Dirección de Desarrollo Digital

Unidad de Científicos de Datos

El futuro es de todos

DNP
Departamento Nacional de Planeación

El futuro es de todos

DNP
Departamento
Nacional de Planeación

ÍNDICE DE RIESGO DE CALIDAD DE AGUA PARA EL CONSUMO HUMANO

Entidad
Departamento Nacional de Planeación
- Dirección de Desarrollo Digital.
- Dirección de Desarrollo Urbano

Sector	**Lenguaje**	**Fuente de datos**
Salud y protección social	Python y R	SSPD, RUPS, SIVICAP

Presentación
En alineación con la propuesta para la estimación del Índice de Riesgo de la Calidad del Agua para Consumo Humano – IRCA a nivel municipal por parte de la Dirección de Desarrollo Urbano, la cual busca conseguir una medida de calidad del agua que considere la población afectada a nivel de municipio, urbano y rural. De tal modo, este proyecto nace a partir de la necesidad de automatizar dicha estimación por medio del desarrolló de un programa (script) que permita calcular el indicador IRCA desde el principio desde que la información ingresa al DNP hasta que se calcula el indicador. Además, se desarrollará una herramienta de visualización, la cual permitirá filtrar los resultados según las principales características de las muestras.

Objetivo general
Estandarizar y automatizar el proceso por el cual se estructura y se calcula el índice de riesgo de calidad de agua para el consumo humano (IRCA), bajo la propuesta del cálculo de la Dirección de Desarrollo Urbano (DDU), así como crear una herramienta de visualización del IRCA para identificar las zonas del país con problemas de calidad del agua y los parámetros relacionados con las características físicas, químicas y microbiológicas de ésta, que impactan negativamente en el indicador, con el fin de plantear soluciones de política pública que permitan mejorar la calidad del agua para consumo humano en el país.

Objetivos específicos
1. Implementar un *script* para realizar la imputación de Id de los prestadores de servicios.
2. Realizar proceso de completitud de datos a partir del cruce de la base de datos Registro Único de Prestadores de Servicios (RUPS), Sábanas de datos de 2017 a 2019 y Número de Suscriptores por prestador.
3. Realizar metodología de predicción para la completitud de la columna Número de Suscriptores.
4. Implementar un *script* que permita calcular el IRCA ponderado por el número de suscriptores.
5. Desarrollar una aplicación web y/o tablero interactivo de visualización con los datos calculados del IRCA.

Metodología

El abordaje metodológico desarrollado consistió en las siguientes fases:
1. Procesamiento de los datos:
    a. Limpieza de bases de datos
    b. Imputación de IDs y número de suscriptores por empresa
    c. Completitud de datos y cálculo del IRCA

2. Desarrollo de una aplicación web:
    a. Imputación de la información geográfica.
    b. Interfaz de la aplicación.
    c. Mapas.

1. Procesamiento de los datos

a. Limpieza de bases de datos

Primero se realizó una exploración inicial, donde fueron calculadas distintas métricas de evaluación. La Tabla 1 hace referencia a un resumen general de las tres bases de datos suministradas por la DDU.

Tabla 1. Resumen general de bases de datos

Métrica de evaluación	Resultado 2017	Resultado 2018	Resultado 2019
Número de filas	47.561	51.671	46.455
Número de columnas	127	124	126
Columnas numéricas	55	55	55
Columnas de texto	74	69	71
Número de filas duplicadas	0	0	0
Número de columnas duplicadas	0	0	0
Columnas con más de la mitad de los datos faltantes	78	81	82
Columnas con más del 10% de datos como extremos	4	0	0

De acuerdo con la exploración inicial se identificaron varias columnas de texto en las tres bases de datos, fue necesario realizar labores de limpieza de texto para cada columna, con el fin de facilitar el tratamiento de los datos. Este procedimiento consistió en remover características que no permiten realizar un análisis efectivo como lo son signos de puntuación, caracteres no alfanuméricos, números, acentos (tildes, eñes), palabras con menos de un número determinado de caracteres (menor a 2 caracteres), espacios innecesarios, y adicionalmente se transformó el texto a minúscula.

Además de realizar un análisis exploratorio de las bases de datos, se seleccionaron solo las variables necesarias para el cálculo del IRCA, las cuales fueron:

- Nombre PP
- Nit
- Departamento PM
- Municipio PM
- Lugar PM
- Ubicación
- Descripción fuente

- IRCA
- IRCA básico

b. Imputación de Id número de suscriptores por empresa

Luego de la limpieza de texto se realizó la imputación de Id y número de suscriptores por empresa, para ello se extrajeron los valores únicos referentes al nombre del prestador de servicio público y se realizó el cruce entre las siguientes bases de datos:

- Sábanas de datos de 2017 a 2019
- Registro Único de Prestadores de Servicios (RUPS)
- Número de suscriptores por prestador de servicio

El primer cruce se realizó entre las sábanas de datos y RUPS a través del NIT, con el objetivo de imputar el Id de las empresas. Sin embargo, tras este cruce solo el 18.25% de las empresas prestadoras de servicios públicos quedaron con un Id asignado, por lo cual se realizó un proceso de imputación realizando uso de técnicas de análisis de texto, se usó la distancia Levensthein (Fórmula 1) la cual consiste en identificar el menor número de operaciones requeridas para transformar una cadena de caracteres a la cadena de caracteres objetivo, por ejemplo, la distancia Levensthein entre casa y capa es de 1 que refiere a la sustitución de p por s. Adicionalmente, la distancia Levensthein fue usada teniendo en cuenta el departamento y el municipio, es decir, que únicamente se hicieron comparaciones entre los nombres de las empresas prestadoras de servicios públicos que aparecen en un municipio de la sabana de datos con las empresas prestadoras de servicios públicos que aparecen en la base del RUPS. Teniendo en cuenta en esta técnica descrita un margen de error del 10%. Para ello fue necesario calcular la distancia Levensthein la cual permite medir la diferencia entre dos secuencias.

