con los objetivos de la investigación. Por ejemplo, en la Figura 16 se observa la distribución geográfica de las variables, los histogramas, los diagramas de caja y la estadística descriptiva en tablas. Por medio de filtros se puede analizar también como cambian estas distribuciones, permitiendo plantear estrategias para entender los datos en relación con el contexto de la arquitectura.

Entre tanto en la Figura 17 se observan dos líneas que se desplazan a lo largo del eje horizontal, que despliega de manera descendente el número de total de viviendas por municipio. La línea de color naranja representa la tendencia de los puntajes globales por municipio, mientras que la amarilla la cobertura general de los servicios básicos públicos domiciliarios. En esta gráfica, se logra observar que existe una correlación en el patrón, es decir en general, para los municipios en donde los servicios públicos suben, los puntajes globales también lo hacen y viceversa. Esta correlación se hace más evidente desde aproximadamente un total de viviendas mayor a 3600.

Después de realizar estas observaciones se procede a graficar los diagramas de dispersión (ver Figuras 18-20) entre la variable dependiente (Pglobal) y la independiente (SPB), primero para todos los municipios, luego para los municipios en donde más de 30 estudiantes presentaron la prueba (debido a representar el mínimo de muestras aceptable para un análisis estadístico) y una última con los municipios con más de 3600 viviendas para explorar el patrón observado en la Figura 17 se observa un coeficiente de correlación de Pearson positivo y significativo (p valor < 0.01) lo cual indica una relación lineal directa entre las variables objeto de análisis. Sin embargo, al evaluar los supuestos del modelo lineal, de manera particular, se encuentra que la prueba de independencia (Durbin-Watson) revela que no hay independencia en los residuales (p valor > 0.01). Lo anterior sugiere la posibilidad de una autocorrelación espacial entre las variables objeto de análisis.

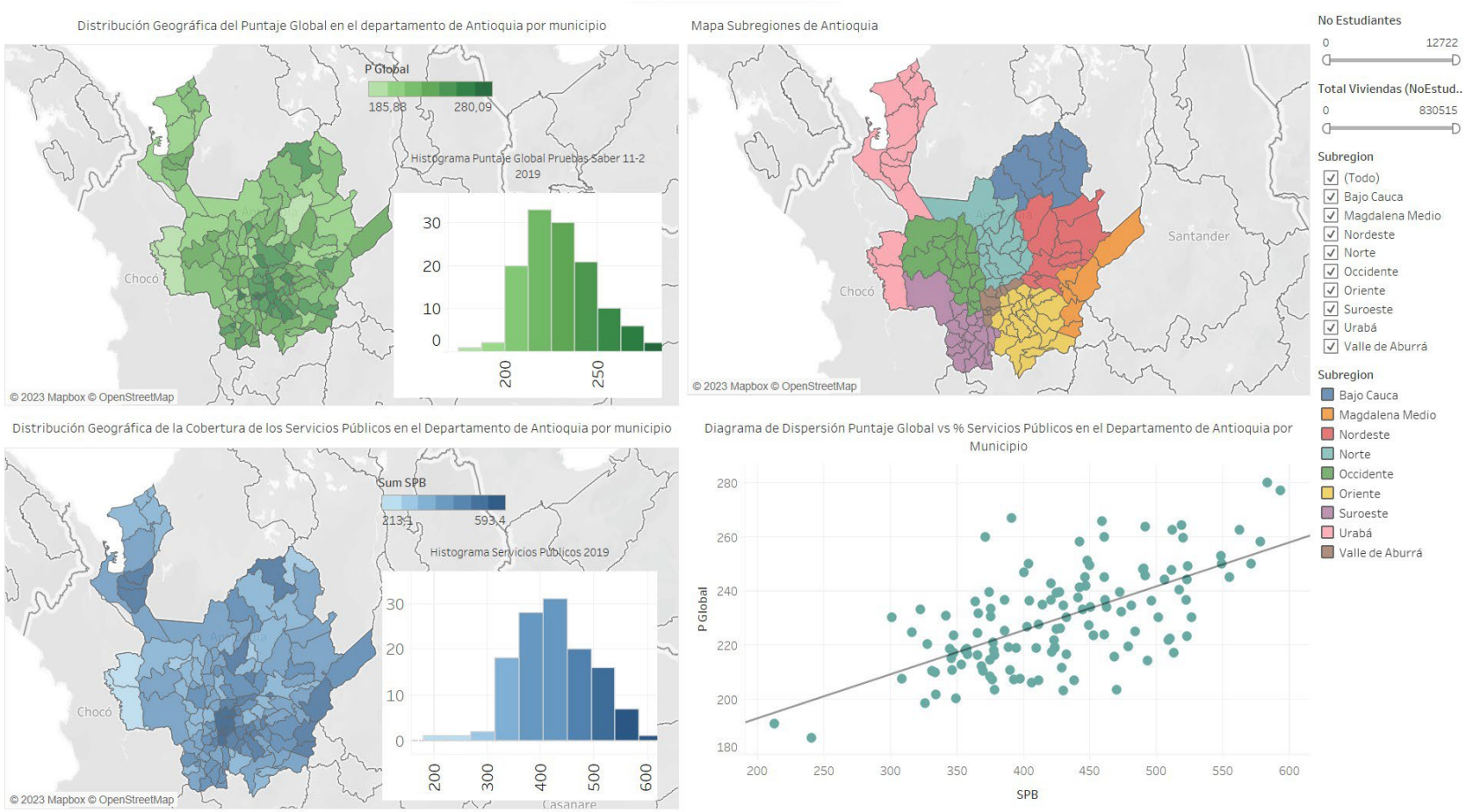


Figura 14. Distribución espacial, histograma y gráfico de dispersión en un cuadro de mando con filtros para su exploración.

Elaboración propia. Ver Dashboard en https://public.tableau.com/views/EDA_Datos_Tesis_Antioquia/Dashboard52

