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 DESARROLLO DE LAS TIC REGIONAL PARA COLOMBIA

INFORME FINAL

Dependencias y entidades involucradas	Departamento Nacional de Planeación • Dirección de Desarrollo Digital (DDD) - Unidad de Científicos de Datos (UCD) • Dirección de Desarrollo Digital (DDD)
Sector	TIC
Tecnologías utilizadas	R, Shiny
Fuentes de datos	ECV, CNPV, Terridata, Colombia TIC

Contenido

1. Presentación 2
2. Objetivos del proyecto 2
3. Metodología 3
4. Resultados 6
5. Herramienta de visualización 9
6. Conclusiones y recomendaciones 10
7. Socialización 11
8. Contacto 11

Anexos 12

Anexo 1 Variables auxiliares 12

1. Presentación

El Índice de Desarrollo de las TIC Regional para Colombia (en adelante IDI Regional) realiza una medición comparativa del desarrollo de las TIC, en términos de acceso, uso y habilidades, para comprender de manera precisa las diferencias regionales en esta materia y, de esta forma, apoyar la priorización y focalización de recursos para el cierre de la brecha digital. Este índice es una adaptación metodológica del *ICT Development Index* (IDI) desarrollado por la Unión Internacional de Telecomunicaciones (ITU por sus siglas en ingles), por lo que se construye adaptando a las realidades nacionales los indicadores propuestos por esta organización en su metodología del 2008.

Debido a que la fuente principal de información para el cálculo del IDI regional es la Encuesta de Calidad de Vida, y teniendo en cuenta que el muestreo probabilístico de la misma no permite obtener estimaciones confiables a nivel de municipio, se hace necesario desarrollar una metodología para obtener estimaciones del índice a este nivel de desagregación. Es así como en este proyecto se desarrolló una metodología para hacer la estimación del índice a nivel municipal usando Análisis de Componentes Principales en el ajuste de modelos de Estimación de Áreas Pequeñas. Como resultado se obtuvieron estimaciones más precisas del índice en los municipios en donde se había recolectado información (viéndose reflejado en el coeficiente de variación) y estimaciones en los municipios en donde no se realizó recolección de información.

The Regional ICT Development Index for Colombia performs a comparative measurement of ICT development, in terms of access, use and skills, to accurately understand regional differences in this matter and, possibly, support the prioritization and targeting of resources to close the digital gap between regions. This index is a methodological adaptation of the ICT Development Index (IDI) developed by the International Telecommunications Union (ITU), which is why it is built by adapting the indicators proposed by this organization in its 2008 methodology to national realities.

Because the main source of information for calculating the regional IDI is the Quality of Life Survey, and taking into account that its probability sampling does not allow obtaining reliable estimates at the municipality level, it is necessary to develop a methodology to obtain estimates of the index at this level of disaggregation. Thus, in this project, a methodology was developed to estimate the index at municipality level using Principal Component Analysis in the fitting of Small Area Estimation models. As a result, more precise estimates of the index were obtained in the municipalities where information had been collected (being reflected in the coefficient of variation), and estimates in the municipalities where no information was collected.

2. Objetivos del proyecto

2.1. General

Realizar el acompañamiento técnico para estimar el Índice de Desarrollo de las TIC Regional para Colombia a nivel municipal a través de la técnica estadística de Estimación en Áreas Pequeñas, así como apoyar el proceso de desarrollo de un tablero de control para la presentación de resultados.