Ecuación 1. Distancia de Levensthein

$$lev_{a,b}(i,j) = \begin{cases} \max(i,j) & if\ \min(i,j) = 0, \\ \min \begin{cases} lev_{a,b}(i-1,j) + 1 \\ lev_{a,b}(i,j-1) + 1 \\ lev_{a,b}(i-1,j-1) + 1_{(a_i \neq b_i)} \end{cases} & otherwise \end{cases}$$

Tras la implementación de esta metodología, los Id asignados a las empresas prestadoras de servicios públicos aumentaron a un 62.06%. Adicionalmente se realizó el cruce entre Sábanas de datos y Número de suscriptores por prestador de servicio, obteniendo el número de suscriptores y algunos Id adicionales.

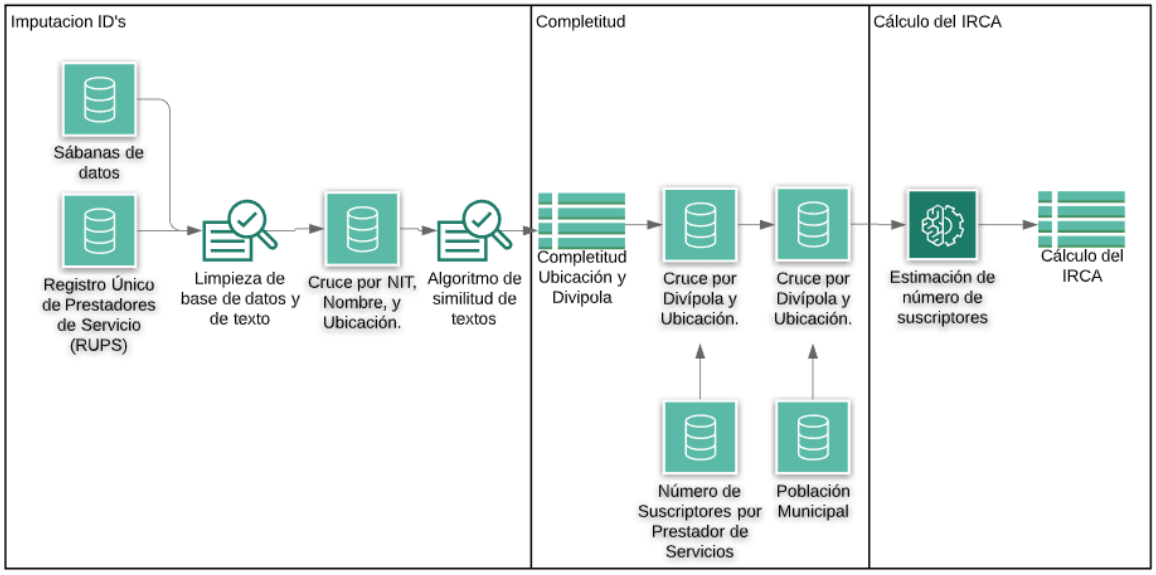


Ilustración 1. Metodología

c. Completitud de datos y Cálculo del IRCA

El primer análisis exploratorio permite percibir una gran cantidad de datos faltantes en columnas claves para realizar el cálculo del IRCA, como lo es la variable de 'Ubicación', por lo tanto, se utiliza una técnica denominada reglas de asignación, para identificar las palabras presentes en la columna 'Descripción de la Fuente' y 'Lugar PM' que hagan exclusiva referencia a registros rurales o urbanos. Posteriormente a la identificación de dichas reglas de asignación se aplica a la base de datos, logrando completar dicha variable.

Asimismo, fue necesario adicionar una columna que indicará el código del municipio y departamento (Divipola) para tener una identificación efectiva en la base de datos la sábana. Para ello, la información se extrajo de la página del Departamento Administrativo Nacional de Estadística y se cruzó con la base de datos la sabana obteniendo así la Divipola.

Posteriormente a la completitud se realizó un cruce con la base de datos de Número de Suscriptores por Prestador de Servicios, obteniendo tan sólo el número de suscriptores pertenecientes al 43% de los municipios y el 40% perteneciente a las empresas prestadoras de servicios, por lo tanto, se decidió realizar un entrenamiento de diferentes modelos de regresión que permitiera realizar el cálculo del número de suscriptores para cada una de las empresas y/o municipios.

También, se incluyó una variable adicional de la base de datos de la población municipal de Colombia del DANE. La correlación de esta variable es del 97% con respecto al número de suscriptores, por lo cual se decidió tenerla en cuenta como variable independiente en el modelo de regresión que se propuso.