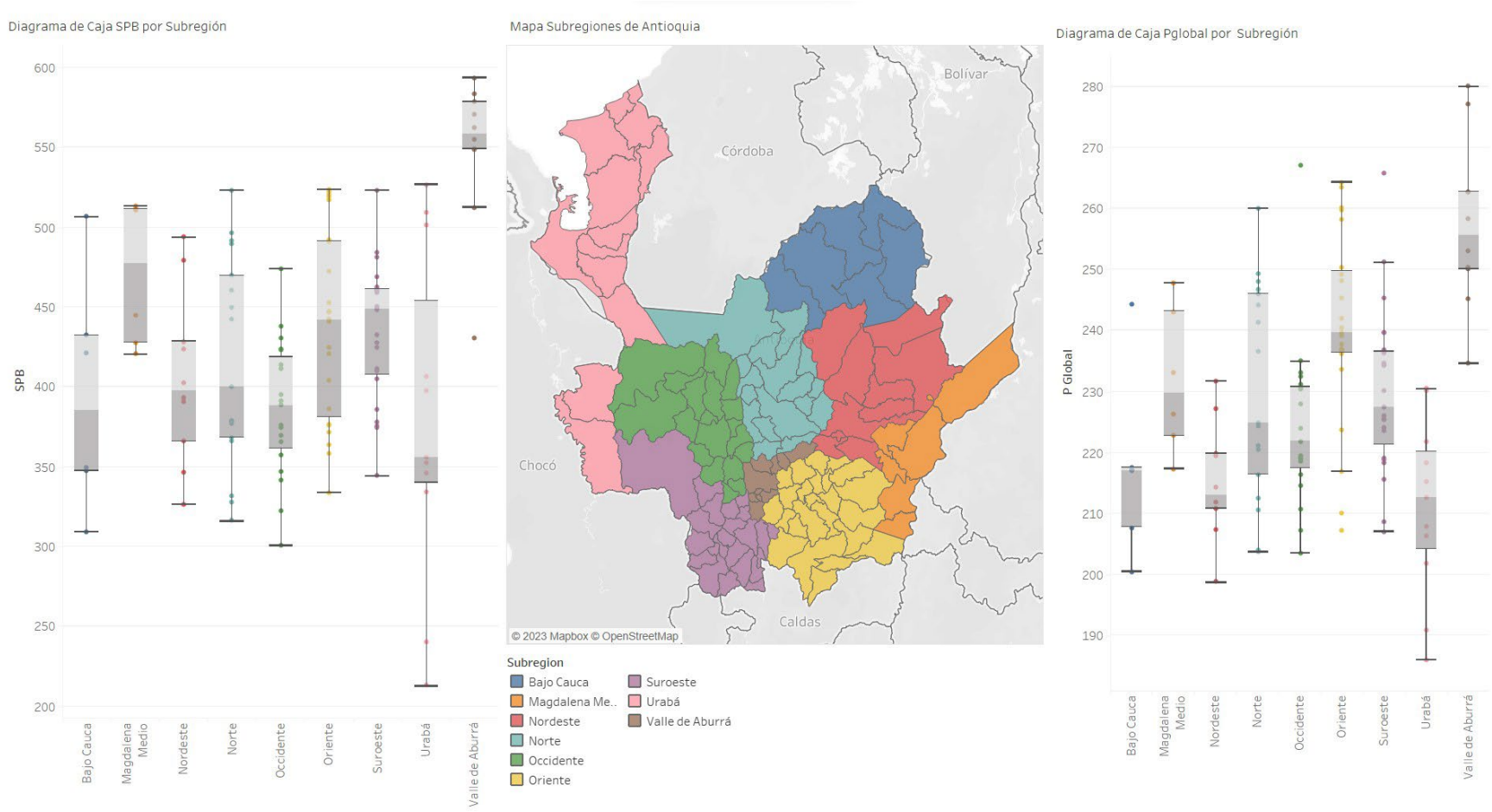


Figura 15. Diagramas de Caja mostrando la distribución de valores por región del puntaje global y los servicios públicos domiciliarios.

Elaboración propia

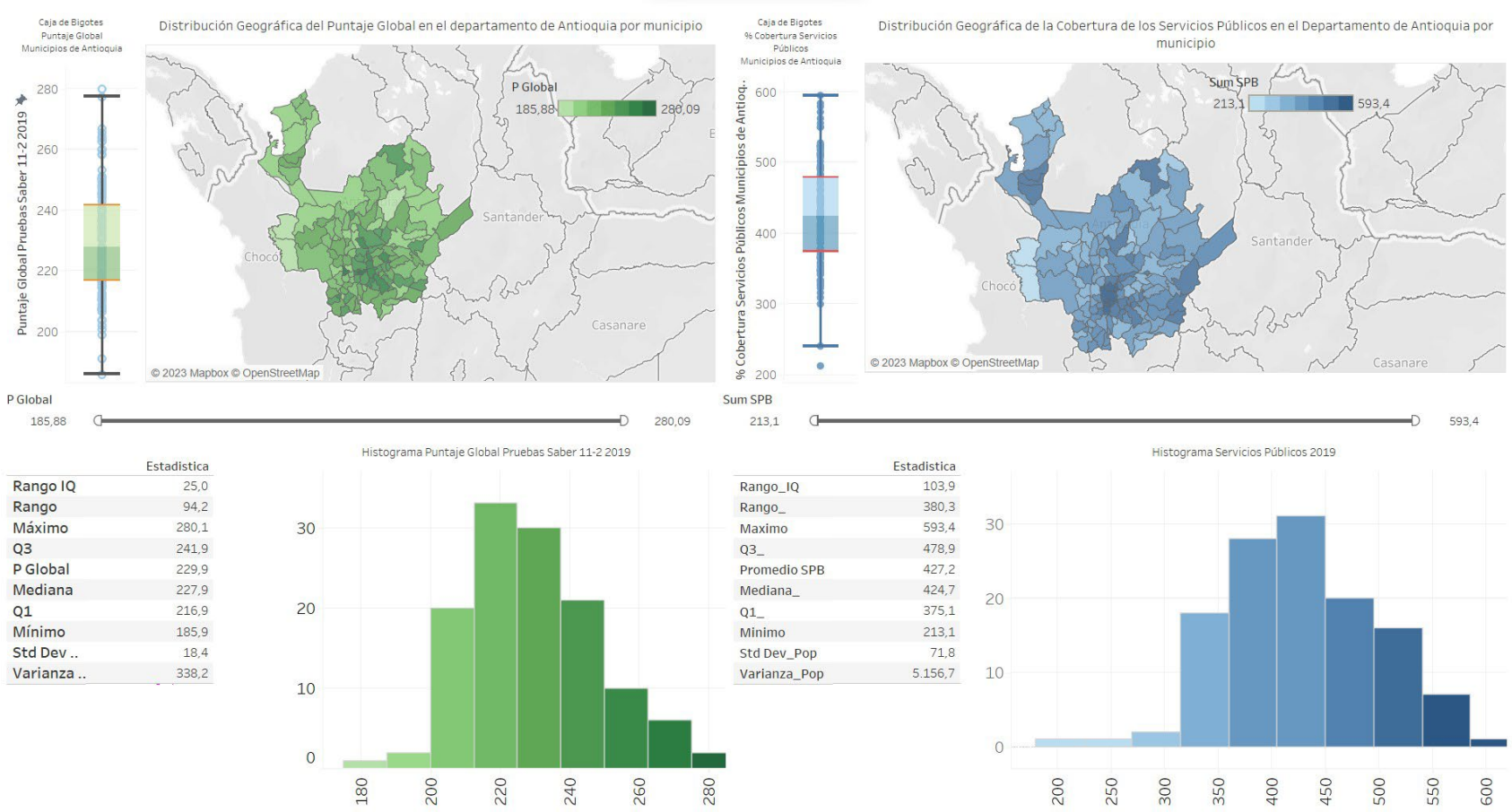


Figura 16. Estadística descriptiva utilizando Tableau Public.

