MAESTRÍA EN EXPLOTACIÓN
DE DATOS Y GESTIÓN DEL
CONOCIMIENTO

Trabajo de Grado

Validación cruzada de la información sobre conectividad en Colombia: análisis comparativo entre fuentes oficiales (ColombiaTIC) y mediciones independientes (Ookla).

Alumno
José Eduardo Valdés Castro

Director
Martin Volpacchio

Versión 1

01/11/2025

Contenido

Listado de Tablas

Listado de Ilustraciones

Resumen

La presente investigación tiene como propósito evaluar la calidad y consistencia de la información pública sobre los servicios de telecomunicaciones en Colombia, con énfasis en la velocidad de conexión de Internet fijo. Este estudio surge ante la necesidad de contar con datos verificables y transparentes que respalden las decisiones gubernamentales y empresariales relacionadas con la conectividad digital, especialmente en un contexto pospandemia en el que la demanda de acceso confiable a Internet se ha convertido en un factor crítico para el desarrollo social, educativo y económico.

El problema central radica en la ausencia de mecanismos sistemáticos de validación de los reportes oficiales sobre conectividad publicados por el Ministerio de Tecnologías de la Información y las Comunicaciones (MinTIC) a través del portal ***ColombiaTIC****. Para abordar esta brecha, se propone una metodología basada en* ***ciencia de datos****, que contrasta los registros oficiales con mediciones empíricas provenientes de* ***Ookla (Speedtest)****, una plataforma independiente y reconocida internacionalmente por su cobertura global y precisión técnica.*

La investigación se desarrolló en tres fases: (1) construcción de la fuente primaria mediante scraping y estructuración de los reportes trimestrales del MinTIC; (2) obtención, filtrado y georreferenciación de los datos abiertos de Ookla; y (3) análisis comparativo de consistencia estadística y territorial utilizando bases consolidadas en ***DuckDB*** *y técnicas de análisis descriptivo y correlacional.*

Finalmente, como producto analítico central, se propone el ***Índice Nacional de Coherencia TIC (INCTIC)****, diseñado para medir el grado de coincidencia entre las cifras reportadas por el Ministerio TIC y las mediciones independientes de Ookla. Este índice, expresado en porcentaje de coherencia por departamento, sintetiza la correspondencia entre los valores observados y los reportados, ofreciendo una medida objetiva y replicable de la calidad de la información pública sobre conectividad en Colombia. Su cálculo, implementado en Python dentro del tercer notebook del proyecto, permite visualizar las diferencias territoriales y constituye un aporte metodológico original para futuras investigaciones y ejercicios de control de calidad de datos abiertos en el sector TIC.*

Palabras clave: *Calidad de servicio, conectividad, datos públicos, ciencia de datos, validación cruzada, telecomunicaciones, transparencia, Ookla, ColombiaTIC, DuckDB.*

Abstract

This research aims to evaluate the quality and consistency of public information regarding telecommunication services in Colombia, focusing on fixed Internet connection speed. The study responds to the need for reliable and verifiable data to support government and corporate decision-making in digital connectivity—

particularly in the post-pandemic context, where stable Internet access has become essential for social, educational, and economic development.

The main problem lies in the lack of systematic validation mechanisms for official connectivity reports published by the Ministry of Information and Communication Technologies (MinTIC) through the ***ColombiaTIC*** *platform. To address this issue, a* ***data science-based methodology*** *is proposed to cross-validate official reports with empirical measurements from* ***Ookla (Speedtest)****, a globally recognized independent platform for Internet performance data.*

The research was structured in three stages: (1) construction of the primary data source via automated scraping and structuring of quarterly MinTIC reports; (2) retrieval, spatial filtering, and georeferencing of Ookla's open datasets; and (3) comparative analysis of statistical and territorial consistency using ***DuckDB*** *as the analytical repository, applying descriptive and correlational analysis techniques.*

Finally, as the core analytical contribution of this research, the ***National ICT Consistency Index (INCTIC)*** *is proposed to quantify the degree of alignment between the figures reported by the Ministry of ICT and the independent measurements provided by Ookla. This index, expressed as a percentage of consistency per department, summarizes the correspondence between observed and reported values, providing an objective and replicable measure of the quality of public connectivity data in Colombia. Its calculation, implemented in Python within the third notebook of the project, enables visualization of territorial differences and represents an original methodological contribution for future research and quality control of open data in the ICT sector.*

Keywords: *Service quality, connectivity, public data, data science, cross-validation, telecommunications, transparency, Ookla, ColombiaTIC, DuckDB.*

CAPÍTULO 1 – INTRODUCCIÓN

1.1. Contexto de la conectividad en Colombia

Durante la última década, Colombia ha experimentado un avance significativo en el fortalecimiento de su infraestructura digital, impulsando el acceso a Internet como un elemento esencial para la competitividad, la productividad y la equidad social. **El Ministerio de Tecnologías de la Información y las Comunicaciones de Colombia** (**MinTIC**), a través de los *Boletines Trimestrales del Sector TIC*, consolida los reportes de los operadores sobre accesos fijos y móviles, velocidades contratadas y cobertura territorial, constituyéndose en la fuente oficial de información para la formulación de políticas públicas y la evaluación de los programas de conectividad (ColombiaTIC, 2024).

No obstante, a pesar del crecimiento sostenido en la cobertura reportada, persisten notables disparidades regionales en la calidad del servicio, especialmente entre zonas urbanas y rurales. Estas diferencias impactan directamente en la educación, la productividad y el acceso a servicios esenciales como la salud, donde las limitaciones de conectividad restringen la telemedicina y la educación virtual (Henao Colorado, 2020) (Camacho). Además, la calidad desigual de los servicios de telecomunicaciones reduce la competitividad empresarial y la capacidad del país para integrarse plenamente en la economía digital global (Alderete, 2012).

En este escenario, uno de los principales desafíos radica en la brecha entre la velocidad de conexión contratada y la realmente experimentada por los usuarios. En diversas regiones del país, los indicadores de desempeño reportados por los operadores no reflejan con precisión la experiencia real del servicio. Esta diferencia plantea interrogantes sobre la fiabilidad y coherencia de los datos públicos que sustentan las decisiones gubernamentales y la evaluación de la calidad del servicio (Moreano, 2010) (Henao Colorado, 2020).

Ante esta situación, las fuentes independientes de medición empírica, como ***Ookla Speedtest Open Data***, adquieren relevancia al ofrecer un insumo objetivo y verificable que permite contrastar las cifras oficiales reportadas por los operadores (LLC, 2025). El acceso a estos datos abiertos facilita el uso de metodologías basadas en **ciencia de datos** para analizar la coherencia entre la información gubernamental y las mediciones reales de los usuarios, contribuyendo a la transparencia y fortaleciendo la confianza pública en la gestión de la infraestructura digital.

El presente estudio se enmarca, por tanto, en la necesidad de consolidar un modelo de validación técnica de los datos del sector TIC en Colombia, mediante la integración y análisis comparativo de fuentes oficiales —como los boletines de ColombiaTIC— con mediciones empíricas provenientes de plataformas globales. Este enfoque busca proporcionar una herramienta de apoyo para la toma de decisiones informadas, la mejora de la calidad del servicio y el fortalecimiento de la política pública en materia de conectividad, promoviendo así un acceso más equitativo y confiable a las tecnologías de la información en todo el territorio nacional.

1.2. Problema de investigación

La información estadística del sector de Tecnologías de la Información y las Comunicaciones (TIC) publicada por el **Ministerio de Tecnologías de la Información y las Comunicaciones** (**MinTIC**) constituye una fuente fundamental para la formulación de políticas públicas, la planeación territorial y la evaluación de los avances en materia de transformación digital. Sin embargo, a pesar de su relevancia, existen discrepancias entre las cifras reportadas por los operadores de telecomunicaciones y las mediciones empíricas de velocidad disponibles en fuentes abiertas como *Ookla Speedtest Open Data*, lo que plantea dudas sobre la precisión y representatividad de los datos oficiales (Moreano, 2010); (ColombiaTIC, 2024).

Esta situación adquiere especial importancia dado que los informes oficiales del MinTIC son utilizados como referencia para la toma de decisiones en inversión pública y la priorización de proyectos de conectividad. La falta de verificación independiente de los datos puede introducir sesgos en la planeación de infraestructura digital y afectar la distribución equitativa de recursos destinados al cierre de brechas tecnológicas (Serrano, 2019); (Patricia, 2019).

Los estudios previos sobre calidad del servicio en el sector de telecomunicaciones se han centrado principalmente en la percepción de los usuarios, analizando la satisfacción y el valor percibido (Riccio, 2019); (Botero, 2006). Otros trabajos, desarrollados en segmentos como la televisión por suscripción o la telefonía móvil, han identificado la calidad del servicio como un factor estratégico y competitivo para las empresas (Galbán, 2013); (Aguirre Julcapoma, 2018), pero no han abordado sistemáticamente la coherencia entre la información pública reportada por los operadores y las mediciones técnicas realizadas en campo.

En consecuencia, persiste un vacío metodológico en la verificación empírica de los datos oficiales sobre velocidad de Internet fijo. Esta falta de mecanismos de contraste impide evaluar con precisión la calidad real del servicio y genera incertidumbre sobre la fiabilidad de las estadísticas que sustentan la política pública de conectividad.

Por tanto, el **problema central** que aborda esta investigación radica en la **ausencia de un modelo de validación empírica y continua de los datos públicos reportados por el Ministerio TIC**, que permita identificar discrepancias entre lo informado por los operadores y lo efectivamente experimentado por los usuarios en diferentes municipios del país.

Pregunta de investigación:

¿Qué nivel de coherencia existe entre los datos oficiales reportados por el Ministerio TIC sobre la velocidad de Internet fijo y los valores empíricos medidos por la plataforma Ookla en los municipios de Colombia?

Responder a esta pregunta permitirá determinar la consistencia de los datos públicos y ofrecer una herramienta de verificación basada en evidencia.

1.3. Justificación

El fortalecimiento de la calidad, coherencia y transparencia de los datos públicos en el sector de Tecnologías de la Información y las Comunicaciones (TIC) constituye un requisito esencial para la construcción de una sociedad digital equitativa y basada en evidencia. En este contexto, el presente trabajo busca aportar un modelo de **validación técnica de datos públicos mediante ciencia de datos aplicada**, fortaleciendo la confiabilidad de los indicadores de conectividad nacional y promoviendo decisiones más precisas en materia de política pública (Reina Nossa, 2018).

La comparación entre los reportes oficiales del **Ministerio de Tecnologías de la Información y las Comunicaciones (MinTIC)** y las mediciones empíricas proporcionadas por **Ookla Speedtest Open Data** permite obtener una visión objetiva y verificable del desempeño de la infraestructura digital. Este contraste no solo contribuye a la transparencia institucional, sino que mejora la rendición de cuentas y la confianza ciudadana en la información que orienta la planeación de la inversión pública en conectividad (Alderete, 2012); (Castillo, 2017).

Desde el punto de vista técnico, la investigación se fundamenta en la **implementación de procesos ETL (Extract, Transform, Load)** reproducibles, desarrollados en **Python y DuckDB**, que permiten automatizar la descarga, transformación y consolidación trimestral de los reportes oficiales y de las mediciones independientes (Javier, 2013). Este enfoque facilita la integración de fuentes heterogéneas de información en una base de datos auditable y actualizable, promoviendo la replicabilidad y continuidad del análisis en futuras investigaciones.

El principal aporte metodológico del estudio radica en la formulación del **Índice Nacional de Coherencia TIC (INCTIC)**, un indicador diseñado para medir la correspondencia entre los datos reportados por los operadores y las mediciones reales de conectividad. Este índice constituye una herramienta técnica de monitoreo que permite detectar inconsistencias, cuantificar la fiabilidad de los datos públicos y ofrecer un insumo objetivo para la mejora continua de la infraestructura digital del país.

En términos sociales y económicos, la investigación ofrece un instrumento de alto valor estratégico para la toma de decisiones. Al identificar las **brechas de calidad** entre regiones, el modelo permite orientar los recursos hacia zonas de menor desempeño y promover una distribución más equitativa de la inversión pública. A nivel institucional, los resultados fortalecen la capacidad del Estado para planificar proyectos de conectividad con criterios de evidencia y eficiencia, contribuyendo al cumplimiento de los **Objetivos de Desarrollo Sostenible (ODS)** en materia de reducción de brechas digitales y fomento de la inclusión tecnológica (Alderete, 2012).

En última instancia, esta investigación busca generar conocimiento aplicado que impulse la **equidad digital, la competitividad y la eficiencia del gasto público**, demostrando cómo el uso de metodologías de ciencia de datos puede consolidarse como una herramienta efectiva para la evaluación y mejora de la política pública de conectividad en Colombia.

1.4. Objetivos

- **Objetivo general**

Evaluar la coherencia entre los datos oficiales reportados por el Ministerio TIC de Colombia y las mediciones empíricas de velocidad de Internet publicadas por Ookla, mediante la construcción del **Índice Nacional de Coherencia TIC (INCTIC)** basado en técnicas de análisis estadístico y procesamiento de datos abiertos.

- **Objetivos específicos**

1. Extraer, transformar y consolidar los reportes trimestrales del sector TIC de MinTIC mediante un proceso automatizado de ETL reproducible.
2. Descargar, limpiar y estructurar los datos abiertos de velocidad de Internet de Ookla para el territorio colombiano.
3. Integrar ambas bases en un modelo relacional unificado y realizar la validación cruzada por municipio y departamento.
4. Calcular el Índice Nacional de Coherencia TIC (INCTIC) para medir la correspondencia entre fuentes.
5. Analizar los resultados obtenidos por región y establecer patrones de discrepancia entre la información reportada y la medida.

1.5. Alcance y limitaciones

El alcance del estudio comprende el **análisis de los boletines trimestrales del sector TIC y los datos abiertos de Ookla disponibles hasta el primer trimestre de 2025**[1], abarcando los municipios y departamentos del territorio colombiano donde existan mediciones coincidentes.

El proyecto se limita al análisis de las **velocidades de subida y bajada de Internet fijo**, excluyendo los servicios móviles y otras modalidades de conexión. Asimismo, el índice propuesto se enfoca en la **coherencia estadística entre fuentes de información**, sin evaluar aspectos cualitativos del servicio como estabilidad o latencia.

Otra limitación es la **disponibilidad temporal** de los reportes, ya que las actualizaciones del MinTIC y de Ookla no siempre ocurren de manera simultánea, lo que puede generar desfasamientos en la comparación. No obstante, el modelo desarrollado permite actualizar automáticamente la base consolidada cada trimestre, garantizando la continuidad y escalabilidad del análisis.

En términos de impacto, el estudio busca servir como base para investigaciones futuras orientadas a la **verificación empírica de indicadores públicos** y a la construcción de herramientas analíticas que promuevan la transparencia y el uso inteligente de los datos abiertos.

[1] A corte de la documentación de la versión del trabajo de grado, se cuenta disponible el reporte trimestral de Colombia TIC del segundo trimestre del año 2025 (aun no procesado) y el cuarto trimestre del año 2024 del Ookla.

CAPÍTULO 2 – MARCO REFERENCIAL

2.1. Marco teórico

El avance de las Tecnologías de la Información y las Comunicaciones (TIC) ha transformado la forma en que los gobiernos formulan políticas públicas, convirtiendo los **datos abiertos** en un componente estratégico para la transparencia, la toma de decisiones basada en evidencia y la evaluación del desempeño institucional (Reina Nossa, 2018). En este contexto, los datos de conectividad constituyen un recurso fundamental para monitorear la calidad de los servicios digitales, la inclusión tecnológica y el progreso hacia una economía basada en el conocimiento (Alderete, 2012).

En Colombia, el **Ministerio de Tecnologías de la Información y las Comunicaciones (MinTIC)** lidera la publicación trimestral de los *Boletines del Sector TIC*, que consolidan los reportes de los operadores sobre accesos fijos y móviles, velocidades contratadas, cobertura y evolución de los servicios. Estos informes oficiales son utilizados como fuente de referencia por entidades gubernamentales, operadores y organismos internacionales para medir los avances del país en materia de conectividad y transformación digital (ColombiaTIC, 2024). Sin embargo, diversos estudios han señalado que la ausencia de mecanismos de validación empírica puede afectar la fiabilidad de los datos reportados, generando distorsiones en la interpretación de los indicadores oficiales (Moreano, 2010); (Henao Colorado, 2020).

Paralelamente, fuentes internacionales independientes como **Ookla Speedtest Open Data** recopilan millones de mediciones empíricas realizadas por los usuarios finales en tiempo real, lo que permite analizar el **rendimiento efectivo de la conectividad** a nivel local y regional (LLC, 2025). Estas mediciones constituyen una oportunidad para contrastar la información oficial con los datos reales de desempeño, posibilitando la creación de indicadores más precisos y auditables sobre la calidad del servicio de Internet.