Para los modelos de regresión se tuvieron en cuenta como variables independientes el IRCA, el número de muestras y la población del municipio y como variable dependiente el número de suscriptores, en total se probaron un varios modelos de regresión para identificar cuál de ellos y que parámetros propios de cada estimador maximizan la correcta estimación del número de suscriptores con respecto a datos de entrenamiento y datos de prueba, cabe aclarar que de los 47561 registros tan sólo 19210 fueron imputados con el número de suscriptores y estos fueron separados en un 80% para entrenar los modelos y 20% para probar los modelos. Para ello se escogieron los siguientes modelos de regresión:

- **Decision Tree Regressor:** es un método de aprendizaje supervisado no paramétrico utilizado para la clasificación y la regresión. El objetivo es crear un modelo que prediga el valor de una variable objetivo mediante el aprendizaje de reglas de decisión simples inferidas de las características de los datos. (scikit-learn, 2019)
- **KNN Regressor:** es un algoritmo simple que almacena todos los casos disponibles y predice el número objetivo basado en una medida de similitud.
- **Gradient Boosting Regressor:** es una técnica de aprendizaje automático para problemas de regresión y clasificación. El objetivo es construir un modelo aditivo de manera progresiva por etapas que permite la optimización de funciones arbitrarias de pérdida diferenciable. (scikit-learn, 2019)

Para la elección del mejor estimador se realizaron iteraciones sobre una grilla de parámetros en busca de los mejores parámetros para el regresor, los resultados de los desempeños por clasificador se presentan en la **Error! Reference source not found.**Tabla 2.

Tabla 2. Métricas de los modelos de predicción

Clasificador	R^2	Error Cuadrático Medio
Decision Tree Regressor	99%	0.05
Gradient Boosting Regressor	94%	0.54
K-Nearest Neighbors	99%	0.10

Dado que el algoritmo Decision Tree tuvo el menor error cuadrático medio (Ecuación 2) se realizó el ajuste, utilizándolo para la predicción del número de suscriptores.

Ecuación 2. Error Cuadrático Medio

$$Error\ cuadrático\ medio = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \overline{y_i})}{n}}$$

2. *Desarrollo de una aplicación web*

La herramienta de visualización fue implementada en el lenguaje de programación R, haciendo uso de las librerías Shiny y Shinydashboard, la creación del tablero, debido a que facilitan la implementación y el despliegue de aplicaciones web.

- Shiny es una herramienta de R que permite elaborar aplicaciones web interactivas, adicionalmente se puede hace uso de otros lenguajes de programación como temas CSS, HTML y JavaScript.
- Por su parte, Shinydashboard es un paquete de R que se utiliza para la construcción del interfaz.

a. Imputación de la información geográfica

A partir de los resultados del IRCA la primera fase fue la imputación de la información geográfica a cada municipio de la base de datos, para ello fue necesario realizar una reorganización de los datos, permitiendo tener en cada columna la información del cálculo del IRCA rural e IRCA urbano. Adicionalmente, a esta tabla y de forma automática se añadieron columnas relacionadas al área, perímetro, código departamental, código municipal, longitud, latitud, entre otras (**Error! Reference source not found.**), estas permiten a los scripts[1] imputar la información geográfica para cada uno de los municipios.

Tabla 3. Desagregación municipal

Nombre de columna	Descripción
DPTO	Código departamental
MPIO	Código municipal
IRCA rural simple	IRCA promedio ponderado por municipio
IRCA rural promedio ponderado por muestra	IRCA anual por municipio, ponderado por muestras
IRCA rural promedio ponderado por suscriptor	IRCA anual por prestador ponderado por suscriptor
Número de muestras rural	Número de muestras tomadas en zonas rurales
IRCA urbano simple	IRCA promedio ponderado por departamento
IRCA urbano promedio ponderado por muestra	IRCA anual por municipio ponderado por muestras
IRCA urbano promedio ponderado por suscriptor	IRCA anual por prestador, ponderado por suscriptor
Número de muestras urbano	Número de muestras tomadas en zonas urbanas

[1] Archivo en el cual se guarda el código escrito en un lenguaje de programación.

Departamento	Nombre de los departamentos de Colombia
Municipio	Nombre de los municipios de Colombia
Capital	Nombre de las capitales de los departamentos
Lng	Longitud de los municipios
Lat	Latitud de los municipios
Lng_cap_ajustada	Longitud departamental
Lat_cap_ajustada	Latitud departamental

Por medio de mapas coropléticos (mapas sombreados de distintos colores), utilizando como marca los polígonos a nivel municipal imputados anteriormente, se visualizan los mapas según los filtros seleccionados por el usuario. En la Ilustración 2 se identifica el flujo de información.