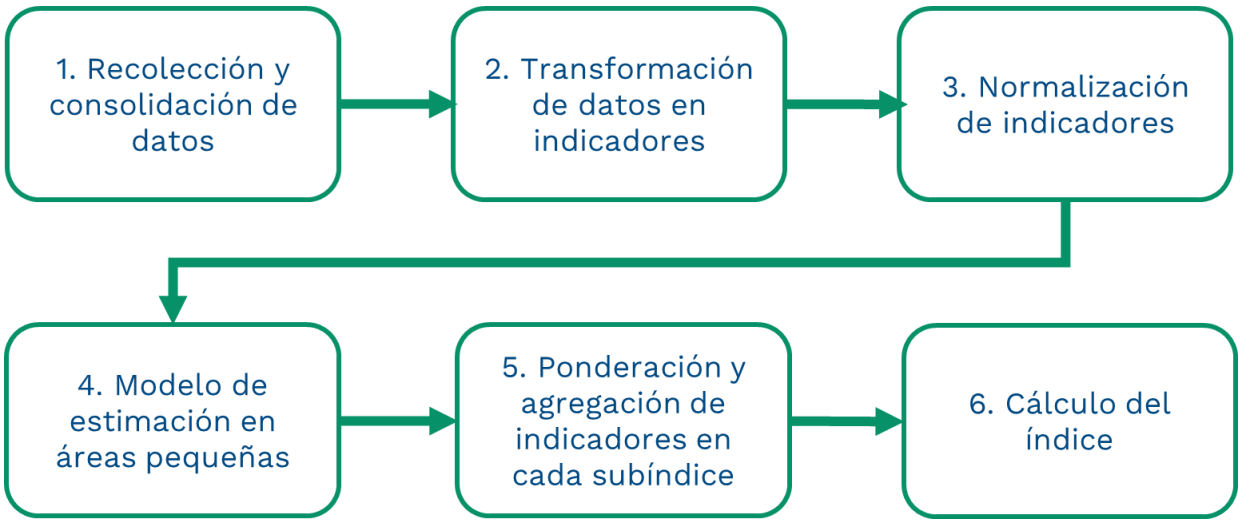
2.2. Específicos

1. Realizar un documento que describa los aspectos técnicos del modelo de Estimación de Áreas Pequeñas utilizado para el cálculo del Índice a nivel municipal.
2. Realizar el código en R, con su respectiva documentación, para hacer la estimación del Índice a nivel municipal, teniendo en cuenta opciones como el Análisis de Componentes Principales (ACP) para el uso de la información auxiliar.
3. Desarrollar el tablero de control para la presentación de los resultados del índice a nivel departamental y municipal.

3. Metodología

Para realizar la estimación a nivel municipal del IDI regional se modificó la metodología actual para calcular el índice, la cual consiste en seis pasos ilustrados en la Figura 1. Los tres primeros pasos consisten en la recolección de los datos desde las diferentes fuentes, y el cálculo y normalización de indicadores. El cuarto paso, el cual consistía en realizar la estimación municipal a partir de variables auxiliares, fue modificado para tener en cuenta la mayor cantidad de información posible. El quinto y sexto paso se mantienen aunque con algunos cambios en el orden de ejecución.

Figura 1. Metodología actual de cálculo del índice



Fuente: Índice de Desarrollo de las TIC para Colombia 2018-2019

El cambio de esta metodología surge de la necesidad de tener estimaciones municipales confiables y comparables en el tiempo. Para dar un poco de contexto de la metodología utilizada para los años 2018 y 2019, en total había 19 variables auxiliares para realizar el ajuste de un modelo de estimación de áreas pequeñas. Para hacer la selección de variables para el modelo, se optó por un proceso automático en donde se ajustaban diferentes modelos y se elegía el modelo con mejores métricas (puntualmente AIC, BIC y Rho para modelos espaciales). Sin embargo, debido a que la cantidad de posibles modelos era muy grande ($19! = 1.22 \times 10^{17}$), se decidió hacer la selección en modelos que consideraban hasta 3 variables en simultaneo. De esta manera, a pesar de que había mucha información en las variables auxiliares no era posible explotarla por temas computacionales.

Como solución a este problema, la nueva metodología consiste en hacer uso del Análisis de Componentes Principales (ACP) para hacer una reducción de variables y de esta manera pasar de 19 variables a un número menor de componentes, bajo el supuesto de expresar la mayor proporción de varianza posible. Para la creación de estos componentes se realizaron los siguientes pasos:

1. *Preprocesamiento:* como su nombre lo dice, este paso consiste en arreglar y procesar los conjuntos de datos para poder hacer el ACP y posteriormente la Estimación en Áreas Pequeñas (SAE). Está compuesto por los siguientes pasos:
   - *Exclusión de municipios*: Como el análisis se hace a nivel de municipio, se eliminan algunos municipios que no colindan con los demás municipios (San Andres y Providencia), y que por lo tanto no se pueden usar para

hacer la estimación de un modelo con autocorrelación espacial. Es importante notar que los modelos espaciales dependen de una matriz de pesos que se calculan a partir de fronteras de los municipios. Como San Andres y Providencia son islas, sus fronteras no se traslapan con ningún municipio por lo que no se pueden calcular los pesos para ellas. Por esta razón, se eliminan del análisis.