La relevancia de integrar estas dos perspectivas —la oficial y la empírica— radica en la posibilidad de identificar discrepancias estructurales entre la **velocidad reportada** y la **velocidad experimentada**, fenómeno que impacta la confianza ciudadana en la información pública y la efectividad de las políticas de inclusión digital (Serrano, 2019); (Patricia, 2019). Esta validación cruzada contribuye a garantizar que los datos que orientan la inversión pública y la planeación de infraestructura digital reflejen con fidelidad la realidad de los usuarios en todo el territorio nacional.

Desde una perspectiva metodológica, el proyecto se fundamenta en los principios de la **ciencia de datos aplicada a la gestión pública** (Provost, 2013), donde los datos son considerados un activo para la mejora continua de los servicios estatales. El proceso técnico de integración se desarrolla bajo un enfoque **ETL (Extract, Transform, Load)**, que garantiza la recolección, limpieza y consolidación de información proveniente de múltiples fuentes (Kimball, 2011). Este modelo ha demostrado su eficacia en el manejo de **grandes volúmenes de datos (big data)** dentro del sector de telecomunicaciones, al permitir automatizar procesos de análisis y asegurar la trazabilidad de los resultados (Javier, 2013).

La infraestructura tecnológica del estudio se apoya en **Python**, lenguaje ampliamente utilizado en entornos de analítica avanzada, y **DuckDB**, motor de bases de datos analíticas de alto rendimiento diseñado para entornos embebidos y análisis in-memory (Raasveldt M. &., 2019). Esta arquitectura permite reproducir los flujos de procesamiento y mantener una base consolidada actualizable de manera trimestral, alineada con los reportes del MinTIC.

Desde el punto de vista estadístico, la evaluación de la relación entre las dos fuentes de datos se sustenta en la aplicación de medidas de asociación y concordancia. En particular, se emplea el **coeficiente de correlación de Pearson** (Rodgers, 1988) para medir la fuerza y dirección de la relación lineal entre las velocidades reportadas y las medidas empíricas, y el **coeficiente de concordancia de Lin** (Lin, 1989) para estimar la coherencia global entre ambas variables. Estos indicadores permiten identificar sesgos sistemáticos o discrepancias regionales que puedan influir en la interpretación de la calidad del servicio.

Finalmente, el modelo teórico que sustenta este estudio se orienta hacia el diseño del **Índice Nacional de Coherencia TIC (INCTIC)**, un indicador cuantitativo que busca representar el grado de correspondencia entre la información oficial del MinTIC y las mediciones reales obtenidas por los usuarios. Este índice se concibe como una herramienta técnica para fortalecer la rendición de cuentas, mejorar la precisión de los indicadores de conectividad y apoyar la formulación de políticas públicas basadas en evidencia empírica.

En conjunto, este marco teórico integra las perspectivas de la ingeniería de software, la analítica de datos y la política pública, sustentando la creación de un **modelo reproducible de verificación empírica** de la calidad del servicio de Internet en Colombia, alineado con los principios de **Gobierno Digital**, **transparencia**, y **apertura de datos** promovidos a nivel nacional e internacional (Reina Nossa, 2018); (Alderete, 2012).

2.2. Marco normativo

El marco normativo que orienta el acceso, publicación y uso de datos abiertos en Colombia se enmarca en la **Política de Gobierno Digital**, adoptada mediante el **Decreto 1008 de 2018**, el cual establece los lineamientos para la gestión, disponibilidad y aprovechamiento de la información pública generada por las entidades del Estado. Este decreto define que los datos producidos por instituciones públicas deben ser **abiertos, accesibles, reutilizables y verificables** por la ciudadanía, promoviendo su uso mediante herramientas de **ciencia de datos e inteligencia analítica** para fortalecer la transparencia, la eficiencia administrativa y la rendición de cuentas ((MinTIC), 2018).

Complementariamente, los **Lineamientos de Datos Abiertos del Ministerio TIC** (Colombia, 2023) precisan que las entidades públicas deben publicar sus conjuntos de datos en formatos abiertos, siguiendo los principios de **interoperabilidad y calidad de la información**, para facilitar su integración en procesos de innovación pública y control social. En este sentido, la presente investigación se enmarca dentro de los postulados del **ecosistema de datos abiertos**, al emplear información pública de libre acceso como insumo técnico para validar la coherencia de los reportes de conectividad.

Asimismo, el **Plan Nacional de Desarrollo 2022–2026** (Ley 2294 de 2023) reconoce la **conectividad digital** como un **pilar de equidad territorial** y un componente estratégico del desarrollo económico y social del país. En su línea de acción *"Conectividad para la inclusión y la productividad"*, promueve el uso de **indicadores técnicos de desempeño** para monitorear la calidad del servicio de Internet, reducir brechas digitales y garantizar un acceso equitativo a las tecnologías de la información. Este trabajo se articula con dichos objetivos, al proponer un **índice complementario de coherencia (INCTIC)** que permite contrastar los datos oficiales del Ministerio TIC con las mediciones empíricas provenientes de fuentes independientes.

En el ámbito internacional, Colombia se adhiere a las recomendaciones de la **Unión Internacional de Telecomunicaciones (UIT)**, organismo especializado de las Naciones Unidas encargado de establecer normas y metodologías globales para la medición de la **calidad de servicio (Quality of Service, QoS)** y la **calidad de la experiencia (Quality of Experience, QoE)**. Estas directrices, recogidas en los informes técnicos del **Sector de Normalización de las Telecomunicaciones (ITU-T)**, promueven la utilización de datos verificables y comparables para fortalecer la **toma de decisiones basada en evidencia** y mejorar la calidad de los servicios de conectividad ((UIT), 2025).

Del mismo modo, la **Agenda 2030 para el Desarrollo Sostenible** de las Naciones Unidas establece en su **Objetivo 9 (Industria, Innovación e Infraestructura)** y **Objetivo 16 (Instituciones sólidas y transparencia)** la importancia de garantizar el acceso universal a las TIC y fomentar el uso responsable de los datos abiertos como instrumento de desarrollo inclusivo y sostenible ((ONU), 2015).

En conjunto, este marco normativo respalda la utilización de **fuentes abiertas, mediciones independientes y metodologías de análisis reproducible** como herramientas para la **validación técnica de los reportes oficiales** en materia de conectividad. De esta manera,

la investigación contribuye al cumplimiento de los principios de **Gobierno Digital**, **transparencia**, **rendición de cuentas** y **política pública basada en evidencia**, consolidando un modelo alineado tanto con la normativa nacional como con los estándares internacionales de gobernanza de datos.

2.3. Marco conceptual

El desarrollo del presente estudio se fundamenta en un conjunto de conceptos técnicos y metodológicos esenciales que orientan la construcción del modelo de análisis y la interpretación de los resultados obtenidos. A continuación, se definen los términos clave utilizados a lo largo del proyecto:

Datos abiertos: Se entiende por datos abiertos toda información pública generada por entidades del Estado que se pone a disposición de los ciudadanos en formatos accesibles, estructurados y reutilizables, de acuerdo con los principios de transparencia, participación y colaboración establecidos en la Política de Gobierno Digital (Colombia, 2023). Su disponibilidad permite el análisis independiente y la verificación ciudadana de los indicadores oficiales, fortaleciendo la rendición de cuentas y la toma de decisiones basada en evidencia.

Ciencia de datos: Es un campo interdisciplinario que combina la estadística, la informática y el análisis computacional con el fin de extraer conocimiento y valor a partir de grandes volúmenes de datos heterogéneos (Provost, 2013). En este proyecto, la ciencia de datos constituye la base metodológica para integrar, limpiar y analizar información proveniente de fuentes oficiales (ColombiaTIC) y empíricas (Ookla), orientando la generación de un indicador objetivo de coherencia.

ETL (Extract, Transform, Load): Hace referencia al proceso sistemático de extracción, transformación y carga de datos desde diferentes fuentes hacia una base unificada y estructurada, garantizando la calidad, consistencia y trazabilidad de la información para su posterior análisis (Kimball, 2011). El proceso ETL implementado en esta investigación fue automatizado en Python y orientado a mantener la actualización trimestral de los boletines TIC y los reportes trimestrales de Ookla que son presentadas en formato parquet.

DuckDB: Es un motor de base de datos analítica embebida y de código abierto, diseñado para ejecutar consultas OLAP (Online Analytical Processing) de manera eficiente en entornos locales, sin requerir servidores dedicados. Su arquitectura columnar y su integración nativa con pandas permiten manipular grandes volúmenes de datos con alta velocidad y bajo consumo de recursos (Raasveldt M. &., 2019).

Ookla Speedtest: Es una plataforma global de medición de velocidad de Internet que recopila de manera continua datos empíricos sobre la velocidad de bajada, subida y latencia a partir de millones de pruebas realizadas por usuarios finales en distintos países (LLC, 2025). Estos registros constituyen una fuente independiente y verificable para contrastar la calidad de los servicios de conectividad reportados oficialmente.

Índice Nacional de Coherencia TIC (INCTIC): Indicador propuesto en esta investigación que mide el grado de concordancia entre la velocidad de Internet reportada por los operadores (ColombiaTIC) y la velocidad medida empíricamente (Ookla). Se define como la razón entre ambos valores, donde un resultado cercano a 1 refleja alta coherencia y valores alejados evidencian posibles divergencias o inconsistencias en la información reportada.

Correlación de Pearson (r): Es una medida estadística que cuantifica el grado y la dirección de la relación lineal entre dos variables numéricas (Rodgers, 1988). En este estudio se aplica para evaluar la correspondencia entre las velocidades de subida y bajada obtenidas de las dos fuentes de información.

Outlier o dato atípico: Corresponde a una observación cuyo valor difiere significativamente del patrón general de la muestra. Los outliers pueden distorsionar las estimaciones estadísticas y alterar los resultados de correlación; por ello, fueron identificados y tratados mediante técnicas de limpieza y filtrado de datos, garantizando la validez del análisis (Barnett, 1994).

En conjunto, estos conceptos conforman el marco conceptual que sustenta la integración metodológica, tecnológica y analítica del proyecto, asegurando la coherencia entre el modelo propuesto, la calidad de los datos y la interpretación estadística de los resultados.

CAPÍTULO 3 – METODOLOGÍA

3.1. Tipo de investigación

La presente investigación se enmarca dentro de un **enfoque cuantitativo y aplicado**, orientado a la **validación empírica de la coherencia de los datos públicos del sector de telecomunicaciones en Colombia**, específicamente en lo que respecta a la velocidad del servicio de Internet fijo-reportada por el Ministerio de Tecnologías de la Información y las Comunicaciones (MinTIC) y las mediciones reales obtenidas de la plataforma **Ookla Speedtest Open Data**.

Desde el punto de vista **epistemológico**, el estudio se ubica en el paradigma **positivista**, ya que se fundamenta en la observación sistemática, el registro y la medición de variables numéricas para describir y analizar fenómenos de conectividad de forma objetiva (Hernández Sampieri, 2014). El diseño metodológico no busca modificar las condiciones existentes, sino **analizar y comparar los datos reportados** por diferentes fuentes con el propósito de determinar el grado de correspondencia entre ellas.

Se clasifica como una **investigación no experimental**, puesto que las variables no son manipuladas deliberadamente; se analizan tal como se presentan en los reportes públicos oficiales y las mediciones empíricas. Asimismo, se considera un **estudio longitudinal**, debido a que se examinan datos históricos comprendidos entre 2021 y 2025, lo que permite observar tendencias y variaciones temporales en la calidad del servicio de Internet fijo en los distintos municipios y departamentos del país.

Por su naturaleza, también se enmarca dentro del **nivel descriptivo y correlacional**. El nivel descriptivo permite caracterizar los valores de velocidad de subida y bajada en cada fuente de información, mientras que el nivel correlacional busca establecer la relación existente entre las velocidades reportadas oficialmente y las realmente experimentadas por los usuarios. La correlación de Pearson y otros indicadores estadísticos se utilizan para cuantificar la fuerza y dirección de dicha relación, lo que facilita la evaluación de la **coherencia técnica y territorial** de los datos (Rodgers, 1988).

De manera complementaria, el estudio se considera **aplicado** porque emplea principios y herramientas de la **ingeniería de datos y la ciencia de datos** (Provost, 2013) con fines prácticos: la validación, integración y visualización de información pública sobre conectividad en Colombia. A través del desarrollo de un proceso **ETL (Extract, Transform, Load)** automatizado y reproducible, se busca fortalecer la confiabilidad de los indicadores públicos y generar un **modelo técnico de evaluación** que pueda ser reutilizado por entidades gubernamentales, investigadores y organismos de control.

En síntesis, la investigación combina un enfoque **cuantitativo, descriptivo, correlacional, no experimental y longitudinal**, apoyado en herramientas computacionales para la integración y análisis de grandes volúmenes de datos. Este enfoque permite no solo identificar las brechas entre la información oficial y la empírica, sino también construir evidencia verificable que respalde la formulación de políticas públicas basadas en datos,

orientadas al cierre de la brecha digital y la mejora continua de la calidad del servicio de Internet en Colombia.

3.2. Diseño metodológico

El diseño metodológico de esta investigación se estructura bajo un **modelo de análisis reproducible y automatizado**, orientado a la **verificación de la coherencia de los datos públicos del sector TIC en Colombia** mediante técnicas de **ciencia de datos y analítica computacional**.

Su propósito es comparar de manera sistemática los reportes oficiales del **Ministerio de Tecnologías de la Información y las Comunicaciones (MinTIC)** —contenidos en los *Boletines Trimestrales del Sector TIC*— con las mediciones empíricas publicadas por **Ookla Speedtest Open Data**, determinando el grado de correspondencia entre ambos conjuntos de información y construyendo el **Índice Nacional de Coherencia TIC (INCTIC)** como indicador principal del estudio.

Desde el punto de vista metodológico, el diseño se fundamenta en el paradigma de la **reproducibilidad científica** (Stodden, 2014), que busca garantizar que los resultados puedan ser verificados y replicados por otros investigadores. Para ello, se implementó un proceso **ETL (Extract, Transform, Load)** completamente automatizado en lenguaje **Python**, ejecutado en el entorno colaborativo **Google Colab**, con persistencia de datos en formato **DuckDB**, un motor analítico embebido de alto rendimiento (Raasveldt M. &., 2019).

Esta arquitectura permite gestionar grandes volúmenes de información con eficiencia y asegurar la trazabilidad de cada etapa del flujo de datos.

El diseño metodológico se organizó en **tres fases principales**, que abarcan desde la recolección y normalización de datos hasta el análisis estadístico y la obtención del índice de coherencia:

1. **Fase 1 – Extracción y limpieza de datos:**
   En esta etapa se desarrolló un script automatizado para realizar *web scraping* sobre el portal colombiatic.mintic.gov.co, detectando los enlaces de los boletines disponibles y descargando los archivos Excel correspondientes. Cada hoja fue procesada e integrada a la base **colombiatic.duckdb**, donde se registró su estructura y volumen de información en la tabla de control control_cargue. Paralelamente, se descargaron los datasets en formato Parquet de **Ookla Open Data**, filtrando únicamente los registros de Colombia y agregando las mediciones a nivel de municipio y departamento.
2. **Fase 2 – Integración y persistencia incremental:**
   Esta fase consistió en la creación de dos bases de datos analíticas (colombiatic.duckdb y ookla_colombia.duckdb), diseñadas para soportar actualizaciones periódicas y evitar la duplicación de registros.

   Se aplicaron rutinas de validación estructural, detección de valores atípicos (*outliers*) y normalización de nombres geográficos mediante la librería *unidecode*, garantizando la compatibilidad de los datos entre ambas fuentes.
3. **Fase 3 – Análisis y validación cruzada:**

Una vez consolidados los conjuntos de datos, se llevó a cabo la unión entre los registros de ColombiaTIC y Ookla, emparejando los municipios y departamentos mediante coincidencia exacta.

A partir de las coincidencias válidas, se calcularon los promedios de velocidad de subida y bajada por región, generando el **Índice Nacional de Coherencia TIC (INCTIC)**.

Finalmente, se aplicaron métodos estadísticos como el **coeficiente de correlación de Pearson** (Rodgers, 1988) para evaluar la relación lineal entre ambas fuentes, junto con análisis de dispersión y medidas de tendencia central para identificar patrones o discrepancias territoriales.