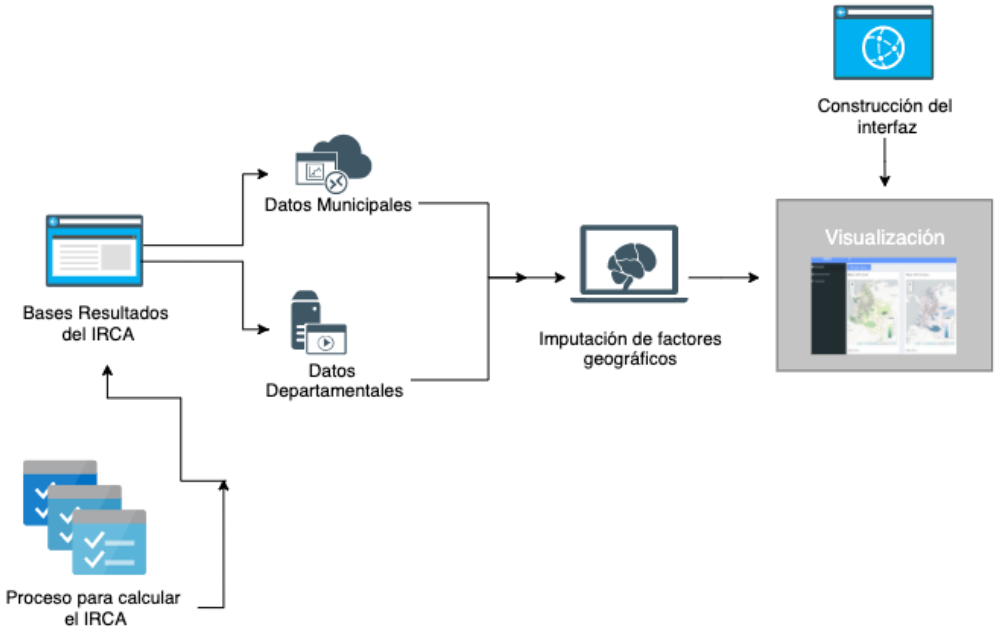


Ilustración 2. Flujo de información

b. Interfaz de la aplicación

A través de la herramienta Shiny apps de R se construyó la interfaz de la aplicación, esta requirió del lenguaje de JavaScript lo cual permite la adaptación de características de visualización de la App, como lo son los colores del Framework y el diseño de los botones dispuestos en la aplicación. En cuanto a la disposición de la App se usó una denominada dashboardPage(), la cual tiene como principal característica la generación de una barra lateral, el desarrollo central y el encabezado (Ilustración 3).

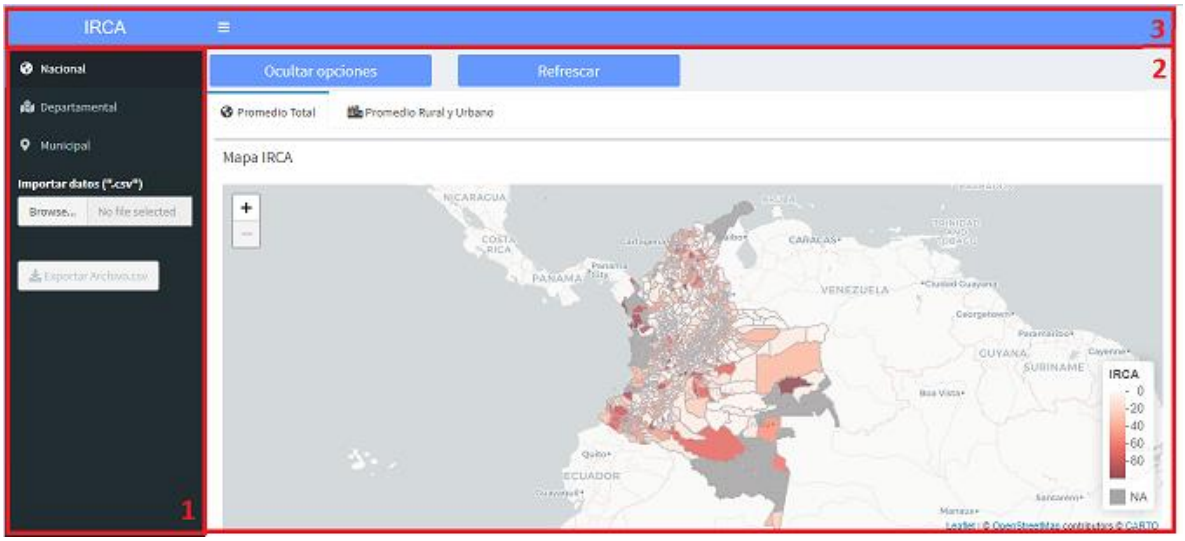


Ilustración 3. Disposición Aplicación Shiny R

Referente a la barra lateral (Ilustración 4) se disponen tres opciones de desagregación, correspondientes a nacional, departamental y municipal, estás podrán ser usadas para filtrar la información e identificar los indicadores precisa.

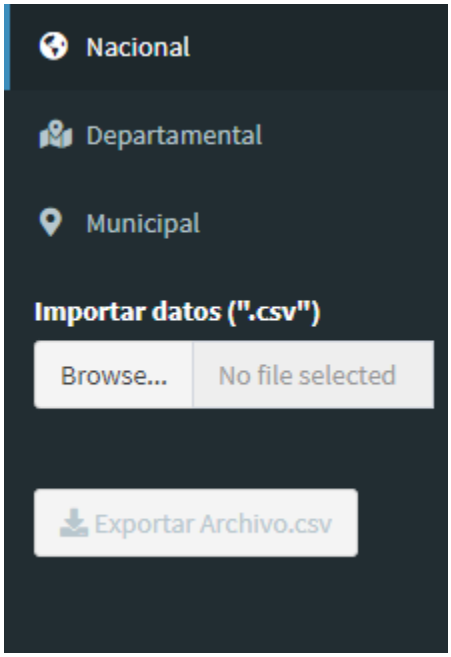


Ilustración 4. Barra Lateral

Adicionalmente la barra lateral cuenta con dos módulos más, los cuales son importar datos (“.csv”) y exportar archivo, el primero permite al usuario subir la base de datos del año que desea calcular y las especificaciones

de esta base de datos se podrán consultar en el manual del usuario. Por otra parte, luego de cargar la base de datos en formato ".csv" la aplicación empezará un proceso del cálculo del IRCA, en este punto el usuario deberá esperar 6 minutos (la aplicación muestra una notificación del tiempo de espera), este cálculo se efectúa bajo el lenguaje de programación Python y es ejecutado por R a través del paquete denominado "reticulate".