Elaboración propia

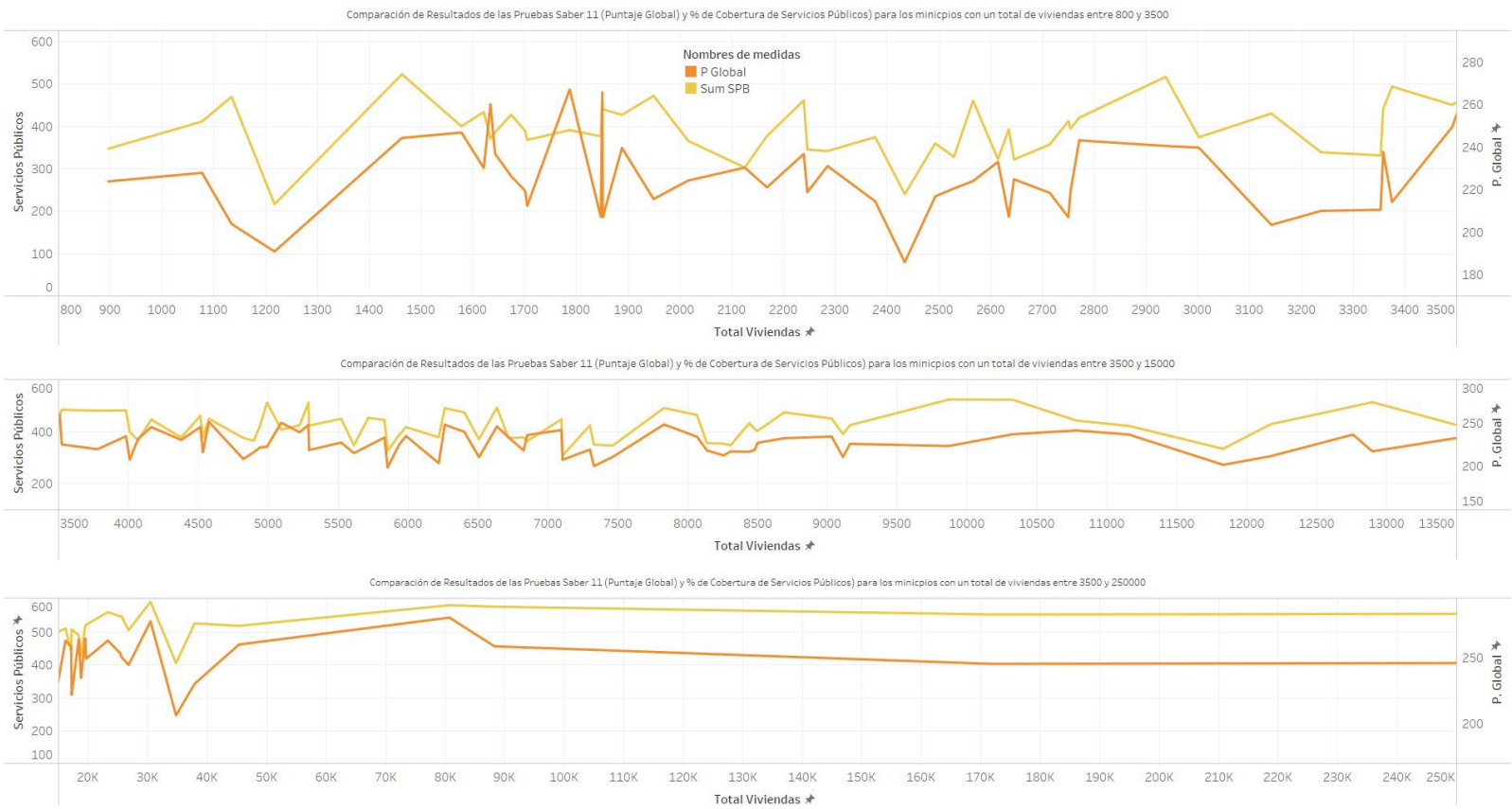


Figura 17. Comparación de tendencias de los puntajes globales y suma de servicios públicos domiciliarios (SPB).

Elaboración propia

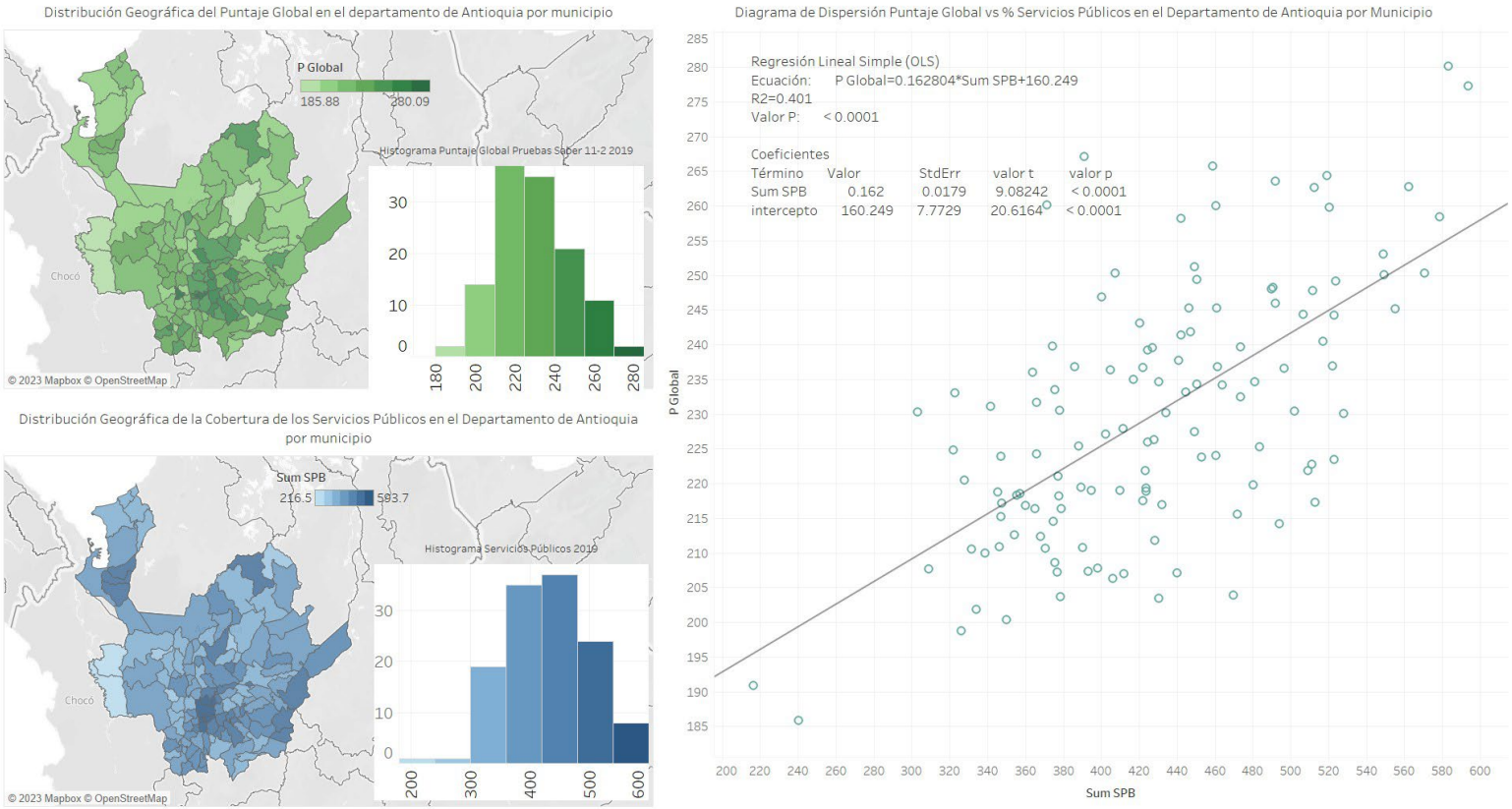


Figura 18. Visualización de diagrama de dispersión para todos los municipios de Antioquia.