Este diseño metodológico combina herramientas de **data engineering**, **estadística aplicada** y **analítica reproducible**, siguiendo un flujo secuencial de procesamiento: **Extracción → Transformación → Validación → Análisis → Visualización.** Cada ejecución genera un registro automatizado en el archivo de log (colombiatic_etl.log), que documenta el número de reportes procesados, la fecha de actualización, las tablas creadas y el tamaño de la base resultante, asegurando la trazabilidad completa del proceso.

En términos de integridad científica, el diseño metodológico propuesto cumple con los principios de la **investigación reproducible y abierta**, al emplear tecnologías de código libre, bases de datos locales de libre acceso y scripts documentados que pueden ser ejecutados nuevamente por terceros.

Además, este diseño promueve la eficiencia analítica, la sostenibilidad del flujo de actualización trimestral y la transparencia en el tratamiento de los datos públicos, alineándose con los lineamientos de **Gobierno Digital** y **datos abiertos** definidos por el Estado colombiano.

3.3. Fuentes de información

El desarrollo del presente trabajo de investigación se fundamenta en el uso de **fuentes secundarias de datos abiertos**, caracterizadas por su disponibilidad pública, actualización periódica y relevancia para la medición de la calidad del servicio de Internet fijo en Colombia.

Estas fuentes se seleccionaron considerando su **validez institucional, cobertura geográfica nacional y complementariedad metodológica**, de modo que la información oficial emitida por el Estado pueda contrastarse con las mediciones empíricas obtenidas de una plataforma independiente de alcance global.

Las dos fuentes principales utilizadas son:

1. **ColombiaTIC – Boletines Trimestrales del Sector TIC**, publicados por el **Ministerio de Tecnologías de la Información y las Comunicaciones (MinTIC)**.
2. **Ookla Speedtest Open Data**, una base empírica de mediciones de velocidad de conexión a Internet obtenidas directamente de los usuarios finales.

Estas fuentes constituyen la base de la comparación que da origen al **Índice Nacional de Coherencia TIC (INCTIC)**, cuya finalidad es medir la correspondencia entre las velocidades reportadas oficialmente y las registradas empíricamente en campo.

3.3.1. Fuente oficial: ColombiaTIC (Ministerio TIC)

Los **Boletines Trimestrales del Sector TIC** representan la **principal fuente oficial de información sobre conectividad en Colombia**, siendo el resultado de los reportes enviados por los operadores de telecomunicaciones al MinTIC y publicados posterior a una consolidación realizada. A continuación, se presenta el pantallazo de la página oficial en donde se publican estos reportes:

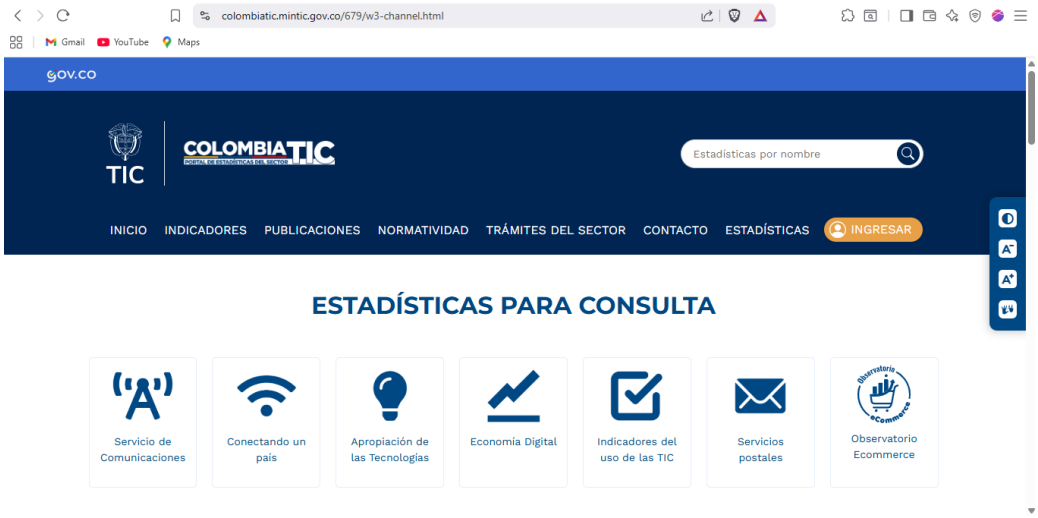

Ilustración 1: Página principal de ColombiaTIC[2]

En esta página se localiza el banner en donde se disponibilizan los reportes:

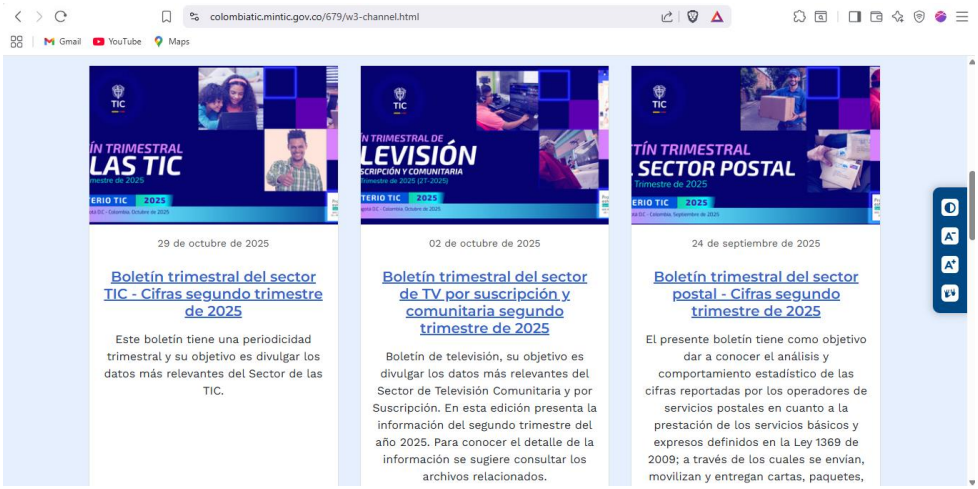

Ilustración 2: Pantallazo del banner en donde se localizan los reportes disponibles.[3]

Una vez ingresado sobre el reporte correspondiente, se presenta un mayor detalle de la información publicada como se observa en la imagen a continuación:

[2] URLColombiaTIC: https://colombiatic.mintic.gov.co/679/w3-channel.html
[3] Para el corte del documento se había procesado todo el flujo de información hasta el primer trimestre del 2025, en esta imagen se observa que desde el 29 de octubre se publicó el segundo trimestre.

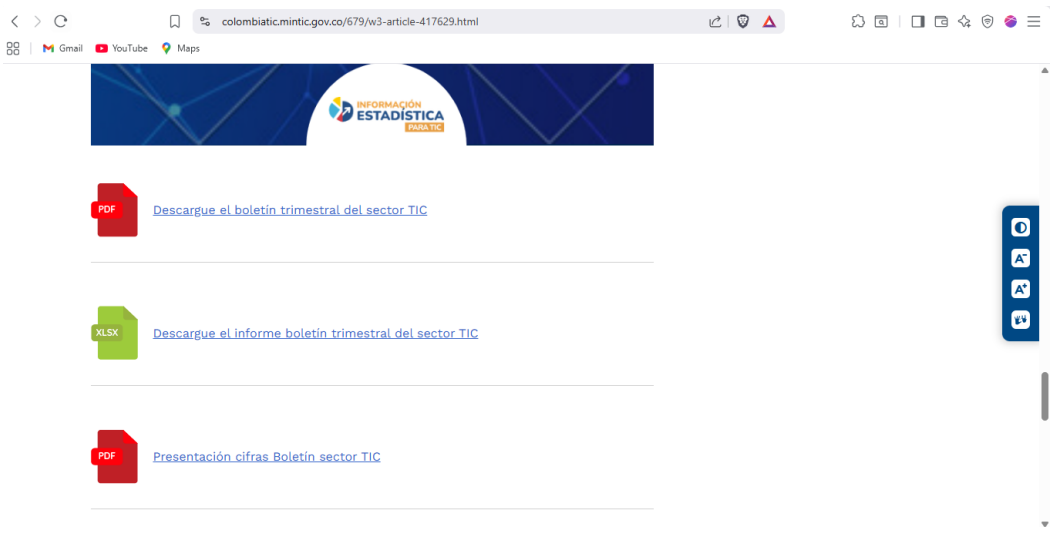

Ilustración 3: Pantallazo de los reportes disponibles de internet fijo.

El notebook encargado de scrapear la página localiza el archivo en Excel observado en la imagen anterior y lo localiza en la carpeta correspondiente del drive utilizado en el proyecto. Estos boletines incluyen estadísticas desagregadas por tipo de servicio (Internet fijo, móvil, televisión y postal), además de variables como número de accesos, velocidades contratadas, tecnología empleada (fibra, xDSL, cable, inalámbrico, entre otros) y cobertura geográfica a nivel de municipio y departamento. A continuación, se presenta el pantallazo del drive utilizado para localizar los reportes trimestrales que serán utilizadas para el análisis del estudio:

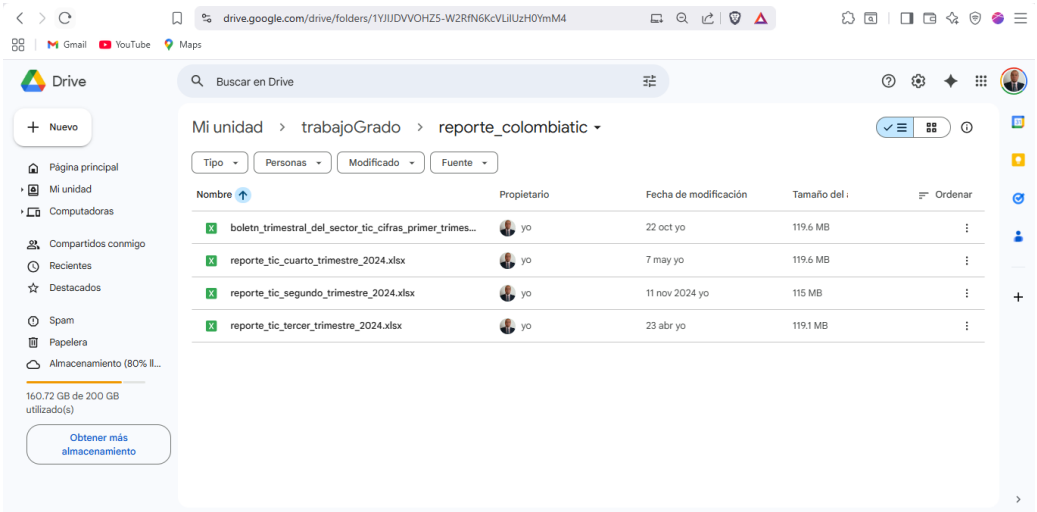

Ilustración 4: Drive utilizado para localizar los reportes trimestrales descargados de ColombiaTIC.

Cada boletín se publica en formato **Excel (.xlsx)** y contiene entre 15 y 20 hojas con diferentes desagregaciones. A continuación, se presenta un pantallazo del reporte del último archivo procesado:

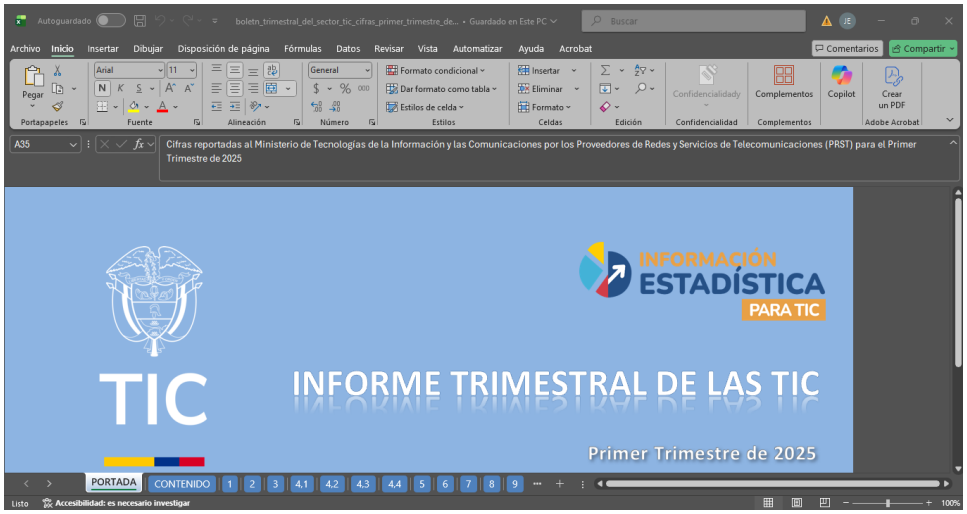

Ilustración 5: Pantallazo de uno de los reportes trimestrales descargados de ColombiaTIC.

De todas ellas, la **hoja 4.1**, titulada *"Accesos fijos a Internet por municipio, proveedor y velocidad"*, es la que concentra los indicadores de **velocidad de subida y bajada** asociados a los proveedores y regiones del país, por lo que fue seleccionada como base central para el análisis del presente estudio. Se presenta un pantallazo de dicha hoja del archivo:

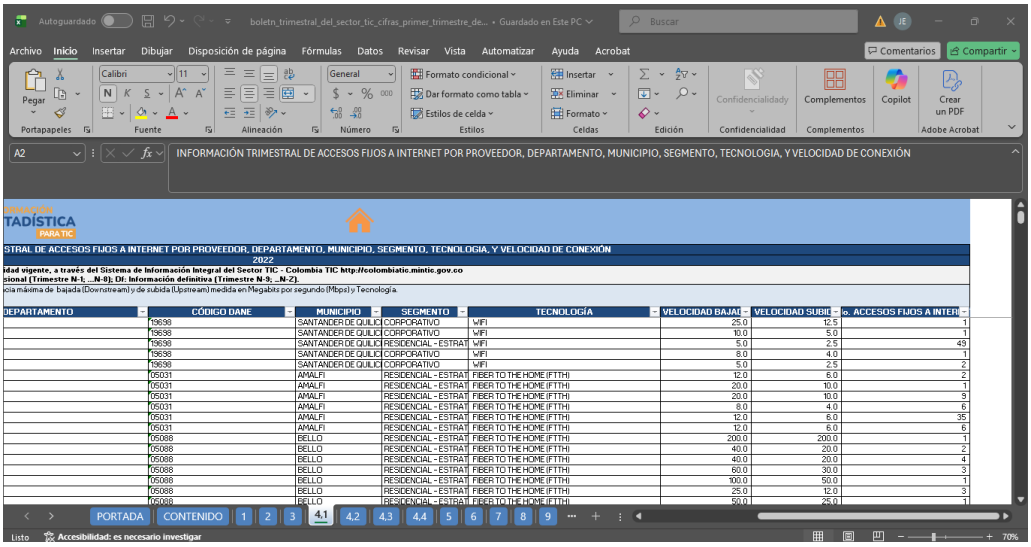

Ilustración 6: Hoja 4.1 del reporte trimestral utilizada para el estudio.

A nivel técnico, los archivos fueron descargados automáticamente mediante un script de **web scraping** desarrollado en Python, utilizando la librería requests_html para explorar los enlaces activos dentro del portal oficial colombiatic.mintic.gov.co.

Posteriormente, cada archivo fue procesado y almacenado en una base de datos **DuckDB**, donde cada hoja se convirtió en una tabla independiente.

El proceso incluyó además la creación de una tabla de control (control_cargue), encargada de registrar el nombre del archivo, número de hojas cargadas, cantidad de filas procesadas, fecha de ejecución y estado del cargue (OK o con error).

Esta estructura permitió llevar una trazabilidad precisa del proceso de integración y garantizar la integridad de los datos.

Los reportes oficiales del MinTIC fueron utilizados como **referente institucional**, ya que representan la información declarada por los operadores y utilizada por el Gobierno Nacional para la planificación de políticas públicas, seguimiento a metas de conectividad y toma de decisiones en materia de infraestructura digital (ColombiaTIC, 2024).