La importación de datos ".csv" realiza el remplazo del archivo ("sabana.csv"), el cual se encuentra en las carpetas de la aplicación, posteriormente al remplazo se ejecuta el código Python el cual tiene este archivo como referencia, al realizar los cálculos establecidos de manera exitosa se reemplaza el archivo de resultados ("sabana_calculada.csv") y al refrescar la página web, los mapas tendrán el cálculo actualizado y el usuario la posibilidad de descargar el archivo de resultados.

El segundo módulo hace referencia al botón de exportar archivo en .csv, este fue desarrollado a través de una función de Shiny y permite la obtención de los resultados arrojados por Python y reflejados en el mapa.

c. Mapas

La disposición central de la interfaz contiene dos pestañas Promedio Total (Ilustración 5) y Promedio Rural y Urbano (Ilustración 6), la primera contiene el mapa con los resultados del cálculo del IRCA ponderados por muestras entre rural y urbano y la segunda contiene dos mapas con los resultados del cálculo del IRCA diferenciados por rural y urbano. A medida que el usuario realiza cambios a través de las opciones presentes en el menú, los mapas podrán cambiar su foco mostrando una desagregación nacional, departamental o municipal.

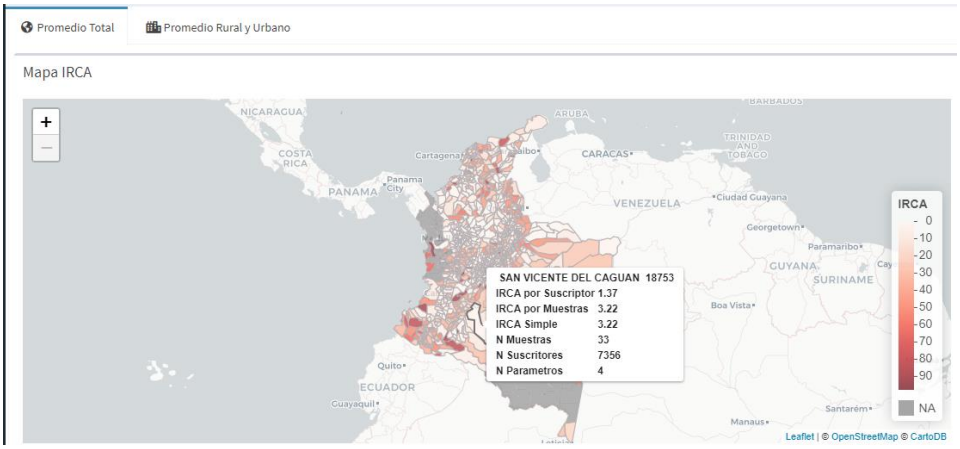


Ilustración 5. Disposición central Promedio Total

Para la visualización de los resultados se desarrolló una etiqueta usando el lenguaje de programación "HTML", está contiene la información de los cálculos por municipio de:

- IRCA por Suscriptor
- IRCA por Muestras
- IRCA Simple

- Número de Muestras
- Número de Suscriptores
- Número de Parámetros

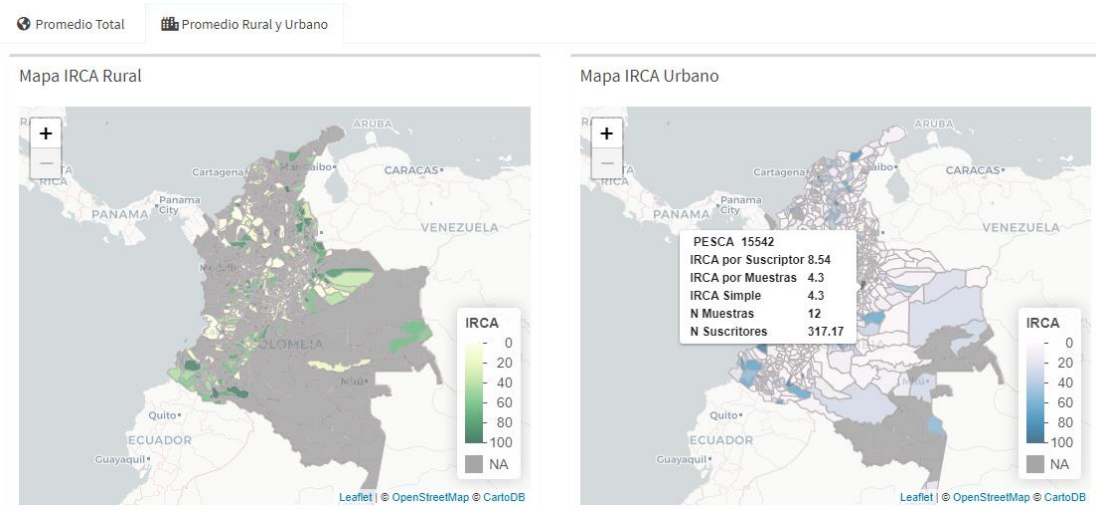


Ilustración 6. Disposición central promedio rural y urbano

Resultados

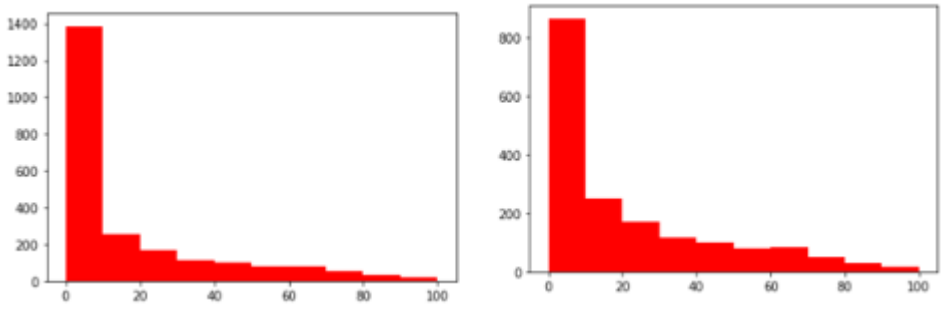
1. Procesamiento de los datos

Por medio de la información suministrada por la Dirección de Desarrollo Urbano se realizó la imputación y completitud de la base de datos la Sabana, con la cual se estableció el calculó del IRCA (promedio simple, ponderado por muestra y ponderado por suscriptor). A partir de dicha imputación se realizó la selección del modelo para calcular el número de suscriptores por empresa prestadora de servicios y posteriormente el cálculo del IRCA PROMEDIO PONDERADO POR NÚMERO DE SUSCRIPTORES, como base de la comparación se tomaron como verdaderos los cálculos del IRCA realizados por la DDU en el 2017

Tabla 4. Estadísticas descriptivas entre los resultados estimados y los de la DDU

		Estimación		**DDU**	
		Media	*Desviación*	*Media*	*Desviación*
Rural	*IRCA Promedio Simple*	29,22	25,13	22,93	24,41
	IRCA Ponderado por Muestra	29,22	25,13	10,34	19,90
	IRCA Ponderado por Suscriptores	26,21	24,61	1,85	7,70
Urbano	*IRCA Promedio Simple*	15,42	22,05	13,26	17,72
	IRCA Ponderado por Muestra	15,42	22,05	11,92	16,96
	IRCA Ponderado por Suscriptores	14,68	21,67	12,39	17,44

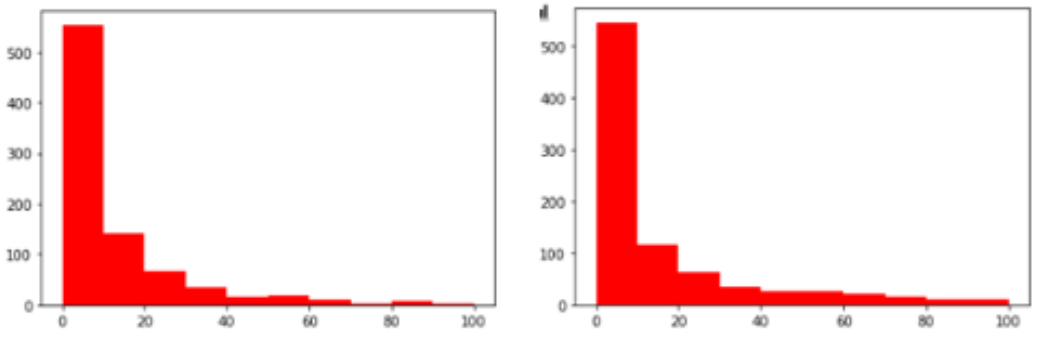
En la Tabla 4 se presentan las estadísticas descriptivas de los valores estimados por el modelo de regresión seleccionado y los valores estimados por la DDU para la sabana de datos del 2017. También se evidencia que la media y desviación estándar entre el IRCA promedio simple, IRCA ponderado por muestra e IRCA ponderado por suscriptores no varía de manera significativa entre los valores estimados de manera automática y los valores estimados por la DDU para el tipo de ubicación Urbano. Sin embargo, de acuerdo con las estimaciones la variación entre el estimado de manera automática y el valor estimado por la DDU para el IRCA Ponderado por Suscriptores es alto, debido a la completitud realizada para la columna ubicación, al igual se evidencia una variación alta entre los diferentes tipos de IRCA calculados por la DDU para el tipo de ubicación Rural, lo cual no se presenta en la ubicación Urbano.



(a) (b)

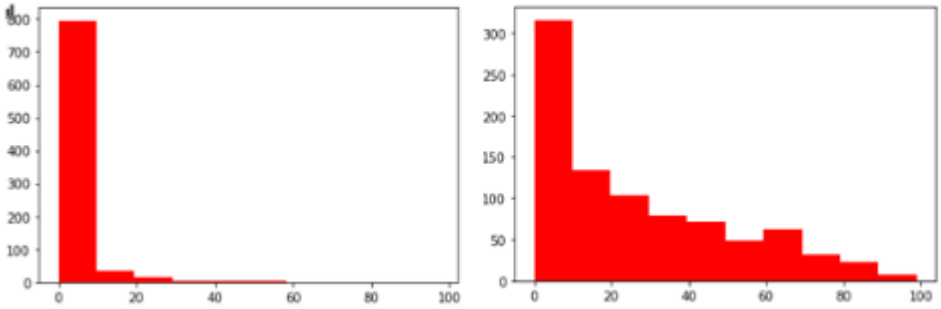
Ilustración 7. (a) Distribución del IRCA por suscriptor de la DDU, (b) Distribución del IRCA por suscriptor Estimado.