Elaboración propia

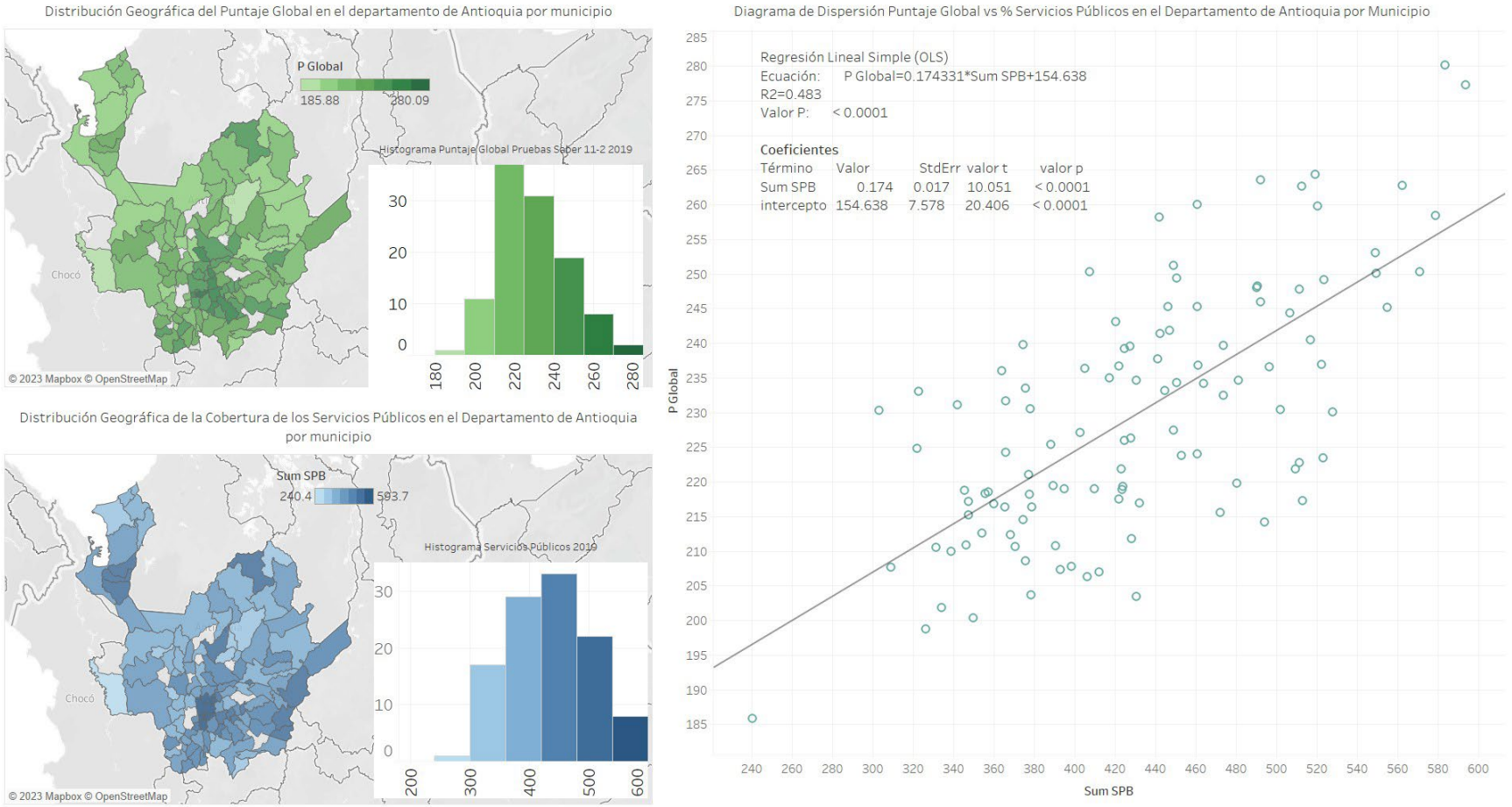


Figura 19. Visualización de diagrama de dispersión para todos los municipios de Antioquia en donde se presentaron más de 30 estudiantes a presentar las pruebas Saber 11-2 2019.

Elaboración propia

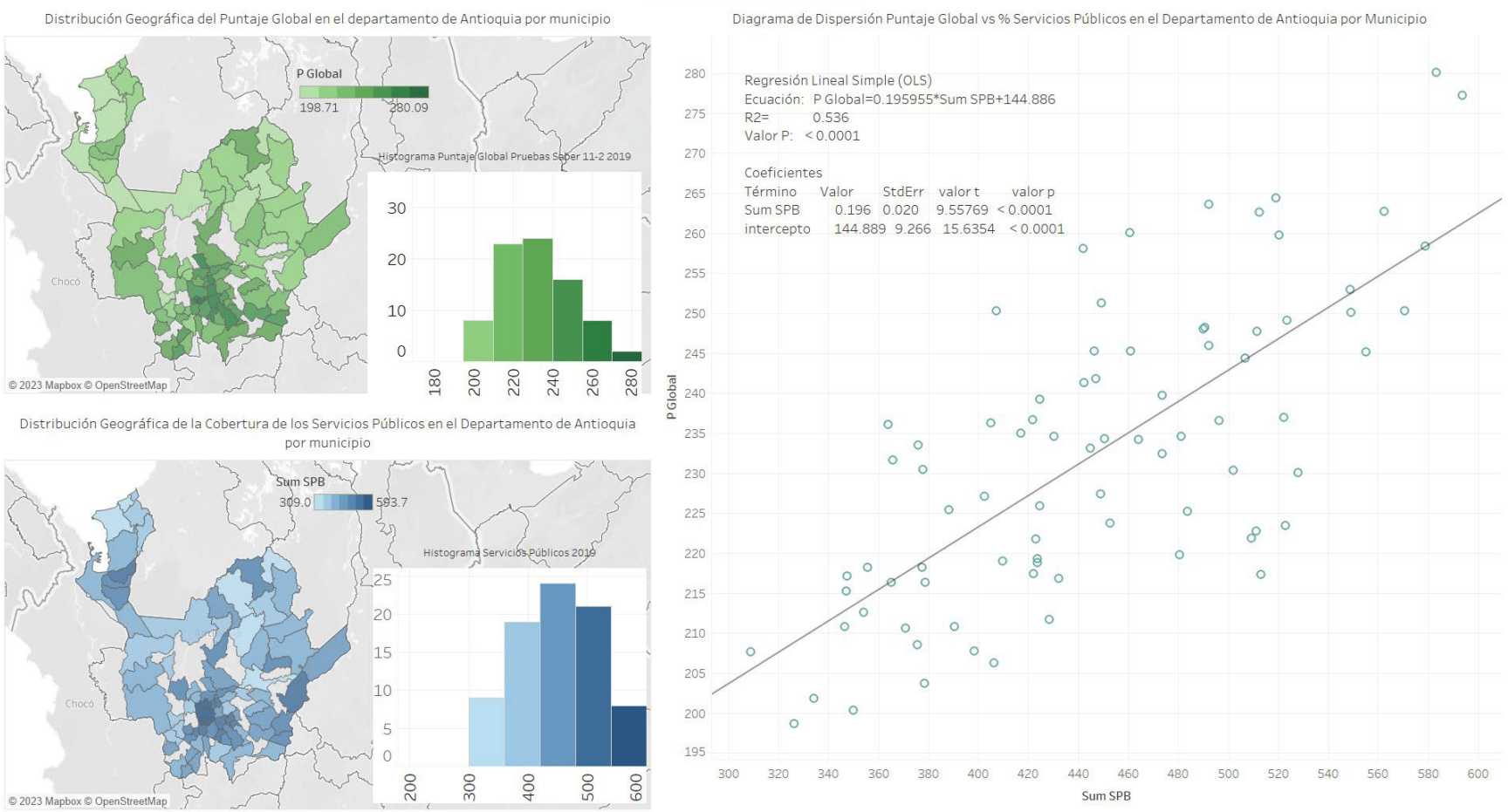


Figura 20. Visualización de diagrama de dispersión para todos los municipios de Antioquia en donde se hay más de 3600 viviendas servidas por los servicios públicos.

Elaboración propia.

8.3 Analítica Geoespacial

Al analizar los resultados y las gráficas en la EDA, se decide seguir la metodología diseñada por Anselin (2005) para realizar un análisis de regresión espacial y así dar respuesta al problema planteado en el contexto de la arquitectura BI. Primero, se generó el modelo de regresión lineal OLS con diagnósticos del programa Geoda, en donde la variable dependiente es el promedio del puntaje global de las pruebas Saber 11-2 2019 por municipio (Pglobal) y la variable independiente es la suma de la cobertura de los servicios públicos (SPB) por municipio, para los tres escenarios analizados: todos los municipios (1), municipios que tenían más de 3600 viviendas en el 2019 (2) y los municipios en donde más de 30 estudiantes presentaron las pruebas Saber 11-2 2019 (3). Como se observa en la tabla 2, en los tres casos se debe generar el modelo de autorregresión espacial simultanea por rezago. Los resultados indican que hay una autocorrelación espacial entre las variables objeto de análisis.