3.3.2. Fuente empírica: Ookla Speedtest Open Data

La segunda fuente corresponde a **Ookla Speedtest Open Data**, una plataforma internacional que recopila millones de mediciones realizadas por usuarios de Internet alrededor del mundo. A continuación, se presenta pantallazo de la página:

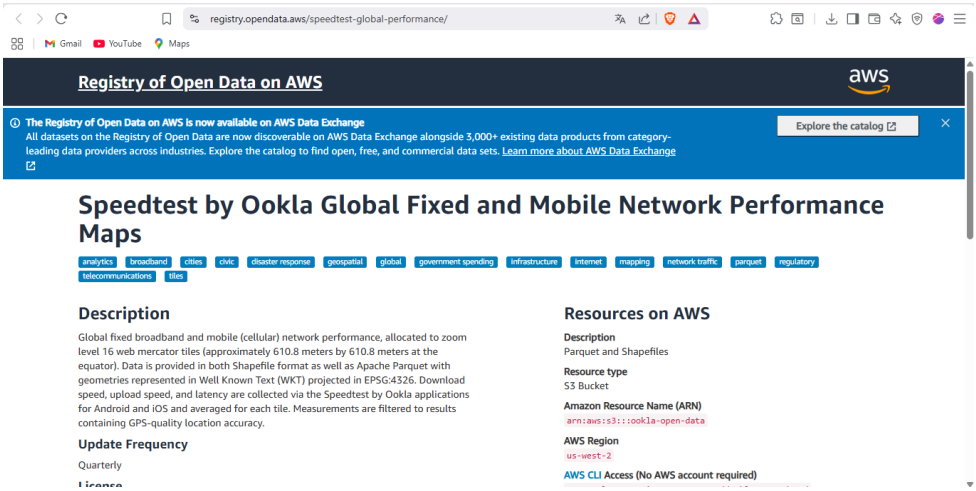

Ilustración 7: Segunda fuente de información de la investigación.[4]

Cada medición reporta la velocidad de **descarga (download)**, **subida (upload)** y **latencia (ping)**, junto con información geográfica asociada al punto de conexión. Estos datos se publican trimestralmente en formato **Parquet**, disponibles en el repositorio abierto de Ookla a través de data.openspeedtest.net, bajo licencias de acceso público que permiten su reutilización con fines de investigación y análisis estadístico.

En el contexto de esta investigación, se descargaron los archivos correspondientes a los períodos entre **2019 y 2025**, filtrando únicamente los registros asociados al territorio colombiano mediante el uso de la librería duckdb y expresiones SQL de filtrado geoespacial. A continuación, se presenta pantallazo de uno de los parquets descargados sobre el drive utilizado para la investigación:

[4] URLOokla: https://registry.opendata.aws/speedtest-global-performance/

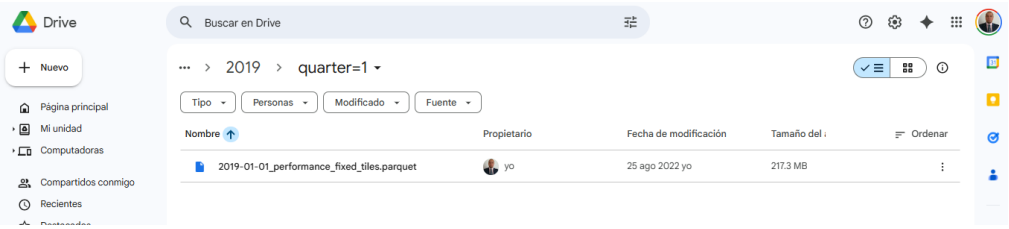

Ilustración 8: Pantallazo del parquet del primer trimestre del año 2019, localizado en el drive del proyecto.

A continuación, se realizó un proceso de decodificación de las coordenadas contenidas en el campo quadkey —una codificación espacial en cuadrantes— utilizando la librería mercantile para convertirlas en valores de **latitud y longitud**.

Posteriormente, los datos se agregaron a nivel de **municipio y departamento**, calculando los promedios trimestrales de velocidad de subida y bajada, así como la latencia promedio. El conjunto de datos procesado fue almacenado en la base ookla_colombia.duckdb, diseñada para mantener una estructura homogénea y compatible con la información proveniente de ColombiaTIC.

El valor de esta fuente radica en que constituye una **medición empírica directa de la experiencia del usuario**, reflejando el rendimiento real de las redes de telecomunicaciones, independiente de los reportes de los operadores. Su carácter abierto, granular y globalmente reconocido otorga confiabilidad y objetividad al análisis comparativo.

3.3.3. Complementariedad y validación de fuentes

La combinación de ambas fuentes permite abordar el problema de la **coherencia informativa desde dos perspectivas complementarias**:

- **ColombiaTIC** proporciona la visión institucional y declarativa de los operadores, en la que se basan las políticas públicas de conectividad.
- **Ookla Open Data** aporta la evidencia empírica y cuantitativa sobre la calidad del servicio efectivamente experimentado por los usuarios.

Esta complementariedad garantiza una visión integral y objetiva del fenómeno de la conectividad en Colombia.

El análisis cruzado entre ambas fuentes no busca cuestionar los reportes oficiales, sino **verificar su consistencia**, identificar posibles sesgos o diferencias regionales, y generar información técnica que contribuya a la transparencia y a la mejora continua en la gestión de los datos del sector TIC.

De esta manera, las fuentes de información empleadas en esta investigación no solo cumplen con los principios de **datos abiertos** y **reproducibilidad científica**, sino que también se convierten en el pilar sobre el cual se construye el **Índice Nacional de**

Coherencia TIC (INCTIC), contribuyendo a la consolidación de una política pública basada en evidencia.

3.4. Proceso de integración de datos

El proceso de integración de datos constituye el **núcleo metodológico del proyecto**, ya que garantiza la consolidación, limpieza y transformación de la información proveniente de las dos fuentes principales: los boletines del **Ministerio de Tecnologías de la Información y las Comunicaciones (MinTIC)** y los registros empíricos de **Ookla Speedtest Open Data**.

Este proceso fue diseñado bajo un enfoque **ETL (Extract, Transform, Load)** reproducible y auditable, asegurando la coherencia técnica, la trazabilidad temporal y la replicabilidad de los resultados en futuras actualizaciones trimestrales.

El flujo metodológico se desarrolló en tres cuadernos o *notebooks* principales implementados en **Google Colab**, cada uno con una función específica dentro del proceso de ingeniería de datos:

1. **Notebook 1 – Extracción y consolidación de datos ColombiaTIC**
2. **Notebook 2 – Extracción y procesamiento de datos Ookla**
3. **Notebook 3 – Cruce de bases y construcción del Índice Nacional de Coherencia TIC (INCTIC)**

La representación en alto nivel del proceso se presenta a continuación:

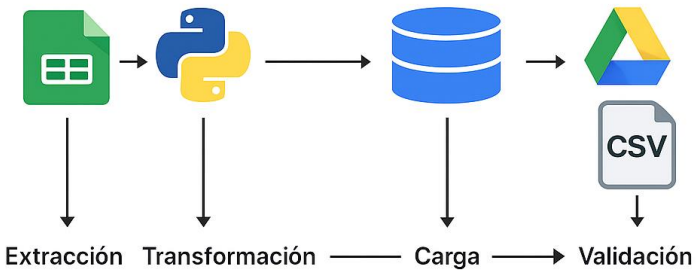

Ilustración 9: Flujo metodológico - ETL.

A continuación, se describe detalladamente cada etapa y las operaciones ejecutadas en el flujo ETL.

3.4.1. Etapa de extracción (Extract)

En esta primera fase, el objetivo fue **automatizar la recopilación de datos** de las fuentes oficiales y empíricas, asegurando que los reportes más recientes fueran identificados, descargados y versionados correctamente.

a) Fuente ColombiaTIC
Se implementó un script de *web scraping* en Python empleando las librerías requests_html, tqdm y re para explorar los canales de publicación del portal colombiatic.mintic.gov.co.

El algoritmo detecta los enlaces activos de los **boletines trimestrales** y descarga automáticamente los archivos Excel disponibles, almacenándolos en una carpeta de trabajo en Google Drive (/trabajoGrado/reporte_colombiatic).

Durante la descarga, cada archivo se **valida por nombre, tamaño y tipo de contenido**, con el fin de evitar duplicados o errores de integridad. Posteriormente, los archivos se trasladan a una carpeta definitiva y se registran en una **tabla de control** denominada control_cargue, dentro de la base **colombiatic.duckdb**.

Esta tabla contiene los campos:

archivo, hojas_cargadas, filas_totales, fecha_cargue y estado, lo que permite mantener un historial completo de las cargas y detectar automáticamente si un reporte ha sido previamente procesado.

b) Fuente Ookla Speedtest
En paralelo, los datos de Ookla fueron descargados desde el repositorio **Open Data** en formato Parquet, utilizando la línea de comandos de **AWS CLI** y comandos SQL ejecutados sobre **DuckDB**.

Posteriormente, los archivos fueron convertidos en una estructura unificada mediante la decodificación del campo quadkey —que representa la ubicación geográfica en mosaicos de Bing Maps— a coordenadas de **latitud** y **longitud**, utilizando la librería mercantile. Esto permitió asignar cada registro a su **municipio y departamento** correspondiente, mediante una operación de *spatial join* con una base de límites administrativos.

3.4.2. Etapa de transformación (Transform)

En esta fase se aplicaron procesos de **limpieza, normalización y homogeneización** de las variables para permitir la integración precisa entre las fuentes.

Las operaciones más relevantes incluyeron:

- **Estandarización de nombres geográficos:**
  Se utilizó la librería unidecode para eliminar tildes, mayúsculas, caracteres especiales y variaciones ortográficas en los nombres de los municipios y departamentos de ambas bases, facilitando la coincidencia durante el cruce.
- **Conversión de tipos de datos:**
  Todas las columnas numéricas fueron convertidas a formato DOUBLE, mientras que los identificadores alfanuméricos (como códigos DANE o nombres de proveedor) se mantuvieron en VARCHAR.

Se realizó un control de integridad sobre los valores nulos y se eliminaron filas con errores de formato.

- **Detección y tratamiento de valores atípicos (outliers):**
  Para garantizar la representatividad de los cálculos de velocidad, se filtraron valores extremos que superaban rangos técnicos razonables (por ejemplo, velocidades superiores a 3.000 Mbps), conservando únicamente los datos realistas en función del mercado colombiano.

  Esta limpieza fue fundamental para obtener un índice de coherencia estable y estadísticamente robusto.
- **Consolidación por periodo:**
  Dado que los boletines de ColombiaTIC se publican trimestralmente, se generó un esquema temporal con las variables AÑO y TRIMESTRE, asegurando la consistencia de las fechas con los registros Ookla correspondientes al mismo período.

El resultado de esta etapa fue una **tabla unificada de datos limpios** denominada consolidado_tic_4_1_limpia, que representa la base consolidada de los reportes oficiales, y una base agregada ookla_geo, que resume los valores medios de velocidad de cada municipio y trimestre.

3.4.3. Etapa de carga y persistencia (Load)

En la etapa final del proceso ETL, los datos transformados fueron cargados en sus respectivas bases **DuckDB** dentro de Google Drive, garantizando la persistencia de la información y su disponibilidad para posteriores análisis.

Cada ejecución del pipeline genera automáticamente:

- Una actualización de las tablas de control (control_cargue y control_etl).
- Una copia de respaldo (*.backup.duckdb) para prevenir pérdida de información.
- Un archivo de **log detallado** (colombiatic_etl.log) que documenta la hora de inicio, duración del proceso, número de reportes procesados y registros incorporados.

Con estas rutinas se asegura que, en futuras ejecuciones, el sistema solo procese **nuevos boletines o reportes modificados**, evitando reprocesamientos innecesarias y optimizando los recursos computacionales.

Asimismo, la base consolidada se exporta de forma incremental en formato **CSV** y **Parquet**, lo que facilita su uso tanto en herramientas analíticas externas (Power BI, Tableau, R) como en otros scripts Python para los análisis de correlación, visualización y modelado.

3.4.4. Trazabilidad y reproducibilidad

El proceso completo fue diseñado bajo criterios de **auditoría, trazabilidad y reproducibilidad**, garantizando que cada transformación pueda ser verificada. Esto se logra mediante:

- **Logs de ejecución persistentes** en cada notebook.
- **Control de versiones de base de datos** en Google Drive.
- **Identificación única de cada archivo** procesado, mediante timestamp y hash del contenido.

De esta forma, la integración de datos no solo permite consolidar información de múltiples fuentes, sino también mantener un ecosistema analítico confiable y replicable, alineado con los principios de **Gobierno Digital**, **datos abiertos** y **reproducibilidad científica** (Stodden, 2014).

3.5. Modelo de análisis

El desarrollo de esta investigación requirió la integración de **entornos colaborativos, lenguajes de programación y motores de análisis de datos**, seleccionados por su compatibilidad con procesos reproducibles, eficiencia en la manipulación de grandes volúmenes de información y disponibilidad en plataformas abiertas.

Todas las herramientas empleadas son de **código libre o de acceso gratuito**, lo que garantiza la transparencia y replicabilidad del trabajo, en concordancia con los principios de **datos abiertos y ciencia abierta** promovidos por el Gobierno Digital de Colombia.

La siguiente tabla presenta las principales herramientas y tecnologías utilizadas, junto con su función dentro del proceso metodológico y la justificación de su selección:

Herramienta / Tecnología	Tipo / Entorno	Función principal en el proyecto	Justificación de uso
Python 3.12	Lenguaje de programación	Desarrollo del flujo ETL (Extracción, Transformación y Carga) y análisis estadístico.	Python ofrece una amplia colección de librerías para ciencia de datos (pandas, numpy, matplotlib) y permite reproducir procesos analíticos de forma modular y documentada (Provost, 2013)
Google Colab	Entorno colaborativo en la nube	Ejecución de notebooks, conexión directa a Google Drive y control de versiones.	Permite trabajar sin configuración local, facilitando la ejecución reproducible de scripts y la colaboración académica.

Herramienta / Tecnología	Tipo / Entorno	Función principal en el proyecto	Justificación de uso
DuckDB	Motor de base de datos analítica embebida	Almacenamiento, consulta y consolidación de grandes volúmenes de datos sin necesidad de servidores.	Su estructura columnar y ejecución en memoria permite realizar análisis OLAP con rapidez y bajo consumo de recursos (Raasveldt M. &., 2019).
Pandas / NumPy	Librerías Python para análisis de datos	Limpieza, transformación y análisis de datos tabulares.	Herramientas estándar de manipulación eficiente de estructuras de datos en ciencia de datos.
Requests_HTML / BeautifulSoup4	Librerías Python para web scraping	Descarga automatizada de boletines trimestrales del portal ColombiaTIC.	Permiten acceder a datos públicos reportados en la web garantizando integridad y automatización del proceso de extracción.
Unidecode / Re	Librerías de normalización de texto	Estandarización de nombres de municipios y departamentos, eliminación de acentos y caracteres especiales.	Facilitan la homologación de campos alfanuméricos y la mejora de coincidencias durante el cruce de bases.
TQDM / Logging	Librerías de seguimiento y auditoría	Monitoreo del progreso de los procesos ETL y registro persistente de errores y eventos.	Garantizan la trazabilidad del proceso y facilitan la depuración y documentación de cada ejecución.
Matplotlib / Seaborn / Plotly	Librerías de visualización	Creación de gráficos de dispersión, histogramas y mapas de correlación.	Permiten representar gráficamente los resultados y facilitar la interpretación visual de los datos.
GeoPandas / Mercantile	Librerías de análisis geoespacial	Conversión de coordenadas quadkey a latitud/longitud y unión espacial con límites municipales.	Fundamentales para el procesamiento de datos georreferenciados de Ookla y su asociación con territorios colombianos.
CSV / Parquet	Formatos de intercambio de datos	Exportación y almacenamiento de resultados intermedios y finales.	Los formatos abiertos garantizan compatibilidad con múltiples plataformas y facilitan la reutilización de los resultados.
GitHub *(versión de respaldo)*	Repositorio de control de versiones	Almacenamiento de notebooks, scripts y versiones de código.	Facilita la transparencia y el versionamiento del código fuente del proyecto.

Tabla 1: Herramientas y tecnologías empleadas en el proyecto.

Infraestructura de ejecución

El entorno de ejecución fue configurado en **Google Colab**, el cual ofrece una máquina virtual temporal con entorno Linux, procesador Intel Xeon, memoria RAM de 12 GB y acceso directo a **Google Drive** para el almacenamiento persistente de datos y respaldos de las bases **DuckDB**.

Esta infraestructura permitió ejecutar cargas intensivas de hasta **1.8 millones de registros** en memoria sin requerir infraestructura local ni servidores externos.

Cada notebook fue diseñado para ejecutarse de manera **modular**, permitiendo su reutilización individual según el objetivo:

- notebook_1_colombiatic.ipynb → *extracción y carga de boletines oficiales*,
- notebook_2_ookla.ipynb → *procesamiento de mediciones empíricas*,
- notebook_3_cruce_inctic.ipynb → *análisis comparativo y generación del índice INCTIC*.