En la Ilustración 7, se evidencia que la distribución del IRCA por suscriptor calculado por la DDU y el calculado por el modelo seleccionado tienen una distribución similar. Por otro lado, al realizar una diferenciación entre el cálculo del IRCA urbano (Ilustración 8) e IRCA rural (Ilustración 9), se evidencia que la mayor similitud se obtiene en la zona urbana y una variación más significativa en la zona rural.



(a) (b)

Ilustración 8. (a) Distribución del IRCA por suscriptor DDU zona Urbana, (b) Distribución del IRCA por suscriptor Estimado zona Urbana



(a) (b)

Ilustración 9. (a) Distribución del IRCA por suscriptor DDU zona rural, (b) Distribución del IRCA por suscriptor Estimado zona rural

2. *Desarrollo de una aplicación web*

En la Ilustración 10 se presentan los dos mapas de la pestaña promedio rural y urbano, el mapa que se encuentra en la parte derecha ilustra el IRCA anual ponderado por suscriptor de las zonas urbanas, siendo el color blanco las zonas con mejor índice de calidad de agua para el consumo humano, por el contrario, entre más oscuro se torne el color azul la calidad del agua va disminuyendo. También se debe tener en cuenta que las zonas que se encuentran de color gris son porque se desconoce información de las muestras, esto dependerá exclusivamente de la calidad de la base de datos que sea importada en la aplicación. En cuanto al mapa que está ubicado en la parte izquierda representa el IRCA anual ponderado por suscriptor de las zonas rurales, la paleta de colores que usa podría ser explicada de la misma manera que el mapa urbano.

El mapa brinda información sobre las siguientes variables:

- El nombre del municipio y su respectivo código municipal
- IRCA anual ponderado por suscriptor
- IRCA anual ponderado por muestra
- IRCA promedio ponderado simple
- Número de muestras
- Número de suscriptores
- Número de parámetros promedio usados por municipio

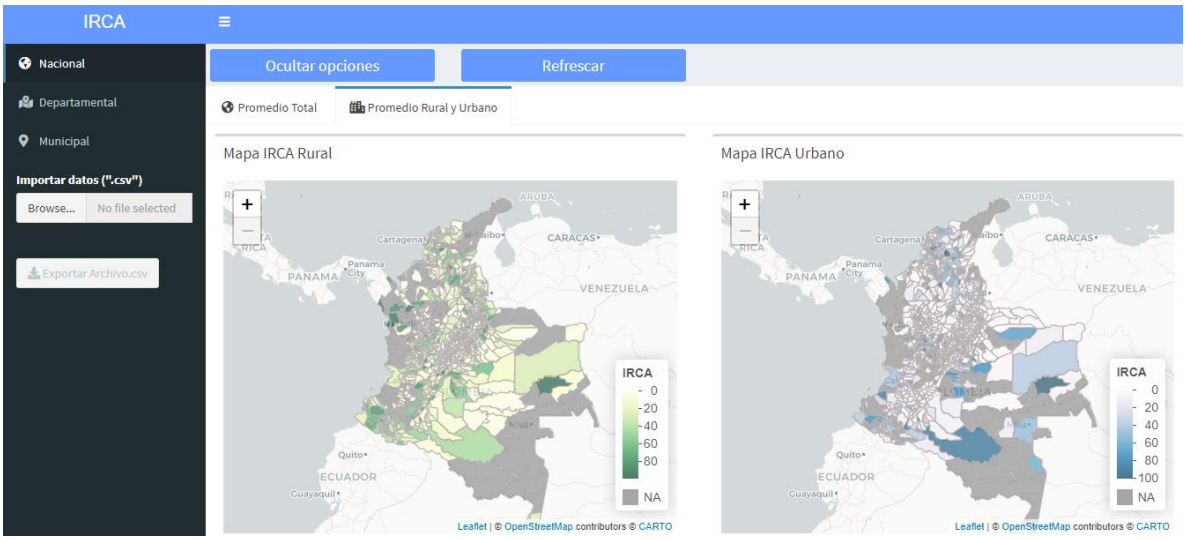


Ilustración 10. Mapa Nacional

Asimismo, la información disponible del tablero se puede visualizar a nivel nacional, departamental o municipal de acuerdo con los filtros utilizados. Véase en la Ilustración 11 e Ilustración 12.