Tabla 2. Tabla comparativa resultados de variables de diagnóstico usando metodología Anselin (2005) con el programa Geoda

Prueba	Modelo 1 (todos los municipios)			Modelo 2 (Más de 3600 viviendas)			Modelo 3 (Más de 30 estudiantes)		
	MI/DF	Valor	p	MI/DF	Valor	p	MI/DF	Valor	p
Moran's I (error)	0,2366	4,4612	0,00001	0,4205	5,6363	0,00000	0,3018	5,1088	0,00000
LM (rezago)	1	26,1901	0,00000	1	41,9763	0,00000	1	32,7940	0,00000
LM Robusto (rezago)	1	9,2053	0,00241	1	14,9501	0,00011	1	10,5497	0,00116
LM(error)	1	17,0572	0,00004	1	27,1944	0,00000	1	22,4587	0,00000
LM (SARMA)	2	26,2626	0,00000	2	42,1445	0,00000	2	33,0084	0,00000

A continuación, en las tablas 3 y 4, se presentan los resultados de los tres modelos, después de aplicar la corrección por rezago ML y su comparación con el modelo OLS:

Tabla 3. Tabla comparativa de parámetros resultantes de modelo OLS

Variable	Modelo 1				Modelo 2				Modelo 3			
	Coef	Error Std	t	p	Coef	Error Std	t	p	Coef	Error Std	t	p
Constante	160.529	7.737	20.74	0,00000	144.886	9.267	15.90	0,00000	154.995	7.554	20.52	0,00000
SPB	0.1623	0,0179	9.09	0,00000	0,130498	0,021	9.56	0,00000	0.174	0,017	10.04	0,00000
r pearson	0.63				0.74				0.69			
R^2 Ajustada	0.39				0.53				0.48			

Tabla 4 comparativa de parámetros resultantes de modelo SAR por rezago

Variable	Modelo 1				Modelo 2				Modelo 3			
	Coef	Error Std	z	p	Coef	Error Std	z	p	Coef	Error Std	z	p
Rho	0,4284	0,092	4,66	0,00000	0,5310	0,0801	6,63	0,00000	0,4749	0,084	5,68	0,00000
Constante	77,9983	19,026	4,10	0,00004	50,8594	15,8914	3,20	0,00137	65,3344	16,852	3,88	0,00011
SPB	0,1248	0,0179	6,98	0,00000	0,1305	0,0182	7,15	0,00000	0,12845	0,017	7,58	0,00000
Pseudo R^2	0.51				0.73				0.62			

Se observa que los resultados confirman que existe una correlación entre la variable dependiente, el promedio del puntaje global de las pruebas Saber 11-2 y la independiente, en este caso el nivel de cobertura de los servicios públicos, como se sugiere en el EDA en las figuras 17-20. Cabe anotar que las figuras 18-20 no contemplan la autocorrelación espacial, pero si mostraron la probable existencia de una correlación positiva entre estas dos variables.

9. Discusión

Los resultados señalan que el diseño de una arquitectura de inteligencia de negocios para apoyar la toma de decisiones en el sector público de Colombia utilizando datos abiertos asociados a entes territoriales debe disponer de datos con la geometría y localización de los mismos (latitud y longitud). Esta característica no suele mencionarse en descripciones de arquitecturas de BI más generales, ya que no es necesaria en casos donde no se involucran datos de áreas. Es por esta razón que el diseño propuesto en esta monografía utiliza una geodatabase en donde se logra integrar archivos shp, que contienen esta información, con datos estructurados que se cargan a la geodatabase como archivos csv y que contienen campos que se pueden enlazar a los entes territoriales. La clave para integrarlos es la nomenclatura que el DANE ha dado a los diferentes municipios.

En la elaboración del diseño y su ejemplificación se logró evidenciar el poder que tiene la visualización de los datos integrados, ya que se pueden encontrar patrones y relaciones no

evidentes en los datos cuando se analizan por separado. Por ejemplo, se encontró una correlación positiva (Pseudo-R^2 > 0.5) y significativa (p < 0.01), es decir, que un aumento en la cobertura de los servicios públicos deriva en un aumento de los promedios de los puntajes globales de las pruebas Saber 11 por municipios. En este escenario la hipótesis nula establece que no hay correlación entre las variables, mientras que la hipótesis alternativa plantea que hay correlación entre ellas. Adicionalmente, incorporar representaciones espaciales en una arquitectura de BI favorece identificar patrones en las variables objeto de análisis. Por ejemplo, las visualizaciones proporcionadas a partir de la arquitectura de BI muestran como los puntajes altos en las pruebas Saber 11-2 y las mayores coberturas de los servicios públicos se concentran en su mayoría alrededor del Valle de Aburrá, mientras que los puntajes más bajos de estas dos variables se observan hacia las periferias del departamento

La arquitectura de BI propuesta solo incorporó variables relacionadas con pruebas educativas y servicios públicos con datos específicos al departamento de Antioquia, no obstante, podría incorporar otras variables. Por ejemplo, se pueden añadir temporalidad u otras variables sociales. Así se podrían incluir variables que representen dimensiones de la salud de la población y niveles de violencia en los municipios e integrarlas al modelo. Por otra parte, se pueden agregar los datos de todos los municipios de Colombia. En este sentido, la arquitectura propuesta se puede extender para considerar no solo otras variables sino otras regiones del país. Es importante recalcar que los datos y análisis deben entenderse de manera contextual y sus resultados no deben aplicarse a nivel individual, ya que la cobertura de los servicios públicos no podrá predecir los resultados individuales en las pruebas Saber 11, esto debido a que hay factores individuales, entre otros que afectan el desempeño escolar (Collazos et al, 2021).

También se debe anotar que la geodatabase utilizada para el desarrollo de esta monografía está alojada localmente. Si se desea tener una geodatabase con los datos descritos anteriormente, que beneficie públicamente a los investigadores en el campo de la educación, debería alojarse en la nube y ofrecerse en el portal de datos abiertos www.datos.gov.co. Es importante que debe ir acompañada de metadatos que contengan detalles de la arquitectura para la cual la geodatabase fue generada.

Finalmente, es importante resaltar que el diseño de una arquitectura BI es útil a la hora de integrar datos de diversas fuentes en el sector público. En esta monografía se logra establecer

que la prestación de los servicios públicos tiene un impacto en el rendimiento en pruebas estandarizadas, tomado como caso de estudio el departamento de Antioquia. Por lo tanto, la cobertura de servicios públicos es una variable que debe ser incluida a la hora de entender los resultados de las políticas públicas en la educación y a la hora de tomar decisiones para mejorarlas, por lo cual se valida la hipótesis planteada en esta monografía. Ahora bien, está ilustración con datos enmarcados en el departamento de Antioquia y al año 2019, puede ser confirmada ampliando espacio-temporalmente los datos, lo que también permitiría la aplicación de analítica predictiva y prescriptiva por medio de algoritmos de ML (Aprendizaje de máquina por sus siglas en inglés o DL (Aprendizaje profundo por sus siglas en inglés) para simular escenarios que involucren mejoras a nivel de servicios públicos u otras variables, mejoras a la calidad de la educación y presupuesto disponible por municipio, para habilitar la búsqueda de soluciones óptimas a la hora, por ejemplo, de adjudicar presupuestos y aprobar proyectos.