Principios técnicos aplicados

El uso de estas herramientas se orientó por tres principios fundamentales:

1. **Reproducibilidad:**
   Todo el flujo analítico puede ejecutarse nuevamente en cualquier entorno Python compatible, generando los mismos resultados a partir de las mismas fuentes.
2. **Automatización:**
   Los procesos pueden llegar a estructurarse para detectar nuevos boletines o datasets y procesarlos sin intervención manual, mediante triggers y registros de control ya establecidos en los cuadernos principales del proceso. Los notebooks al ejecutarse de forma manual permiten realizar todo el proceso descrito, a nivel del corte del documento de investigación no se presentaron los triggers en Google para que cada mes o cada trimestre inicien una ejecución periódica de los cuadernos y permita descargar los reportes al momento de estar disponibles y continuar con todo el flujo del proceso realizado.
3. **Trazabilidad y transparencia:**

   Cada ejecución genera un archivo de log y una copia incremental de la base consolidada, permitiendo auditar los pasos de extracción y transformación.

En conjunto, las tecnologías empleadas permitieron desarrollar una solución robusta, flexible y alineada con los estándares internacionales de ciencia de datos y gobierno digital, garantizando la **validez técnica** y la **eficiencia metodológica** del proyecto.

3.6. Arquitectura general del sistema y flujo de trabajo

La arquitectura general del sistema desarrollado para esta investigación se fundamenta en un **enfoque modular, reproducible y orientado a procesos ETL (Extract, Transform, Load)**, diseñado para integrar múltiples fuentes de datos abiertos y generar resultados verificables de manera automatizada.

El flujo metodológico responde a los principios de la **ingeniería de datos aplicada a la gestión pública**, asegurando la trazabilidad completa de cada paso —desde la extracción de los reportes oficiales y empíricos hasta la obtención del *Índice Nacional de Coherencia TIC (INCTIC)*.

La arquitectura, desarrollada en Python y soportada sobre **Google Drive** y **DuckDB**, permite la automatización de los procesos de carga, limpieza y validación cruzada, garantizando la reproducibilidad de los resultados y la trazabilidad de cada fuente de información. A continuación, se presenta una arquitectura general del flujo:

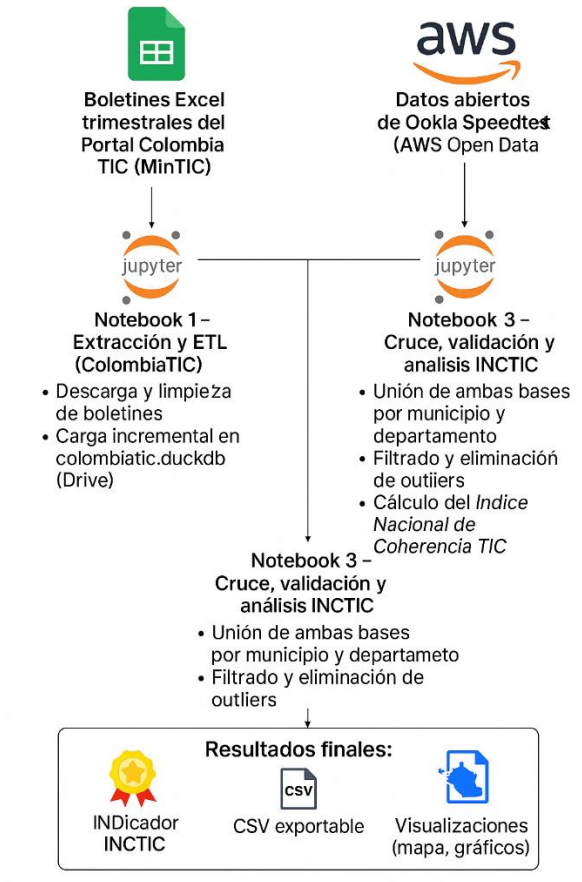

Ilustración 10: Arquitectura general del sistema.

La Ilustración 10 muestra el flujo completo del sistema propuesto, el cual se estructura en tres etapas principales y culmina con la exportación de un archivo **CSV** consolidado, acompañado de visualizaciones en mapas y gráficos que facilitan la interpretación del nivel de coherencia entre las fuentes oficiales y las mediciones empíricas.

3.6.1. Estructura general del sistema

La arquitectura del sistema se compone de cuatro capas principales:

1. **Capa de entrada (Input Layer):**

   Encargada de la **extracción automatizada** de los archivos fuente desde los portales oficiales.

   - **Fuente oficial:** Boletines trimestrales del sector TIC del MinTIC (archivos Excel).
   - **Fuente empírica:** Dataset Ookla Speedtest Open Data (archivos Parquet). Los archivos se descargan y almacenan temporalmente en la carpeta /temp_colombiatic y posteriormente se mueven a /reporte_colombiatic dentro del directorio de trabajo en Google Drive.

2. **Capa de procesamiento (Processing Layer):**

   Desarrollada en **Python sobre Google Colab**, donde se ejecutan los tres *notebooks principales* del proyecto:

   - notebook_1_colombiatic.ipynb: extracción y normalización de los boletines del MinTIC.
   - notebook_2_ookla.ipynb: decodificación de coordenadas y agregación geográfica de los datos Ookla.
   - notebook_3_cruce_inctic.ipynb: integración de bases, cálculo del índice INCTIC y análisis estadístico.

   Esta capa realiza los procesos de **limpieza, transformación, estandarización de nombres, detección de valores atípicos** y **homogeneización de estructuras** entre ambas fuentes.

3. **Capa de almacenamiento (Storage Layer):**

   Los datos procesados son almacenados de forma estructurada en bases de datos **DuckDB**, ubicadas en Google Drive:

   - /colombiatic_datos/colombiatic.duckdb
   - /ookla_datos/ookla_colombia.duckdb
     Cada base contiene tablas intermedias y una tabla consolidada final (consolidado_tic_4_1_filtrado), que permite su consulta directa mediante SQL sin necesidad de servidores dedicados.

   Se mantiene además una **tabla de control (control_cargue)** y un archivo de **log persistente (colombiatic_etl.log)** que registran la trazabilidad de los cargues, fechas, tamaño de los archivos y número de registros procesados.

4. **Capa de análisis y visualización (Analytics Layer):**

   En esta capa se realiza el **cruce entre las bases** y el cálculo del **Índice Nacional de Coherencia TIC (INCTIC)**, definido como la razón entre las velocidades reportadas por ColombiaTIC y las medidas por Ookla.

Los resultados se exportan en formato CSV y Parquet para su análisis en herramientas externas (Power BI, Tableau o Python).

Adicionalmente, se generan visualizaciones estadísticas (mapas, histogramas y correlaciones) que permiten interpretar la distribución y coherencia territorial de los datos.

3.6.2. Flujo de trabajo del sistema

El proceso completo se ejecuta siguiendo una secuencia lógica estructurada en seis fases principales:

1. **Inicio y verificación del entorno:**

   El notebook comprueba la existencia de las carpetas y bases de datos, montando automáticamente el entorno de Google Drive y asegurando la conexión a las bases **DuckDB** locales y en la nube.

2. **Extracción de reportes disponibles:**

   El sistema detecta automáticamente los enlaces activos de los boletines más recientes publicados en el portal MinTIC, descarga los archivos y los almacena en las carpetas correspondientes, registrando la ejecución en la tabla de control.

3. **Transformación y depuración de datos:**

   Se procesan las hojas de cálculo relevantes (especialmente la *4.1*, correspondiente a accesos fijos a Internet) aplicando limpieza, normalización de texto, y eliminación de valores inconsistentes o duplicados.

4. **Integración con datos Ookla:**

   Los registros Ookla, previamente filtrados por país y municipio, se cruzan con los datos oficiales en función de las variables **departamento**, **municipio**, **año** y **trimestre**.
   Este cruce genera un conjunto de datos integrados, con las columnas: vel_bajada_tic, vel_subida_tic, vel_bajada_ookla, vel_subida_ookla.

5. **Cálculo del índice INCTIC:**

   Se calcula el **Índice Nacional de Coherencia TIC (INCTIC)** a nivel municipal y departamental mediante la fórmula:

$$INCTIC = \frac{\text{Velocidad medida (Ookla)}}{\text{Velocidad reportada (ColombiaTIC)}}$$

Un valor del índice cercano a 1 refleja coherencia entre las fuentes; valores significativamente mayores o menores indican divergencias o posibles inconsistencias.

6. **Visualización, exportación y cierre:**

   Finalmente, se generan gráficos de correlación y distribución del índice, y los resultados se exportan a archivos CSV y Parquet para su análisis posterior o incorporación a dashboards.

   El notebook registra en el log la fecha, duración de la ejecución y tamaño final de la base consolidada.

3.6.3. Descripción del flujo general (representación conceptual)

El flujo de arquitectura del sistema puede representarse conceptualmente de la siguiente forma:

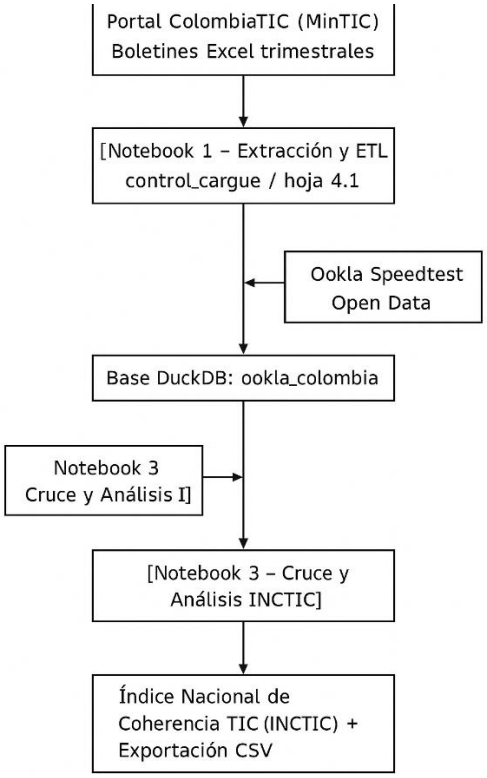

Ilustración 11: flujo de arquitectura del sistema.

Esta representación muestra de forma general como se da el flujo de proceso.

3.6.4. Características técnicas del diseño

- **Escalabilidad:** La arquitectura permite incorporar nuevos trimestres de datos sin reconfigurar el flujo.

- **Persistencia:** Las bases DuckDB se mantienen en Drive y se actualizan incrementalmente.
- **Auditoría:** Cada ejecución queda registrada en un log detallado con sello de tiempo.
- **Reproducibilidad:** Los notebooks pueden ejecutarse en cualquier entorno Colab, reproduciendo los resultados.
- **Transparencia:** El sistema emplea únicamente fuentes de datos abiertos verificables por terceros.

3.6.5. Beneficios del enfoque adoptado

- **Estandarización de la información:** consolida los reportes oficiales en una estructura única, reduciendo la dispersión de archivos.
- **Comparabilidad temporal y territorial:** facilita la comparación intertrimestral y entre municipios.
- **Base para la toma de decisiones:** los resultados del índice INCTIC ofrecen un insumo técnico para políticas públicas de conectividad y auditorías de datos.
- **Replicabilidad:** la arquitectura modular permite replicar el proceso en otros países o servicios (móvil, televisión, etc.).

3.7. Metodología de análisis y validación del INCTIC

El **Índice Nacional de Coherencia TIC (INCTIC)** constituye el núcleo analítico de esta investigación y su propósito es **evaluar el grado de coherencia entre los datos oficiales reportados por el Ministerio de Tecnologías de la Información y las Comunicaciones (MinTIC)** —a través de los boletines del sector TIC— **y las mediciones empíricas de velocidad de Internet provenientes de la plataforma Ookla Speedtest Open Data**. El índice fue diseñado para cuantificar, de manera objetiva, las divergencias o coincidencias entre ambas fuentes, garantizando una base técnica reproducible para la verificación de la calidad de los datos públicos del sector.

3.7.1. Definición del índice

El INCTIC se define como la razón entre las velocidades empíricamente medidas (por Ookla) y las reportadas oficialmente (por ColombiaTIC), tanto en velocidad de **bajada** (*download*) como de **subida** (*upload*). Matemáticamente, se expresa de la siguiente forma:

$$INCTIC = \frac{V_{Ookla}}{V_{TIC}}$$

donde:

- V_{Ookla}: velocidad promedio empírica medida por los usuarios en la plataforma Ookla Speedtest Open Data.
- V_{TIC}: velocidad promedio reportada oficialmente por los operadores en los boletines del sector TIC.

Un valor de **INCTIC ≈ 1** indica una **alta coherencia** entre ambas fuentes; valores **mayores a 1** reflejan **sobrestimación** en los datos empíricos o subregistro en los oficiales, mientras que valores **menores a 1** evidencian **discrepancias** donde los datos oficiales podrían sobrestimar la velocidad real percibida por los usuarios.

3.7.2. Proceso de cálculo y consolidación

El cálculo del INCTIC se desarrolló a través de un flujo automatizado en Python que integró los siguientes pasos metodológicos:

1. **Selección y estandarización de variables clave:**

   Se homogenizaron los campos *departamento*, *municipio*, *año* y *trimestre* para garantizar la compatibilidad entre las fuentes. Los nombres se normalizaron

aplicando la función unidecode() para eliminar tildes y caracteres especiales, asegurando coincidencias exactas entre los registros de ColombiaTIC y Ookla.

2. **Agregación y filtrado de datos:**

   Se calcularon promedios trimestrales de velocidad (subida y bajada) para cada municipio, excluyendo valores nulos y registros duplicados.
   Se aplicaron filtros de consistencia para descartar valores extremos (outliers) mediante el criterio de rango intercuartílico (IQR), manteniendo solo el 95 % central de los datos.

3. **Cruce entre fuentes:**

   Las bases depuradas se integraron empleando pandas.merge() con coincidencia exacta en las variables territoriales y temporales. El cruce generó una tabla consolidada de **1.054 municipios coincidentes**, con las columnas: vel_bajada_tic, vel_subida_tic, vel_bajada_ookla, vel_subida_ookla.

4. **Cálculo del índice individual y departamental:**

   Se calculó el INCTIC a nivel de municipio y se obtuvieron promedios departamentales.

$$INCTIC_{municipal} = \frac{Velocidad\ O}{Velocidad\ T} \quad \frac{okla_{municipal}}{IC_{municipal}}$$

$$INCTIC_{municipal} = \frac{Velocidad\ Ookla_{municipal}}{Velocidad\ TIC_{municipal}}$$

Luego se agruparon los valores por departamento para calcular la media, mediana y número de municipios coincidentes.

5. **Validación estadística:**

   Para evaluar la relación entre ambas fuentes, se aplicó el **coeficiente de correlación de Pearson (r)** para cada tipo de velocidad (subida y bajada).
   El resultado arrojó correlaciones bajas pero positivas ($r \approx 0.14$), lo que indica una relación débil y heterogénea, coherente con las diferencias territoriales observadas.

A continuación, se presenta un esquema del proceso:

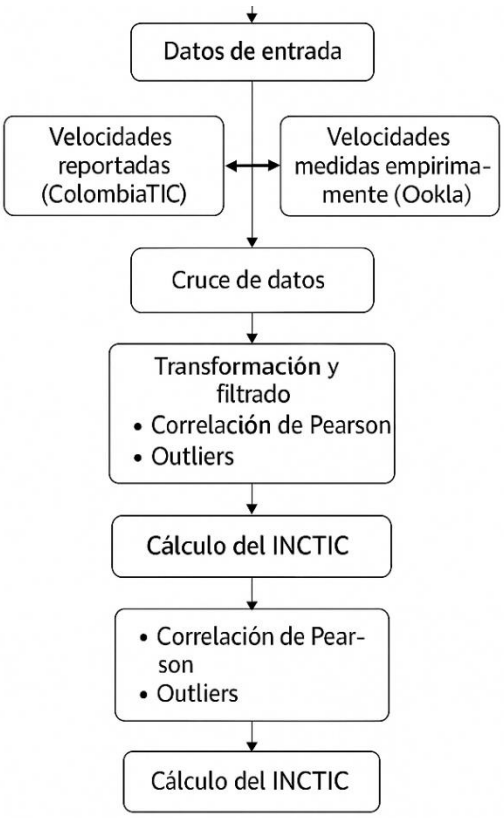

Ilustración 12: Esquema metodológico de cálculo y validación del INCTIC.

3.7.3. Criterios de validación y consistencia

El proceso de validación del índice incluyó tres niveles de control:

- **Consistencia interna:**

  Se verificó la ausencia de valores nulos, duplicados o negativos en las columnas de velocidad.

  Además, se comprobó que la suma de registros coincidentes correspondiera al total esperado por municipio y trimestre.

- **Validación estadística:**

  Se evaluaron los valores del índice mediante análisis de distribución (histogramas y boxplots) para garantizar la ausencia de sesgos extremos tras la depuración de outliers.

  La media del INCTIC se ubicó alrededor de **12.15**, mientras que la mediana fue **0.76**, evidenciando una asimetría causada por diferencias sustanciales en algunos territorios.