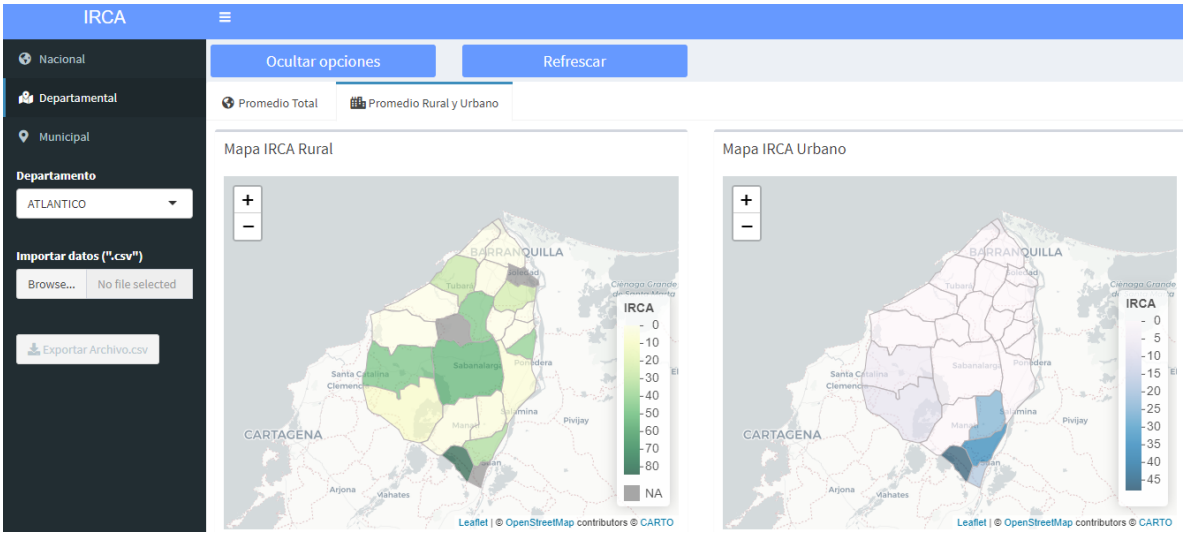


Ilustración 11. Mapa Departamental

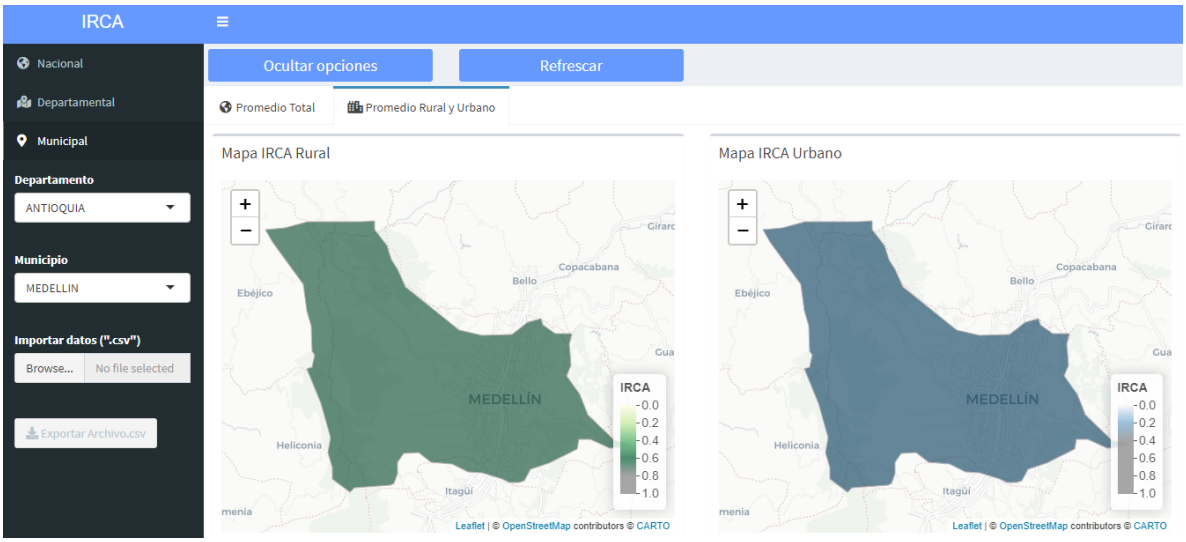


Ilustración 12. Mapa Municipal

Conclusiones y recomendaciones

1. Se implemento un script para el procesamiento de las bases de datos (limpieza de texto, completitud, imputación de Id por empresa, imputación del número de suscriptores, cálculo del IRCA promedio simple, IRCA ponderado por muestras e IRCA ponderado por suscriptor, identificación del número promedio de parámetros usados por municipio e identificación del número de muestras por municipio)
2. Se utilizó un modelo de regresión denominado Decision Tree Regresor para la imputación del número de suscriptores para empresas, dado los valores de entrenamiento y testeo usados se obtuvo un valor del R^2 de 99%.
3. Se obtuvieron distribuciones similares en los resultados generales del cálculo del IRCA, sin embargo, a causa de los datos faltantes y la completitud automática de la variable ubicación se obtiene una gran diferencia entre el cálculo del IRCA rural estimado automáticamente con los valores estimados por la DDU.
4. Los resultados del IRCA y algunos indicadores secundarios como lo son el número de muestras y el número de parámetros se puede evidenciar a través de la herramienta de visualización.
5. El usuario deberá tener en cuenta el formato de la base de datos que se le plantea en el manual del usuario, para que la aplicación funcioné adecuadamente al momento de importar nuevos datos.
6. Se recomienda obtener mejor información de las variables de interés para el cálculo, esto podría potenciar significativamente la mejora del desempeño del modelo.

Socialización

Los resultados del presente proyecto se socializaron con la DDU.