10. Orientaciones metodológicas para el análisis geoespacial en el marco de una arquitectura BI

La arquitectura BI diseñada en esta monografía para apoyar la toma de decisiones en el sector público de Colombia, tomando como caso de estudio Antioquia, se fundamenta en arquitecturas tradicionales (ver Ong et al., 2011). Está diseñada bajo una estructura que no requiere de programación alguna para permitir el flujo de información desde la fuente de los datos hasta la visualización y aplicación de analítica. Además, permite crear un cuadro de mando fácilmente con los elementos generados en el EDA. Sin embargo, existen otras opciones como es el caso de introducir RStudio para acceder a la geodatabase. En RStudio es posible realizar el EDA, generar modelos SAR y hacer análisis de autocorrelación espacial con las librerías especializadas tales como *RGDAL, sp, sf, spdep* entre otras. También es posible elaborar visualizaciones y crear reportes en HTML y PDF (Bivand et al, 2013, Lovelace et al 2019, Xie et al, 2023).

La analítica geoespacial tradicionalmente se ha trabajado exclusivamente con sistemas de información geográfica, pero actualmente, los programas de minería de datos están incluyendo módulos para generar modelos que contengan la estructura espacial de los datos. Este es el caso de Knime, cómo se puede leer en su página web (https://www.knime.com/), que en su versión 4.7 incluye una extensión especializada en analítica geoespacial, la cual permite leer

directamente la geodatabase y al igual que R y Python permite hacer el EDA, elaborar visualizaciones, generar modelos SAR y hacer análisis de autocorrelación espacial.

También es posible diseñar una arquitectura en donde la capa de almacenamiento y usuario final estén contenidas en su totalidad en los SIG. Sin embargo, la versión de ArcGIS Pro utilizada en esta monografía no cuenta con el modelo SAR y por eso se recurrió a Geoda. En cambio, si tiene el análisis de autocorrelación de Moran. Otros sistemas de información geográfica de código abierto, tales como QGIS, permite la elaboración de extensiones, lo que facilita que el experto en programación agregue algoritmos que permitan realizar los análisis geoespaciales. Por otro lado, Carto es un sistema de Información geográfica en la nube, que se conecta a los servicios de AWS para correr algoritmos de analítica geoespacial una vez se haya conectado a una geodatabase. Todos los SIG tienen la capacidad de generar visualizaciones combinando mapas, con gráficos estadísticos y tablas. ArcGIS Pro ofrece la posibilidad de generar cuadros de mando interactivos.

Aunque en el sector público pueden tratarse datos que no cuenten con o no necesiten georreferenciación, la mayoría debería tenerlo en cuenta porque sus decisiones están arraigadas en las entidades territoriales. Sin embargo, es posible guardar datos estructurados no georreferenciados en una geodatabase y aplicar analítica no espacial con las herramientas de la arquitectura aquí propuesta o con las alternativas dadas.

Por otro lado, existen otro tipo de geodatabase que no se aloja en un sistema de información geográfica, sino en bases de datos relacionales que incluyen una columna con la geometría espacial, tal es el caso de PostGIS, una extensión de la base de datos PostgreSQL. Varios programas de analítica o SIG tienen conectores a alguna base de datos espacial tipo SQL, lo cual influiría en la escogencia de otras herramientas en una arquitectura BI cuya capa de almacenamiento se base en este tipo de solución.

También, es importante señalar que se necesita un equipo interdisciplinario para llevar a cabo la tarea de implementar una solución de inteligencia de negocios y en este caso particular debe incluirse expertos en modelamiento espacial. Este es uno de los retos identificados por los investigadores de la aplicación de la BI en el sector público (Hartley y Seymour, 2011; Abai et al, 2019; Elbashir et al, 2022). En las Figura 21 se observa esquemáticamente la interacción y

responsabilidades de diferentes actores en la arquitectura BI propuesta en esta monografía y en la figura 22 se muestran varias profesiones que intervienen en la construcción de soluciones BI.

Por otro lado, se necesita una investigación comparativa para determinar cuál es la solución de arquitectura BI más viable en el sector público que permita acelerar la investigación en temas relacionados con las políticas públicas, su diseño y monitoreo. Este análisis no solo debe incluir la simplicidad como factor relevante, pero también costos de software, de implementación, de mantenimiento y entrenamiento. Finalmente, es necesario la publicación de un documento oficial en donde se sienten las directrices para la integración de datos abiertos por parte de las entidades públicas y por ente territorial.

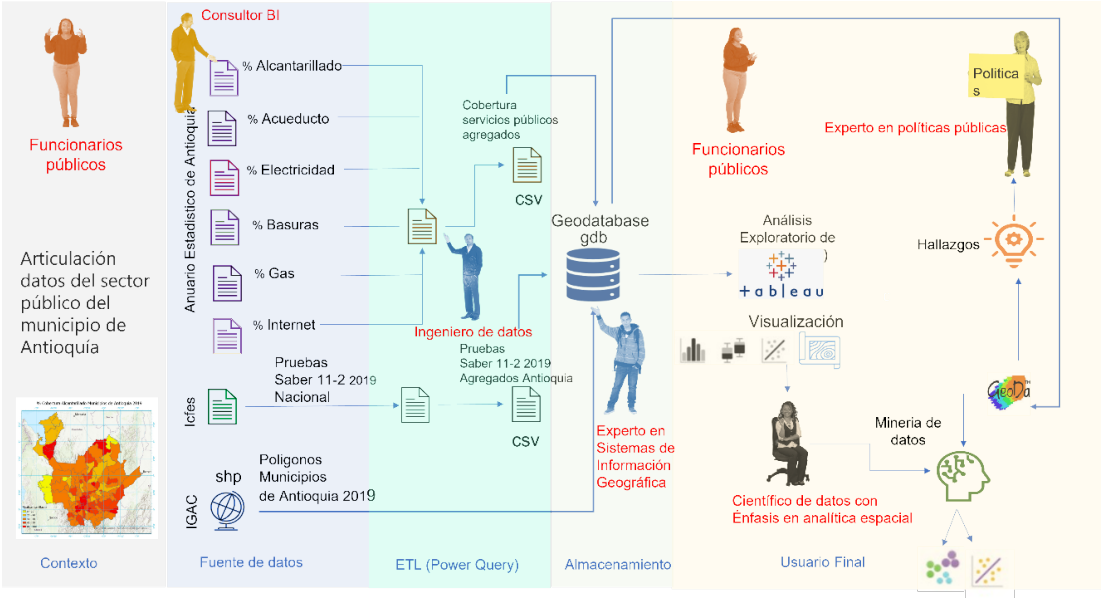


Figura 21. Lista del mínimo de profesionales que deben interactuar a lo largo de la implementación de una solución BI.

Elaboración propia

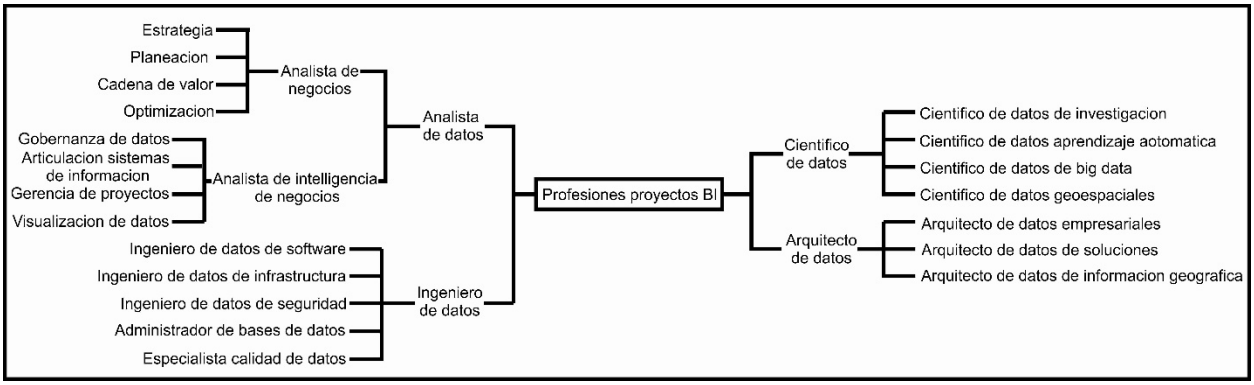


Figura 22. Gráfico ilustrando diferentes profesiones involucradas en el desarrollo de una solución de BI. Elaboración propia.