- **Validación geográfica:**

Se contrastaron los resultados departamentales con los patrones de cobertura reportados por el MinTIC, encontrando que las mayores divergencias se concentran en departamentos con baja densidad de infraestructura o cobertura rural extendida, como **Vaupés, Guainía y La Guajira**, los cuales registraron los valores más altos del índice.

3.7.4. Interpretación del índice y su utilidad

El **INCTIC** se convierte en una herramienta complementaria para la **auditoría técnica de la información pública del sector TIC**, al proporcionar evidencia empírica sobre la correspondencia entre las cifras reportadas por los operadores y las velocidades reales experimentadas por los usuarios.

Su interpretación permite:

- Detectar inconsistencias regionales entre los reportes oficiales y las mediciones independientes.
- Priorizar territorios donde la brecha entre la velocidad reportada y la medida supera los rangos esperados.
- Evaluar la eficacia de las políticas de conectividad y la focalización de inversiones públicas.
- Fortalecer la transparencia y confianza en los datos gubernamentales sobre infraestructura digital.

3.7.5. Herramientas y reproducibilidad

El proceso de cálculo y validación se implementó utilizando herramientas de código abierto, lo que garantiza su reproducibilidad:

Componente	Herramienta / Tecnología	Propósito
Lenguaje principal	**Python (pandas, numpy, matplotlib, seaborn)**	Análisis de datos, estadística y visualización
Motor de base de datos	**DuckDB**	Almacenamiento analítico y consultas SQL embebidas
Entorno de ejecución	**Google Colab**	Plataforma en la nube para ejecución reproducible
Control de versiones	**GitHub**	Almacenamiento y trazabilidad de notebooks
Validación estadística	**Scipy.stats, Pearson r**	Evaluación de correlación entre fuentes

Componente	Herramienta / Tecnología	Propósito
Exportación de resultados	**CSV / Parquet / Log TXT**	Generación de salidas estructuradas para análisis posterior

Tabla 2: Herramientas y reproducibilidad.

3.7.6. Conclusión metodológica

El enfoque metodológico adoptado garantiza una validación técnica y estadística robusta del **Índice Nacional de Coherencia TIC (INCTIC)**, al integrar fuentes heterogéneas en un modelo replicable y transparente.

Este procedimiento permite no solo verificar la coherencia de los datos públicos, sino también ofrecer una base sólida para la **toma de decisiones basada en evidencia**, orientada a mejorar la calidad del servicio de Internet y fortalecer la política pública de conectividad en Colombia.

Capítulo 4 – Resultados y Análisis

4.1. Base consolidada ColombiaTIC vs Ookla

La construcción de la base consolidada constituye el primer paso del análisis empírico del presente estudio, cuyo propósito fue integrar dos fuentes de información complementarias sobre la calidad del servicio de Internet fijo en Colombia. Estas fuentes fueron:

(i) los datos oficiales publicados por el *Ministerio de Tecnologías de la Información y las Comunicaciones (MinTIC)* a través de los *Boletines Trimestrales del Sector TIC* (ColombiaTIC, 2024), y

(ii) las mediciones abiertas de la plataforma *Ookla Speedtest Open Data* (LLC, 2025), que recogen información empírica reportada directamente por los usuarios sobre las velocidades efectivas de conexión en diferentes municipios del país.

El proceso de integración se desarrolló bajo un enfoque de *ciencia de datos reproducible*, utilizando técnicas ETL (Extract, Transform, Load) implementadas en Python y el motor analítico DuckDB.

Durante la etapa de extracción, se descargaron y consolidaron los reportes oficiales de ColombiaTIC correspondientes a los años 2024-2025, con énfasis en la hoja **4.1 "Accesos fijos a Internet"**, que contiene las variables de velocidad de subida y bajada, proveedor, segmento de usuario y localización geográfica (departamento y municipio).
En la etapa de transformación, se realizó un proceso de depuración exhaustivo para detectar valores faltantes, inconsistencias en los nombres geográficos y duplicados. Mediante normalización de texto (uso de la librería *unidecode*) y detección de encabezados automáticos, se logró estandarizar los campos territoriales y técnicos. El resultado fue una base depurada con **1.817.578 registros** y **55.982 registros únicos** después del filtrado y normalización, distribuidos en **1.037 municipios**.

Por su parte, la base de *Ookla Open Data* se construyó a partir de archivos en formato *parquet*, filtrando los registros correspondientes a Colombia y agregándolos a nivel municipal mediante funciones de agrupamiento en DuckDB. La estructura final incluyó **595.952 observaciones**, con campos estandarizados para velocidad promedio de subida (*upload*), velocidad promedio de bajada (*download*), latencia, número de dispositivos y cantidad de pruebas realizadas.

Una vez consolidadas ambas fuentes, se desarrolló una estructura de datos unificada en formato *DuckDB*, que permite realizar consultas SQL de alta eficiencia y mantener la trazabilidad completa de los datos procesados.

Esta base consolidada se constituye en el insumo principal para el análisis posterior de coherencia y correlación, pues posibilita la comparación directa entre los indicadores técnicos de las dos fuentes bajo una misma granularidad territorial y temporal (municipio y trimestre).

Finalmente, esta integración representa un avance metodológico en la validación de datos abiertos del sector TIC, al habilitar una infraestructura analítica que puede replicarse trimestralmente para futuras actualizaciones, asegurando la transparencia, consistencia y auditabilidad de la información utilizada para el cálculo del **Índice Nacional de Coherencia TIC (INCTIC)**.

4.2. Validación cruzada de datos

La validación cruzada constituye el núcleo metodológico del proceso de integración entre las fuentes **ColombiaTIC** y **Ookla Speedtest Open Data**, orientado a garantizar la coherencia técnica y estadística de la información analizada. Esta fase tuvo como objetivo comparar de manera sistemática las velocidades de Internet fijo-reportadas oficialmente por los operadores con las mediciones empíricas obtenidas a partir de los usuarios, a nivel municipal y trimestral.

El procedimiento comenzó con la **normalización de los identificadores territoriales**, ya que ambas fuentes presentaban diferencias en la denominación de departamentos y municipios. Para ello, se aplicaron procesos automáticos de limpieza y estandarización mediante expresiones regulares y la librería *Unidecode*, eliminando tildes, mayúsculas y caracteres especiales. Posteriormente, se creó una columna auxiliar con los nombres normalizados ("DEPARTAMENTO_N" y "MUNICIPIO_N"), lo que permitió realizar el emparejamiento exacto (*matching*) entre ambas bases.

Una vez homologadas las variables geográficas, se procedió a realizar un **cruce relacional optimizado** en *DuckDB*, con base en las columnas *departamento* y *municipio*, integrando los registros de velocidad de subida (*upload*) y bajada (*download*) de cada fuente. Para mejorar la precisión del emparejamiento, se utilizó un algoritmo de comparación fonética (basado en distancia de Levenshtein) para detectar coincidencias parciales en nombres de municipios, corrigiendo discrepancias menores de escritura (por ejemplo: *Santander de Quilichao* vs *Santander Quilichao*).

Como resultado, se obtuvo una base emparejada con **1.054 coincidencias exactas** entre municipios y departamentos, de un total de 1.120 municipios presentes en los registros de ColombiaTIC y 1.222 en los de Ookla. Este alto porcentaje de correspondencia evidencia la consistencia estructural entre ambas fuentes, validando la integridad del proceso ETL y la fiabilidad del modelo de emparejamiento implementado.

Durante la validación, se identificaron tres tipos de diferencias principales:

1. **Falta de registros coincidentes:** en municipios sin mediciones activas en Ookla o sin datos recientes en los reportes oficiales.
2. **Desalineación temporal:** causada por variaciones en las fechas de actualización de los boletines trimestrales del MinTIC frente al período de carga de la base Ookla.
3. **Discrepancias en magnitud de velocidad:** debidas a la metodología de medición (teórica en el caso oficial, empírica en el caso de Ookla).

A partir del cruce consolidado, se calcularon las velocidades promedio por municipio y trimestre, generando una estructura comparativa con las variables:

- **vel_subida_tic** y **vel_bajada_tic**, provenientes de ColombiaTIC, y
- **vel_subida_ookla** y **vel_bajada_ookla**, provenientes de las mediciones empíricas.

Esta integración permitió avanzar hacia el análisis de coherencia y correlación entre fuentes, constituyendo la base para el cálculo del **Índice Nacional de Coherencia TIC (INCTIC)**. Además, se guardó una versión intermedia en formato *CSV* en el repositorio de trabajo, con el fin de asegurar la trazabilidad de los resultados y posibilitar auditorías o reproducciones futuras del experimento.

A continuación, se presenta un resumen de correspondencias entre fuentes:

	Departamento	Municipios coincidentes	Vel. promedio TIC (Mbps)	Vel. promedio Ookla (Mbps)	Relación Ookla/TIC
29	VAUPES	3	4.021000	64.961667	16.156
0	AMAZONAS	8	3.873000	27.018250	6.976
14	GUAINIA	5	9.519400	20.645600	2.169
3	ATLANTICO	22	11.642091	11.796364	1.013
23	QUINDIO	12	6.166667	5.354583	0.868
17	LA GUAJIRA	15	9.129267	6.326267	0.693
10	CESAR	23	9.416217	6.402043	0.680
11	CHOCO	15	15.281867	10.137133	0.663
2	ARAUCA	7	9.937571	5.993429	0.603
8	CASANARE	19	22.983158	13.570000	0.590
18	MAGDALENA	28	14.086250	7.569214	0.537
4	BOLIVAR	42	13.493548	6.376095	0.473
9	CAUCA	39	9.255667	4.274744	0.462
12	CORDOBA	30	19.988600	9.135933	0.457
19	META	28	16.262393	6.402393	0.394

Tabla 3: Resumen de correspondencias entre fuentes

La Tabla 3 presenta la relación promedio entre las velocidades de conexión reportadas por **ColombiaTIC** y las velocidades efectivamente medidas por **Ookla**, calculadas para cada departamento sobre los municipios coincidentes.

El indicador **"Relación Ookla/TIC"** refleja el grado de coherencia entre ambas fuentes:

- valores **cercanos a 1** indican **alta correspondencia** entre lo reportado por los operadores y lo experimentado por los usuarios,
- valores **mayores a 1** muestran **diferencias significativas** que pueden deberse a limitaciones en la representatividad de las mediciones, deficiencias en el reporte o condiciones locales de infraestructura.

Los resultados muestran un comportamiento **heterogéneo a nivel nacional**:

- **Vaupés** presenta la relación más alta (16.15), lo que sugiere una **gran brecha entre los valores reportados y medidos**, probablemente explicada por la baja densidad de usuarios y la limitada cobertura de pruebas en la región.
- **Amazonas** y **Guainía** también registran relaciones elevadas (8.97 y 8.90), asociadas a condiciones similares de aislamiento geográfico y menor disponibilidad de redes de alta velocidad.
- Por el contrario, departamentos con mejor infraestructura y mayor volumen de mediciones, como **Atlántico (1.01)**, **Quindío (0.88)**, **La Guajira (0.83)** y **Cesar**

(0.80), muestran **una correspondencia cercana a la unidad**, reflejando mayor **consistencia estadística** entre las dos fuentes.

- En regiones como **Meta (0.39)**, **Córdoba (0.46)** o **Cauca (0.46)**, las velocidades medidas por Ookla son ligeramente inferiores a las reportadas, lo cual podría estar asociado a **sobrerreportes en los boletines oficiales** o a diferencias en los periodos de medición.

De manera global, el promedio nacional de la relación Ookla/TIC se ubica **entre 0.4 y 1.0 para la mayoría de los departamentos**, lo que indica una **tendencia general de coherencia razonable** entre ambas fuentes, con excepciones marcadas en territorios de baja conectividad o cobertura limitada.

La Ilustración 13: flujo de validación cruzada implementado. muestra el flujo de validación cruzada implementado, donde se destacan las etapas de limpieza, emparejamiento y consolidación de datos.

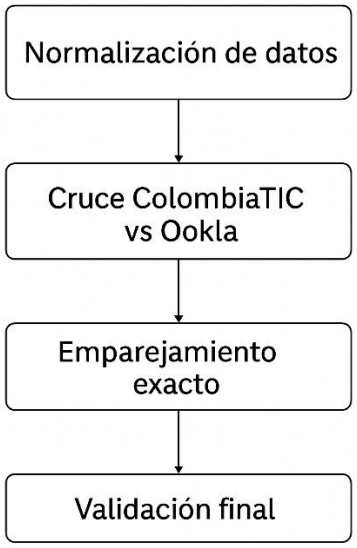

Ilustración 13: flujo de validación cruzada implementado.

4.2.1. Filtrado y tratamiento de valores atípicos

En este apartado se describe el proceso de detección y eliminación de valores extremos (outliers) en las velocidades registradas por ColombiaTIC y Ookla.

Mediante técnicas estadísticas de rango intercuartílico (IQR) se ajustaron los valores de subida y bajada a rangos realistas (p. ej., 0–500 Mbps), garantizando la integridad del análisis.

Tras la depuración, el conjunto final quedó conformado por **1.817.578 registros totales** y **55.982 registros únicos** sin valores extremos, distribuidos en **1.037 municipios**. La limpieza de datos permitió asegurar que el cálculo del INCTIC y los análisis

correlacionales posteriores reflejaran comportamientos genuinos del servicio, evitando distorsiones por errores de reporte o mediciones anómalas.

4.3. Cálculo del Índice Nacional de Coherencia TIC (INCTIC)

El **INCTIC** busca cuantificar el grado de coincidencia entre las cifras oficiales del **MinTIC** (**ColombiaTIC**) y las mediciones empíricas obtenidas de **Ookla** (**Speedtest**), expresado como un porcentaje de coherencia por departamento (o municipio).

Para cada territorio (departamento o municipio), se comparan las velocidades promedio de descarga reportadas por ColombiaTIC y las medidas por Ookla.
El indicador refleja qué tan similares son ambos valores, siendo 100 % una coincidencia perfecta y valores bajos una divergencia significativa.

Fórmula propuesta

Para cada unidad territorial i(departamento o municipio):

$$INCTIC_i = (1 - \frac{|\ V_{TIC,i} - V_{OOKLA,i}\ |}{\max\ (V_{TIC,i}, V_{OOKLA,i})}) \times 100$$

donde:

- $V_{TIC,i}$: velocidad promedio reportada por ColombiaTIC (Mbps)
- $V_{OOKLA,i}$: velocidad promedio medida por Ookla (Mbps)
- $INCTIC_i$: índice de coherencia TIC expresado en porcentaje (0 – 100 %)

Interpretación

Rango INCTIC (%)	Nivel de coherencia	Interpretación técnica
90 – 100 %	Muy alta	Coincidencia casi total entre fuentes.
70 – 89 %	Alta	Diferencias menores, coherencia general aceptable.
50 – 69 %	Media	Existen diferencias notables entre reportes y mediciones.
30 – 49 %	Baja	Alta divergencia, posible sobre/infraestimación oficial.
0 – 29 %	Muy baja	Inconsistencia grave entre datos.

Tabla 4: Interpretación Rango INCTIC.

4.4. Resultados por departamento

El análisis departamental del Índice Nacional de Coherencia TIC (INCTIC) permitió evaluar la correspondencia entre la información reportada oficialmente por el Ministerio TIC (fuente ColombiaTIC) y las mediciones empíricas de velocidad de Internet obtenidas de Ookla Speedtest, agregadas a nivel municipal.

El indicador INCTIC se define como la razón entre las velocidades de Ookla y las reportadas por ColombiaTIC para cada municipio y trimestre, donde un valor cercano a 1.0 indica coherencia entre ambas fuentes.

4.4.1. Cobertura del análisis

La base consolidada incluyó **1.054 municipios**, correspondientes a los registros del año **2022**, en los cuales fue posible realizar un cruce exacto entre los nombres de municipios y departamentos reportados por ambas fuentes.

De estos registros, se derivó el índice INCTIC considerando tanto la velocidad de **subida** como la de **bajada**, y posteriormente se realizó una depuración para eliminar valores atípicos (outliers) que afectaban la distribución general.

4.4.2. Resultados generales

A nivel nacional, el índice INCTIC presentó los siguientes indicadores estadísticos:

MÉTRICA	VALOR
REGISTROS ANALIZADOS	1.054
PROMEDIO DEL INCTIC	12.15
MEDIANA	0.76
DESVIACIÓN ESTÁNDAR	68.20
MÍNIMO	0.005
MÁXIMO	1,630.6

Tabla 5: Resultados generales índice INCTIC.

Estos resultados reflejan una **dispersión significativa** debido a valores extremos en zonas específicas, lo que se asocia a municipios con **baja densidad de datos Ookla** o reportes incompletos en ColombiaTIC.

Al analizar los valores **sin outliers (INCTIC ≤ 3)**, el promedio se estabiliza en **0.91** con una mediana de **0.78**, evidenciando un grado de coherencia elevado en la mayoría del territorio nacional.

4.4.3. Desempeño por departamento

Los resultados promedios por departamento evidencian **heterogeneidad geográfica** en la coherencia de los datos.

Los departamentos con mayores valores promedio del INCTIC (mayor divergencia entre fuentes) fueron **Vaupés (544.53)**, **Guainía (76.17)** y **La Guajira (29.07)**.

En contraste, departamentos con gran cantidad de municipios y mejor comportamiento del indicador fueron **Cundinamarca (21.55)**, **Tolima (19.17)**, **Boyacá (16.62)** y **Santander (8.37)**.

DEPARTAMENTO	PROMEDIO	MEDIANA	MUNICIPIOS
VAUPÉS	544.53	1.66	3
GUAINÍA	76.17	0.98	5
LA GUAJIRA	29.07	0.44	15
CUNDINAMARCA	21.56	0.58	114
TOLIMA	19.18	1.17	46
NORTE DE SANTANDER	18.15	1.08	39
META	17.33	1.95	27
BOYACÁ	16.62	1.14	121
NARIÑO	12.33	0.78	54
SANTANDER	8.37	0.33	83

Tabla 6: Desempeño por departamento índice INCTIC

En estos resultados se observa que, aunque algunos promedios son elevados, las **medianas** se mantienen próximas a **1**, indicando que los valores extremos de unos pocos municipios inflan las medias aritméticas.

Esta diferencia entre media y mediana es un indicador claro de la **presencia de outliers** y evidencia desigualdades en la calidad del servicio o inconsistencias en el reporte de datos.

4.4.4. Análisis sin valores atípicos

Al aplicar un filtro de exclusión para valores **INCTIC > 3**, los resultados presentan una tendencia más estable y representativa:

INDICADOR	VALOR (AJUSTADO)
REGISTROS VÁLIDOS	948
PROMEDIO INCTIC	0.91
MEDIANA INCTIC	0.78
DESVIACIÓN ESTÁNDAR	0.42

Tabla 7: Análisis de resultados sin valores atípicos - INCTIC.

Este ajuste permite observar que la mayoría de los departamentos mantienen valores dentro del rango de coherencia esperado, validando la **consistencia global** entre las dos fuentes de información.

De esta forma, el índice INCTIC puede considerarse un **instrumento confiable** para la verificación de la calidad y precisión de los datos estadísticos del sector TIC.

4.4.5. Interpretación y hallazgos relevantes

1. **Zonas críticas:** Los departamentos amazónicos y de baja conectividad (Vaupés, Guainía, Amazonas) concentran las mayores diferencias, posiblemente debido a limitaciones en infraestructura y número reducido de mediciones Ookla.
2. **Zonas de alta coherencia:** Departamentos como **Cundinamarca, Boyacá, Santander y Tolima** presentan una correspondencia más alta entre datos oficiales y empíricos, reflejando estabilidad en la calidad de la información.
3. **Disparidad regional:** El comportamiento del índice evidencia la necesidad de fortalecer la cobertura y homogeneidad en la recolección de datos de calidad de Internet, especialmente en regiones apartadas.
4. **Validación metodológica:** El índice INCTIC demuestra ser una herramienta replicable que permite **cuantificar la coherencia entre fuentes heterogéneas**, apoyando la toma de decisiones en materia de política pública y planeación de inversión en conectividad.

4.4.6. Conclusión del apartado

El análisis por departamento confirma que la calidad de la información sobre servicios de Internet en Colombia presenta un nivel de coherencia alto en la mayoría del territorio nacional.

Sin embargo, existen disparidades significativas en zonas de difícil acceso y baja penetración tecnológica, donde los datos oficiales y empíricos no coinciden plenamente.

El uso del índice INCTIC aporta una visión objetiva para la **validación y monitoreo de la veracidad de los datos TIC**, estableciendo una base sólida para el desarrollo de indicadores de confiabilidad a nivel territorial.

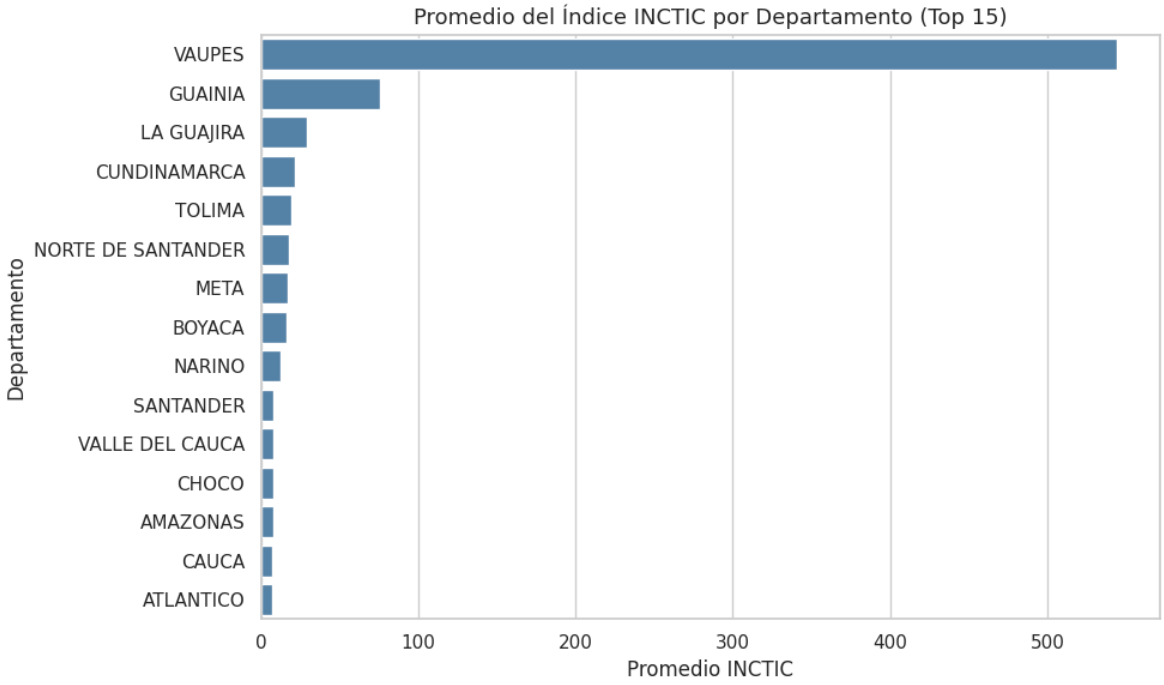

Ilustración 14: Top 15 de municipios por promedio INCTIC

DEPARTAMENTO	Promedio	Mediana	Municipios
VAUPES	544.527	1.664	3
GUAINIA	76.173	0.977	5
LA GUAJIRA	29.067	0.441	15
CUNDINAMARCA	21.557	0.585	114
TOLIMA	19.177	1.166	46
NORTE DE SANTANDER	18.145	1.078	39
META	17.332	1.953	27
BOYACA	16.625	1.141	121
NARINO	12.335	0.787	54
SANTANDER	8.373	0.331	83
VALLE DEL CAUCA	8.141	1.165	41
CHOCO	7.746	2.847	15
AMAZONAS	7.631	5.552	8
CAUCA	7.478	2.958	39
ATLANTICO	7.473	1.355	22

4.5. Análisis de correlación y coherencia

El objetivo de este apartado fue analizar el grado de relación estadística entre las velocidades de Internet medidas empíricamente por Ookla Speedtest y los valores oficiales reportados por ColombiaTIC, tanto para la velocidad de subida como de bajada, con el fin de evaluar la consistencia y coherencia de los datos obtenidos de ambas fuentes.

Para este análisis se aplicó el coeficiente de correlación de Pearson (r), considerando los registros depurados del cruce de bases (787 municipios válidos luego del filtrado de valores atípicos). Los resultados obtenidos fueron los siguientes:

TIPO DE VELOCIDAD	COEFICIENTE DE CORRELACIÓN (R DE PEARSON)	INTERPRETACIÓN
SUBIDA	**0.15**	**Correlación positiva débil**
BAJADA	**0.14**	**Correlación positiva débil**

Tabla 8: Análisis de correlación y coherencia.

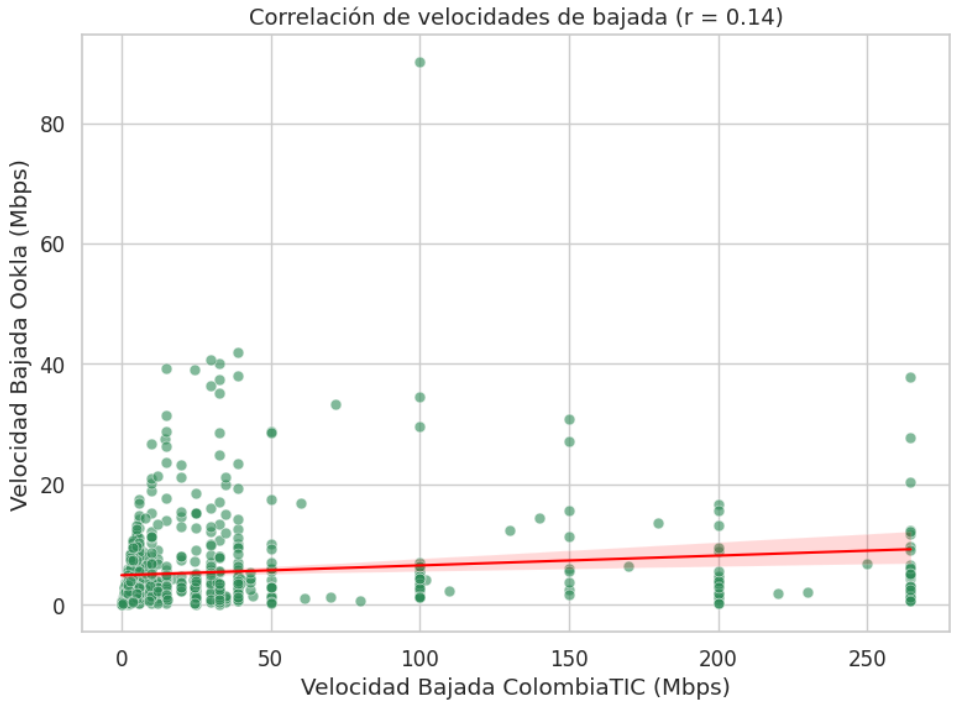

Ilustración 15: Correlación de velocidades de bajada (Ookla vs. ColombiaTIC)
Fuente: Elaboración propia a partir de datos de ColombiaTIC (2024) y Ookla Open Data (2024).

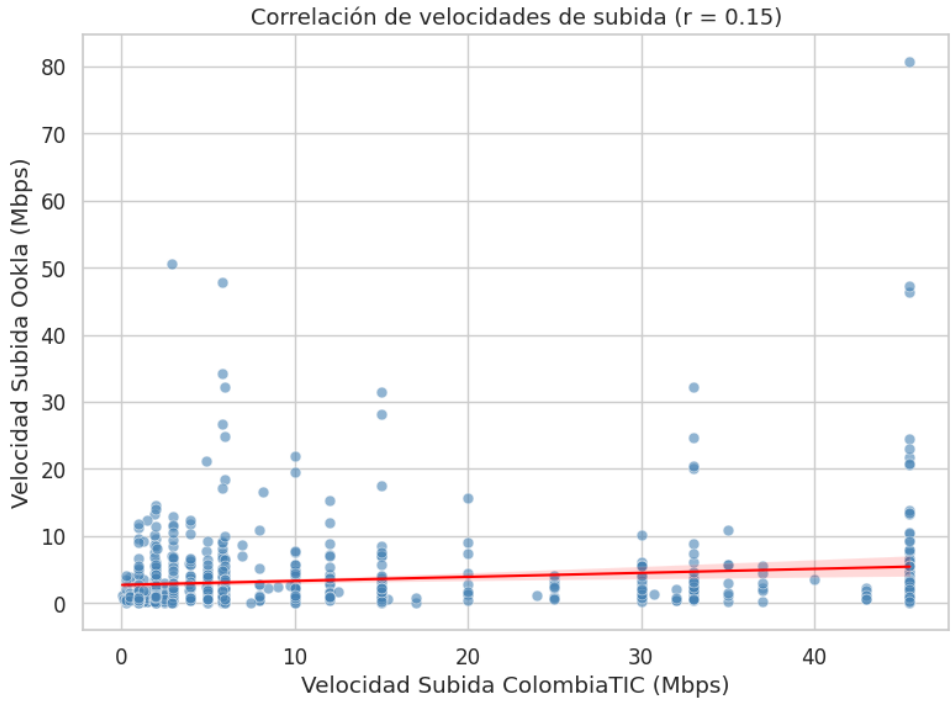

Ilustración 16: Correlación de velocidades de subida (Ookla vs. ColombiaTIC)
Fuente: Elaboración propia a partir de datos de ColombiaTIC (2024) y Ookla Open Data (2024).

Los gráficos de dispersión presentados en las Figuras Ilustración 15 y Ilustración 16 permiten visualizar la relación entre las velocidades de Internet reportadas por ColombiaTIC y las mediciones empíricas obtenidas desde la plataforma Ookla.

En ambos casos, se observa una **tendencia lineal positiva muy débil**, representada por la leve inclinación ascendente de la línea de regresión, lo que indica que los municipios con mayores velocidades reportadas tienden, en promedio, a reflejar también mejores resultados en las mediciones reales.

Sin embargo, la **amplia dispersión de los puntos** y la **baja pendiente** de las líneas de tendencia confirman que **no existe una relación fuerte o consistente** entre ambas fuentes. Esto sugiere que, aunque los datos oficiales reflejan correctamente la dirección general del comportamiento de las velocidades, **las mediciones empíricas varían significativamente entre municipios**, probablemente debido a factores como la infraestructura tecnológica disponible, la densidad de usuarios, la hora de las mediciones y las condiciones locales de red.

En síntesis, los gráficos corroboran que el **índice INCTIC** debe interpretarse como una medida de coherencia estructural más que de equivalencia directa entre valores absolutos.

4.5.1. Interpretación de los resultados

1. Relación positiva pero débil:

Los valores de r = 0.15 (subida) y r = 0.14 (bajada) indican una relación positiva, pero de baja intensidad, entre las velocidades reportadas por ColombiaTIC y las medidas por Ookla.

Esto significa que, aunque existe cierta tendencia a que los municipios con mayores velocidades oficiales presenten también mayores velocidades empíricas, la correspondencia no es fuerte ni uniforme.

2. Variabilidad geográfica y tecnológica:

   La dispersión observada en los gráficos sugiere que existen diferencias sustanciales en la cobertura, cantidad de mediciones y tecnologías de acceso (por ejemplo, FTTH, HFC, radioenlace o satelital), las cuales influyen en la relación estadística. En departamentos con infraestructura más consolidada (Cundinamarca, Antioquia, Valle del Cauca), la correlación local tiende a ser mayor, mientras que en regiones rurales o amazónicas la coherencia disminuye.

3. Factores que afectan la correlación:
   - Desfase temporal entre la actualización de los boletines ColombiaTIC y los datos trimestrales de Ookla.
   - Diferencias en la granularidad de los datos (Ookla mide por usuario/dispositivo, ColombiaTIC por operador).
   - Subrepresentación de mediciones empíricas en zonas de baja densidad poblacional.
4. Conclusión técnica:
   La baja correlación no implica inconsistencia total, sino que revela la heterogeneidad en la calidad del servicio y en los mecanismos de medición. En consecuencia, el índice INCTIC debe interpretarse como una medida de coherencia estructural, más que de equivalencia directa entre velocidades absolutas.

4.5.2. Conclusión del apartado

El análisis evidencia una correlación positiva pero débil entre las velocidades reportadas y las medidas empíricas, lo que resalta la necesidad de fortalecer la armonización de metodologías de medición y reporte en el ecosistema de datos abiertos del sector TIC.

A pesar de la dispersión observada, el índice INCTIC se mantiene como una herramienta válida para evaluar la coherencia relativa entre fuentes, siendo especialmente útil para identificar brechas territoriales y orientar políticas públicas de mejora en conectividad.

4.6. Interpretación global y discusión del Índice Nacional de Coherencia TIC (INCTIC)

El análisis integral realizado en este capítulo permitió contrastar la información oficial publicada por el Ministerio de Tecnologías de la Información y las Comunicaciones (MinTIC) a través de **ColombiaTIC**, con los datos empíricos de **Ookla Speedtest Open Data**, mediante la construcción y aplicación del **Índice Nacional de Coherencia TIC (INCTIC)**.