11. Conclusiones y trabajo futuro

Para el desarrollo de este trabajo se encontró material bibliográfico que permitió sustentar el desarrollo de la monografía, encontrando que en Colombia hay avances importantes en la publicación y utilización de datos abiertos de entes gubernamentales. También se determinó que no hay una metodología que oriente al usuario en cómo integrar información de varias entidades mediante una arquitectura BI para una mejor toma de decisiones en el sector público. Por otro lado, se tomaron referencias bibliográficas para proponer una arquitectura BI que facilite esta tarea.

Igualmente, se determinó que debido a la naturaleza geoespacial de los datos que normalmente se utilizan para el diseño de las políticas públicas, lo más apropiado es utilizar una geodatabase como repositorio, por ejemplo, la de ArcGIS Pro, en la que se pueden alojar archivos que contengan la estructura espacial de los datos y también importar archivos csv a tablas que pueden ser integradas a la estructura espacial, en este caso los municipios del departamento de Antioquia. Programas de BI como Tableau y de analítica geoespacial como Geoda pueden acceder directamente a la geodatabase y permiten a la capa de usuario final realizar la EDA y la analítica de datos.

Al realizar la EDA en Tableau, se pudo hacer la exploración de datos de manera interactiva y dinámica, permitiendo descubrir patrones que resultaron útiles a la hora de modelar los datos. Así, se determinó que una regresión tipo SAR utilizando la metodología de Anselin (2005) sería la más apropiada para describir la relación entre la cobertura de los servicios públicos domiciliarios, los puntajes globales de las pruebas Saber 11-2 2019 y las condiciones geográficas

del departamento de Antioquia. Para el caso de estudio en esta monografía, los resultados del modelo SAR demuestran que el puntaje global de las pruebas Saber 11-2 2019 es una variable dependiente de la estructura espacial y de la cobertura de los servicios públicos básicos.

Lo que implica que, para el desarrollo de las políticas educativas de los municipios de Antioquia, la variable de la cobertura de los servicios públicos debe ser incluida como factor importante para entender los alcances de las políticas. Es necesario identificar estrategias que ayuden a mitigar el impacto de una baja cobertura en servicios públicos en el rendimiento promedio en las pruebas estandarizadas de los estudiantes de los entes territoriales. Así mismo, se requiere establecer indicadores de medición para monitorear el impacto de las políticas públicas relacionadas con las variables que quieren relacionar.

De esta manera se ilustra, que una arquitectura BI que integre datos de diversos entes gubernamentales y que incluya la característica geoespacial de los datos es una herramienta útil para la toma de decisiones públicas basadas en datos puesto que permite articular información de diferentes entes gubernamentales para una región específica y así analizar posibles dependencias de diferentes variables. Estas dependencias pueden facilitar y optimizar la distribución de recursos públicos.

Se recalca que los datos utilizados en esta monografía están restringidos temáticamente a los servicios públicos y a los resultados de las pruebas Saber 11, espacialmente a un departamento (Antioquia) y temporalmente solo a un año (2019), así que como trabajo futuro se propone la conexión de más perspectivas al modelo de datos. Al subir archivos con datos de salud, vivienda, economía, violencia entre otros a la geodatabase se logra caracterizar a los municipios desde diferentes ángulos y explorar más relaciones que ayuden a fomentar políticas públicas, no solo en educación, sino también en otros campos.

Igualmente se propone la ampliación de la geodatabase a todos los municipios colombianos, lo que facilitaría el análisis de todo el territorio nacional. Así este tipo de análisis será factible para cualquier departamento o región del país. Por último, también convendría agregar la dimensión temporal, al cargar datos de diferentes años, tanto de la cobertura de los servicios públicos, como de los resultados de las pruebas Saber 11, así como de cualquier otro conjunto de datos que se logre conectar a la arquitectura. Esta variable permitiría realizar análisis

espacio-temporales, también útiles a la hora de medir el impacto de las políticas públicas en las regiones a lo largo del tiempo. Finalmente, con una cantidad mayor de datos, se podrían realizar modelos de inteligencia artificial que permitan generar soluciones para que en las entidades territoriales se pueda gobernar con base en los datos.

12.Referencias

Abai, N. H. Z., Yahaya, J. H., & Deraman, A. (2015, August). Incorporating business intelligence and analytics into performance management for the public sector issues and challenges. In 2015 International Conference on Electrical Engineering and Informatics (ICEEI) (pp. 484-489). IEEE.

Acevedo, M. F. (2012). Data analysis and statistics for geography, environmental science, and engineering. Taylor & Francis Group.

Anselin, L. (2003). Exploring Spatial Data with GeoDaTM : A Workbook. Center for Spatially Integrated Social Science.

Bivand, R., Pebesma, E., & Gómez-Rubio, V. (2013). Applied Spatial Data Analysis with R. In Springer eBooks. https://doi.org/10.1007/978-1-4614-7618-4

Burgos, J. R. (2022). Lluvias, servicios públicos y mortalidad infantil en Colombia (Rain, Public Services and Child Mortality in Colombia). Social Science Research Network. https://doi.org/10.2139/ssrn.4195056

Burleson, S. E., & Thoron, A. C. (2014). Maslow's hierarchy of needs and its relation to learning and achievement. Retrieved November, 12, 2019.

Castillo, A., Rodriguez D, Carrillo, J.M. (2015). Lattice Data: Plugin de QGIS que implementa análisis estadístico exploratorio de datos lattice para la identificación de correlación espacial. Tesis de Grado. Universidad Distrital Francisco José de Caldas. Bogotá.

Castro, M y Ruiz, J (2019). La educación secundaria y superior en Colombia vista desde las pruebas Saber. Prax. Saber [online]. 2019, vol.10, n.24, pp.341-366. ISSN 2216-0159. https://doi.org/10.19053/22160159.v10.n25.2019.9465.

Cervera, J. P. C (2021). Comportamiento del uso de datos abiertos en Colombia (2016-2021). Ciencia y Poder Aéreo, 17(1), 137-149. https://doi.org/10.18667/cienciaypoderaereo.742

Colmenares-Quintero, R. F., Maestre-Gongora, G. P., Pacheco-Moreno, L. J., Rojas, N., Stansfield, K. E., & Colmenares-Quintero, J. C. (2021). Analysis of the energy service in non-interconnected zones of Colombia using business intelligence. Cogent Engineering, 8(1), 1907970.

Collazos, A., Quintero, M., & Trujillo, K. (2021). Determinants of academic performance of the Saber 11 test during the 2014 - 2019 period in Colombia. Panorama, 303.

Consejo Nacional de Política Económica y Social [Conpes]. (2019). Política Nacional Para La Transformación Digital e Inteligencia Artificial. Conpes 3975 de 2019.

Consejo Nacional de Política Económica y Social [Conpes]. (2018). Política Nacional Para Política Nacional De Explotación De Datos (Big Data). Conpes 3920 de 2018.

Diaz, Y. M., Valentin, T. F., & Alvarez, J. R. (2021). Influencia del Internet en el Rendimiento Académico de los Estudiantes de Educación Básica Regular. Ciencia Latina Revista Científica Multidisciplinar, 5(3), 2477–2490. https://doi.org/10.37811/cl_rcm.v5i3.465

Elbashir, M. Z., Sutton, S. G., Arnold, V., & Collier, P. A. (2022). Leveraging business intelligence systems to enhance management control and business process performance in the public sector. Meditari Accountancy Research, 30(4), 914-940.