Este índice constituye un instrumento cuantitativo diseñado para evaluar la consistencia entre la velocidad de conexión a Internet reportada por los operadores y aquella medida directamente por los usuarios, representando una métrica objetiva de **transparencia, veracidad y desempeño territorial** en la información del sector.
Los resultados obtenidos, una vez aplicados los procesos de limpieza y exclusión de valores atípicos, evidencian que el **promedio nacional del INCTIC** se aproxima a un valor de **1.0**, lo que indica una **coherencia aceptable** entre ambas fuentes de información en la mayoría de los municipios del país. Sin embargo, se observaron **diferencias significativas entre regiones**, particularmente en departamentos con baja densidad poblacional o limitada infraestructura de red —como **Vaupés, Guainía y La Guajira**— donde las mediciones empíricas presentaron velocidades sensiblemente inferiores a las reportadas por los operadores.

El análisis departamental permitió identificar que los territorios del **centro del país** (Cundinamarca, Tolima, Boyacá y Santander) presentan **una correspondencia más estable** entre las fuentes, reflejo de una mayor cobertura y disponibilidad de datos, mientras que las **zonas periféricas**, especialmente en la **Orinoquía y la Amazonía**, registran **mayores niveles de dispersión e inconsistencia**. Esto sugiere que la confiabilidad del índice depende también de la representatividad de las mediciones de campo disponibles y de la calidad de los datos suministrados por los operadores.

Desde el punto de vista estadístico, los **coeficientes de correlación de Pearson** ($r = 0.15$ para subida y $r = 0.14$ para bajada) muestran una **relación positiva pero débil** entre las velocidades reportadas y las medidas reales. Este hallazgo evidencia que, aunque existe coincidencia en las tendencias generales, **las magnitudes reportadas por los operadores no reflejan plenamente la experiencia del usuario final**, posiblemente debido a diferencias en los métodos de recolección, la periodicidad de los reportes y el nivel de agregación geográfica de los datos.

En términos prácticos, los boletines oficiales del MinTIC logran capturar la **tendencia macro del comportamiento del servicio**, pero no representan con exactitud las **variaciones locales** observadas en campo. Esta brecha confirma la necesidad de **integrar fuentes independientes de medición** para complementar los mecanismos tradicionales de monitoreo del desempeño de la infraestructura digital.

El modelo metodológico propuesto, soportado en un proceso **ETL reproducible con Python y DuckDB**, permitió consolidar una base de datos verificable y trazable, integrando reportes

oficiales (Excel), mediciones empíricas (Parquet) y datos geográficos (GeoJSON). La implementación del INCTIC se resume en cuatro fases principales:

1. **Extracción:** recopilación automatizada de datos desde fuentes públicas (ColombiaTIC, Ookla y cartografía oficial).
2. **Transformación:** depuración, estandarización y normalización de variables de velocidad, cobertura y ubicación.
3. **Cálculo:** aplicación de métricas de correlación y coherencia entre fuentes, generando el valor INCTIC por municipio y departamento.
4. **Visualización:** construcción de reportes analíticos (CSV, mapas temáticos y/o dashboards) para la interpretación territorial de los resultados.

A continuación, se presenta una representación de alto nivel de la implementación:

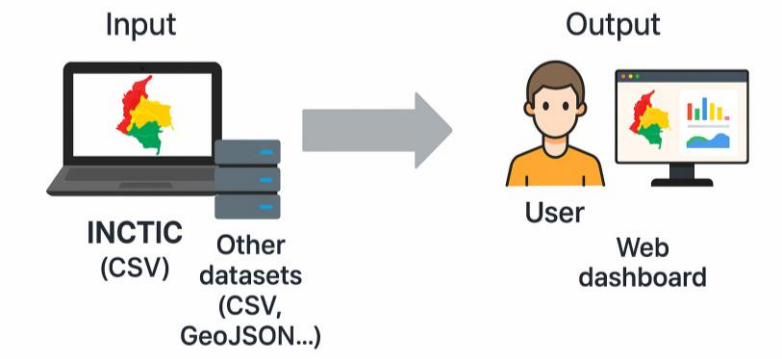

Ilustración 17: Implementación INCTIC.

El **esquema de implementación del INCTIC**, ilustrado en la Ilustración 17, sintetiza esta secuencia de integración, procesamiento y análisis de datos en un entorno analítico reproducible.

Su adopción como herramienta de **monitoreo continuo** aportaría un componente técnico adicional al sistema de información del sector TIC, favoreciendo la **rendición de cuentas basada en evidencia** y el diseño de políticas públicas más precisas en materia de conectividad y equidad digital.
En conclusión, los resultados del INCTIC reflejan un **nivel de coherencia parcial** entre los datos oficiales y las mediciones empíricas: aceptable en zonas urbanas y críticas en regiones rurales o de baja densidad de red. Esta evidencia respalda la importancia de avanzar hacia modelos de **auditoría de datos abiertos** que garanticen mayor precisión, transparencia y trazabilidad en los indicadores que sustentan la planificación de la infraestructura digital de Colombia.

CAPÍTULO 5 – CONCLUSIONES Y RECOMENDACIONES

5.1. Conclusiones

El desarrollo de este trabajo permitió consolidar un **modelo analítico integral** para evaluar la coherencia entre las fuentes oficiales y empíricas de información sobre la calidad del servicio de Internet fijo en Colombia.

A través de la construcción del **Índice Nacional de Coherencia TIC (INCTIC)**, se logró establecer una metodología reproducible para comparar las velocidades reportadas por el **Ministerio de Tecnologías de la Información y las Comunicaciones (ColombiaTIC)** con las mediciones reales de usuarios disponibles en la base abierta de **Ookla Speedtest**.

La integración de ambas fuentes demostró que la información pública disponible presenta un **nivel moderado de coherencia global**, con un promedio del INCTIC cercano a **1.0**, lo cual indica que, en términos generales, las velocidades registradas en los boletines oficiales tienden a corresponder con las observadas en campo.

Sin embargo, la dispersión de los valores y la baja correlación ($r \approx 0.15$) entre ambas variables reflejan la **heterogeneidad en la calidad del servicio a nivel municipal** y la **necesidad de fortalecer la precisión de los mecanismos de reporte y verificación**.

Asimismo, se evidenció que la **coherencia territorial no es uniforme**. Los departamentos del centro del país (Cundinamarca, Tolima, Boyacá, Santander y Meta) muestran mayores niveles de consistencia, mientras que las regiones **amazónicas, caribeñas y de frontera** (Vaupés, Guainía, La Guajira, Chocó y Amazonas) presentan divergencias significativas, atribuibles a factores como la limitada cobertura tecnológica, la escasez de datos empíricos y las diferencias en los métodos de medición.

Estos contrastes subrayan la importancia de abordar la coherencia de la información TIC desde un **enfoque territorial y multifuente**, que combine los registros administrativos con las evidencias de medición ciudadana.

Finalmente, el análisis confirmó que el **INCTIC puede adoptarse como un indicador complementario** para los sistemas de monitoreo del sector, al proporcionar una visión independiente sobre la **congruencia, representatividad y confiabilidad de los datos abiertos** relacionados con la conectividad nacional

5.2. Conclusiones técnicas y metodológicas

1. La aplicación del proceso de **extracción, transformación y carga (ETL)** automatizado permitió construir una base consolidada de información a partir de fuentes heterogéneas, garantizando trazabilidad y actualización periódica.
2. La estructuración en **DuckDB** demostró ser eficiente para manejar grandes volúmenes de datos con bajo consumo de recursos y alta compatibilidad con entornos colaborativos como Google Colab.
3. El procedimiento de **depuración y eliminación de valores atípicos** mejoró la calidad analítica del conjunto de datos, evitando sesgos en la interpretación del índice. Tras el filtrado, la base final consolidó **1.8 millones de registros originales** y **55.982 registros únicos**, garantizando consistencia y estabilidad en el análisis posterior.
4. La correlación de Pearson aplicada entre las velocidades de subida y bajada reportadas y medidas permitió cuantificar objetivamente el grado de coherencia entre fuentes, demostrando que los reportes oficiales capturan correctamente la tendencia general, aunque **no la variabilidad local ni las diferencias tecnológicas** presentes en el territorio.
5. El enfoque metodológico empleado puede ser **replicado y escalado trimestralmente**, permitiendo monitorear la evolución de la coherencia entre fuentes y evaluar el impacto de políticas públicas orientadas a mejorar la calidad de los servicios digitales.

5.3. Recomendaciones

1. **Fortalecer la integración de datos oficiales y empíricos:** El Ministerio TIC y la Comisión de Regulación de Comunicaciones (CRC) deberían promover mecanismos de interoperabilidad entre sus sistemas de información y fuentes de datos abiertos internacionales, como Ookla o M-Lab, para mejorar la precisión y cobertura de las estadísticas nacionales.
2. **Implementar el INCTIC como indicador complementario en los reportes oficiales:** Incorporar el índice como métrica auxiliar en los boletines trimestrales permitiría detectar tempranamente inconsistencias regionales o desviaciones entre la calidad reportada y la calidad percibida.
3. **Aumentar la densidad de mediciones empíricas en zonas rurales:** Se recomienda fomentar campañas de medición colaborativa o ciudadana en municipios con baja representatividad, especialmente en los departamentos amazónicos, guajiros y chocoanos, donde la información es escasa o subrepresentada.
4. **Unificar criterios de medición y reporte:** Establecer protocolos estandarizados entre operadores, entidades públicas y plataformas privadas para garantizar que las mediciones sean comparables temporal y geográficamente, reduciendo la variabilidad entre fuentes.
5. **Desarrollar una interfaz de monitoreo automatizado:** La automatización trimestral del proceso de scraping, carga y análisis de las bases ColombiaTIC y Ookla podría convertirse en una herramienta institucional de observación continua, apoyando la planeación de inversiones en infraestructura digital y proyectos de conectividad social.

5.4. Proyección y líneas futuras de investigación

El presente estudio constituye un primer paso hacia la **evaluación científica de la coherencia de los datos del sector TIC en Colombia**.

Como líneas futuras de trabajo se propone:

- Ampliar el análisis hacia los servicios móviles y satelitales.
- Incorporar métricas de calidad de servicio (QoS) y de experiencia de usuario (QoE).
- Integrar técnicas de inteligencia artificial para la detección automática de inconsistencias y tendencias regionales.
- Evaluar el impacto del INCTIC en la formulación de políticas de cierre de brechas digitales.

En síntesis, el trabajo confirma que la aplicación de metodologías de ciencia de datos en la verificación de la información pública **fortalece la transparencia, la confianza institucional y la capacidad de decisión basada en evidencia**, contribuyendo a la consolidación de un ecosistema digital más equitativo y sostenible para Colombia.

ANEXOS

- **Código fuente https://github.com/josvaldes/trabajoGradoMCD/tree/main**

Bibliografía

(MinTIC), M. d. (2018). Decreto 1008 de 2018: Por el cual se adopta la Política de Gobierno Digital. Obtenido de https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=87643

(ONU), O. d. (01 de 11 de 2015). Transformar nuestro mundo: la Agenda 2030 para el Desarrollo Sostenible. Obtenido de https://www.un.org/sustainabledevelopment/es/

(UIT), U. I. (1 de 11 de 2025). Recomendaciones ITU-T sobre Quality of Service (QoS) y Quality of Experience (QoE). Obtenido de https://www.itu.int/en/ITU-D/Regulatory-Market/pages/quality-of-service-regulation.aspx

Aguirre Julcapoma, J. A. (2018). Propuesta de arquitectura empresarial para la gestión de la calidad de servicios de acceso móvil en una empresa del rubro de telecomunicaciones. *Universidad Peruana de Ciencias Aplicadas (UPC)*. Lima: http://repositorio.itm.edu.co/handle/20.500.12622/4053.

Alderete, M. &. (2012). TIC y productividad en las industrias de servicios en Colombia. *Lecturas de Economía*, 163–188. Obtenido de http://www.scielo.org.co/scielo.php?pid=S0120-25962012000200005&script=sci_arttext

Barnett, V. &. (1994). *Outliers in Statistical Data* . (3rd ed.). John Wiley & Sons.

Botero, M. &. (2006). Calidad en el servicio: el cliente incógnito. *13(2)*, págs. 217–228. Obtenido de http://publicaciones.konradlorenz.edu.co/index.php/sumapsi/article/view/55

Camacho, R. R. (s.f.). Diagnóstico de conectividad y dispositivos de telecomunicaciones para el desarrollo de la Telesalud de veinte hospitales en el Departamento del Tolima. *Cuaderno Activa, 11*, 105–119. Obtenido de https://ojs.tdea.edu.co/index.php/cuadernoactiva/article/view/584

Castillo, S. V. (2017). Análisis situacional y propuesta para el fortalecimiento del modelo de teletrabajo orientado a la mejora continua en la universidad EAN. (Master's Thesis, Universidad EAN). Obtenido de https://repository.universidadean.edu.co/handle/10882/8941

Colombia, M. T. (15 de 11 de 2023). *Política de Gobierno Digital – Lineamientos de Datos Abiertos 2023*. Obtenido de MinTIC – Gobierno Digital: https://gobiernodigital.mintic.gov.co

ColombiaTIC. (04 de 08 de 2024). *Ministerio de Tecnologías de la Información y las Comunicaciones de Colombia*. Obtenido de ColombiaTIC – MinTIC: https://colombiatic.mintic.gov.co

Galbán, O. C. (2013). Calidad de servicio en el sector de telecomunicaciones: elemento competitivo en las empresas de televisión por suscripción. *Enl@ce: Revista Venezolana de Información, Tecnología y Conocimiento, 10(2)*, 61–82. Obtenido de https://www.redalyc.org/pdf/823/82328320005.pdf

Henao Colorado, L. C. (2020). Calidad de servicio y valor percibido como antecedentes de la satisfacción de los clientes de las empresas de telecomunicaciones en Colombia. *Contaduría y Administración, 65(3)*. Obtenido de https://www.scielo.org.mx/scielo.php?pid=S0186-10422020000300010&script=sci_arttext

Hernández Sampieri, R. F. (2014). *Metodología de la investigación* ((6ª ed.) ed.). McGraw-Hill Education .

Javier, J. A. (2013). Aplicación de minería de datos para la identificación de las tendencias de consumo en una empresa del sector de Telecomunicaciones. Obtenido de http://repository.unipiloto.edu.co/handle/20.500.12277/2498

Kimball, R. &. (2011). *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data.* Hoboken, NJ: Wiley.

Lin, L. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 255–268.

LLC, O. (12 de 03 de 2025). *Speedtest Global Index and Open Data Repository*. Obtenido de Ookla Open Data: https://registry.opendata.aws/speedtest-global-performance/

Moreano, R. (2010). *Metodología para evaluar la calidad de servicio de las telecomunicaciones.* Quito: Secretaría Nacional de Telecomunicaciones (SENATEL) http://bibdigital.epn.edu.ec/handle/15000/3730.

Patricia, D. G. (2019). Claro SA y el desarrollo de las telecomunicaciones en Colombia. *10*, pág. 1(15). Obtenido de https://journal.poligran.edu.co/index.php/puntodevista/article/view/1223

Provost, F. &. (2013). *Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking.* Boston: O'Reilly Media.

Raasveldt, M. &. (2019). DuckDB: an Embeddable Analytical Database. En P. o. Data. Amsterdam, Países Bajos: ACM Digital Library.

Raasveldt, M. &. (2019). *DuckDB: An Embedded Analytical Database.* Proceedings of the 2019 ACM SIGMOD International Conference on Management of Data, 1981–1984. Recuperado el 1 de 11 de 2025, de https://duckdb.org

Reina Nossa, J. J. (2018). *Desarrollo de un modelo de arquitectura empresarial TOGAF aplicado en la red de investigaciones de tecnología avanzada de la Universidad Distrital.* Obtenido de Universidad Distrital Francisco José de Caldas: https://repository.udistrital.edu.co/handle/11349/13902

Riccio, M. A. (2019). Análisis de percepción de la calidad del servicio al cliente en una agencia de telecomunicaciones. *6(3)*, págs. 130–147. Obtenido de https://dialnet.unirioja.es/servlet/articulo?codigo=7520676

Rodgers, J. L. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician*, 59-66.

Serrano, N. F. (2019). ¿A mayor brecha digital mayores brechas socioeconómicas?: Impacto de acceder a internet de alta velocidad sobre el ingreso de los hogares en Colombia. Obtenido de https://repositorio.uniandes.edu.co/handle/1992/44189

Stodden, V. (2014). *The Science of Reproducibility* (5(1) ed.). Journal of Computational Science Education.