Ezeamama, A. E., Friedman, J. F., Acosta, L. P., Bellinger, D. C., Langdon, G. C., Manalo, D. L., Olveda, R. M., Kurtis, J. D., & McGarvey, S. T. (2005). HELMINTH INFECTION AND COGNITIVE IMPAIRMENT AMONG FILIPINO CHILDREN. American Journal of Tropical Medicine and Hygiene, 72(5), 540–548. https://doi.org/10.4269/ajtmh.2005.72.540

Ezeamama, A. E., Bustinduy, A. L., Nkwata, A., Martinez, L., Pabalan, N., Boivin, M. J., & King, C. H. (2018). Cognitive deficits and educational loss in children with schistosome infection—A systematic review and meta-analysis. PLOS Neglected Tropical Diseases, 12(1), e0005524. https://doi.org/10.1371/journal.pntd.0005524

Ezeamama, A. E., McGarvey, S. T., Acosta, L. P., Zierler, S., Manalo, D. L., Wu, H., Kurtis, J. D., Mor, V., Olveda, R. M., & Friedman, J. F. (2008). The Synergistic Effect of Concomitant Schistosomiasis, Hookworm, and Trichuris Infections on Children's Anemia Burden. PLOS Neglected Tropical Diseases, 2(6), e245. https://doi.org/10.1371/journal.pntd.0000245

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2000). Quantitative geography: Perspectives on spatial data analysis. SAGE Publications, Limited.

Greasley, A. (2019). Simulating Business Processes for Descriptive, Predictive, and Prescriptive Analytics (1st ed.). De Gruyter. https://www.perlego.com/book/1357831/simulating-business-processes-for-descriptive-predictive-and-prescriptive-analytics-pdf

Groebner, D. F., & Shannon, P. W. (1993b). Business Statistics: A Decision-making Approach. Prentice Hall.

Hartley, K., & Seymour, L. F. (2011, October). Towards a framework for the adoption of business intelligence in public sector organisations: the case of South Africa. In Proceedings of the South African Institute of Computer Scientists and Information Technologists Conference on Knowledge, Innovation and Leadership in a Diverse, Multidisciplinary Environment (pp. 116-122).

Hernández, J., Ramírez, M. J., & Ferri, C. (2004). Introducción a la minería de datos. Pearson Prentice Hall.

Herrero, J. G., López, J. M. M., De Jesús, A. B., Guisado, M. Á. P., Bustamante, Á. L., & R, W. P. (2018). Ciencia de datos: técnicas analíticas y aprendizaje estadístico. Un enfoque práctico.

Hutcheson, G. D., & Sofroniou, N. (1999). The multivariate social scientist: Introductory statistics using generalized linear models. SAGE Publications, Limited.

Instituto Colombiano para la Evaluación de la Educación [Icfes]. (2022). Brechas en aprendizaje: una mirada desde las pruebas de Estado. Apuntes Del Icfes Para La Política Educativa.

Instituto Colombiano para la Evaluación de la Educación [Icfes]. (2023). Inequidades en el logro académico en las regiones: resultados en la prueba de lectura. Apuntes Del Icfes Para La Política Educativa.

Idrovo, A.J (2012). Diagnóstico Nacional de Salud Ambiental. Prosperidad Para Todos. Ministerio de Ambiente y Desarrollo Sostenible. República de Colombia.

Inmon, W., & Linstedt, D. (2014). Data Architecture: A Primer for the Data Scientist: Big Data, Data Warehouse and Data Vault. Morgan Kaufmann.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R. Springer Science & Business Media.

Larose, D. T., & Larose, C. D. (2014). Discovering Knowledge in Data: An Introduction to Data Mining (Wiley Series on Methods and Applications in Data Mining) (2nd ed.). Wiley.

Loshin, D. (2012). Business intelligence: The savvy manager's guide. Elsevier Science & Technology.

Lovelace, R., Nowosad, J., & Muenchow, J. (2019). Geocomputation with R. In Chapman and Hall/CRC eBooks. https://doi.org/10.1201/9780203730058

Matthews, S. (2006). Geoda and Spatial Regression Modeling. Population Research Institute Presentation. Pennsylvania State University

Matias Camargo, S. R. (2015). La regulación económica de los servicios públicos domiciliarios en Colombia. Diálogos De Saberes, 42, 63–78. https://doi.org/10.18041/0124-0021/dialogos.42.2015.6061

Martínez Mateus, W. A., & Turriago Hoyos, Á. (2015). Análisis de distribución geográfica y espacial de los resultados de las Pruebas Saber 11 del Instituto Colombiano para el Fomento de la Educación Superior (ICFES). 2005-2012. Colombia. Cuadernos Latinoamericanos de Administración, XI(21), 39-49.

Munné, R. (2016). Big Data in the Public Sector. Springer eBooks, 195-208. https://doi.org/10.1007/978-3-319-21569-3_11

Kalelkar, M., Churi, P., & Kalelkar, D. (2014). Implementation of model-view-controller architecture pattern for business intelligence architecture. International Journal of Computer Applications, 102(12).

Ong, I. L., Siew, P. H., & Wong, S. C. (2011). A Five-Layered Business Intelligence Architecture. Communications of the IBIMA, 1-11. https://doi.org/10.5171/2011.695619

Ordóñez, M. G., & González, M. P. (2021). Caracterización de marcos de referencia que apoyan la implementación del gobierno de datos propuesto por MinTIC para entidades públicas. Investigación e Innovación en Ingenierías, 9(2), 42-58.

Olson, D. L. (2018). Data Mining Models, Second Edition (2nd ed.). Business Expert Press.

Sedkaoui, S. (2018). Data analytics and big data. John Wiley & Sons, Incorporated.

Sharda, R., Delen, D., & Turban, E. (2018). Business Intelligence, Analytics, and Data Science: A Managerial Perspective (4th ed.). Pearson.

Sherman, Rick (2014). Business Intelligence Guidebook: From Data Integration to Analytics, Elsevier Science & Technology.

Siabato, W y Guzmán-Manrique J (2019). “La autocorrelación espacial y el desarrollo de la geografía cuantitativa.” Cuadernos de Geografía: Revista Colombiana de Geografía 28 (1): 1-22. doi: 10.15446/rcdg.v28n1.76919.

Schmitt, M. (2023). Deep learning in business analytics: a clash of expectations and reality. International Journal of Information Management Data Insights, 3(1), 100146.

Van Der Lans, Rick (2012). Data Virtualization for Business Intelligence Systems: Revolutionizing Data Integration for Data Warehouses, Elsevier Science & Technology

Vargas-Merino, J. A., & Valladares-Castillo, S. O. (2019). Aplicaciones de inteligencia de negocios en el sector público: Una revisión sistemática de la literatura [Trabajo de investigación]. Universidad Peruana Unión.

Varona-Taborda, M.A, Mosquera-Ramírez. J.C, Medina-Moreno, C.V, Lemus-Muñoz, D.F, Muñoz-Hernández, C.J, Arias-Iragorri, C.G, “Business Intelligence for the Programs of the Secretaries of Health, Education and Planning in a Territorial Entity,” Revista Facultad de Ingeniería, vol. 30 (58), e13826, 2021. https://doi.org/10.19053/01211129.v30.n58.2021.13826

Vidal, J (2014) Consideraciones procesos ETL en entornos Big Data: Caso Hadoop | Dataprix.

Williams, S., & Williams, N. (2010). The Profit Impact of Business Intelligence. Elsevier.

Xie, Y., Allaire, J., & Grolemund, G. (2018). R Markdown: The Definitive Guide. https://www.amazon.com/Markdown-Definitive-Guide-Chapman-Hall/dp/1138359335