- *Variables auxiliares:* estas variables se usan para hacer la estimación SAE a nivel municipal y por lo tanto deben estar disponibles para todos los 1.122 municipios del país (incluyendo áreas no municipalizadas). La idea consiste en ajustar un modelo en donde la variable respuesta es el índice y las covariables son estas variables auxiliares o los componentes que salen de ellas. Este modelo se estima a partir de los municipios en donde sí hay información para calcular el índice, y posteriormente se usa para predecirlo en los municipios en donde no hay información suficiente. Generalmente los municipios en donde no se ha recolectado información en la Encuesta de Calidad de Vida (ECV), son los que necesitan la predicción de este modelo SAE. El procesamiento de esta información consiste en agregar etiquetas para facilitar su manipulación dentro del código e imputar información faltante en algunos municipios. En la Tabla 4 se pueden ver las variables consideradas
- *Consolidación de datos:* en este paso se hace la distinción entre los municipios en donde se tiene la estimación directa del índice[1], y los municipios en donde no. De esta manera, para cada año se obtienen dos conjuntos de datos diferentes.

2. *Exploración de los datos:* debido a que la estimación del IDI regional se hace para cada año, se realiza una exploración de la distribución de los variables en los municipios en donde sí se tiene estimación directa. Esta exploración consiste en graficar las densidades de las variables para los años analizados, graficar la distribución acumulada y hacer la prueba estadística *Kolmogorov-Smirnov* (KS) para identificar si hay diferencias estadísticamente significativas en la distribución de las variables a través de los años. Es importante aclarar que esta comparación de distribuciones a través de los años no es del todo justa, pues los municipios considerados por la ECV para un año son generalmente diferentes para otro. Sin embargo, la representación visual es útil para saber que tanto cambia una variable de un año a otro.

3. *Análisis de componentes principales:* como se mencionó anteriormente, la principal motivación para aplicar una técnica de reducción de dimensionalidad es sacar provecho de todas las variables auxiliares disponibles, pues en la metodología actual solo se usan tres variables de las 19 que hay en total (ver Tabla 4). Debido a que cada variable pertenece a una dimensión conceptual diferente, se decidió hacer el análisis de componentes de dos formas diferentes:
   - *Opción de precisión:* en esta opción se calculan los componentes usando todas las variables auxiliares simultáneamente, es decir, se obtienen máximo 19 componentes. Luego se elige una cantidad menor de componentes que expliquen una proporción deseable de varianza y posteriormente se ajustan los modelos SAE[2]. Para la selección del modelo final se consideran todas las posibles combinaciones de componentes principales, y se toma el que mejores métricas tenga (AIC, BIC y Rho para modelos espaciales). En esta opción se le da prioridad a tener un porcentaje alto de varianza explicada en la selección de la cantidad de componentes, y se deja a un lado la interpretabilidad de los resultados.

[1] Entiéndase estimación directa al cálculo del índice que proviene de la información de la ECV, y estimación SAE a la estimación del índice a partir del modelo mencionado.
[2] Entiéndase modelos SAE como el modelo Fay-Herriot (FH) y el modelo Fay-Herriot Espacial (SFH).

- *Opción de interpretabilidad:* en esta opción se realizan tres análisis de componentes correspondientes a las tres dimensiones definidas en las variables auxiliares (ver Tabla 4). En cada dimensión se busca obtener un(os) índice(s) compuesto(s) que se pueda interpretar desde el punto de vista económico. En la selección del modelo final se dará prioridad a la experticia temática con apoyo de las métricas que resulten.

4. *Estimación de modelos SAE:* una vez obtenidas las soluciones para cada una de las opciones en el paso de componentes principales, y teniendo como referencia las coordenadas de las proyecciones de los municipios en el nuevo plano factorial creado en cada una de ellas, se hace la estimación de los modelos SAE. En este caso se consideraron tres opciones:
   - *Opción de precisión:* En esta parte se toman los componentes seleccionados en el paso anterior y para la selección de variables se hace una búsqueda exhaustiva entre todos los posibles modelos que se pueden crear con estos componentes. Por ejemplo, si se tienen tres componentes principales (CP1, CP2 y CP3), entonces se estiman los modelos:
     - *IDI = CP1*
     - *IDI = CP2*
     - *IDI = CP3*
     - *IDI = CP1 + CP2*
     - *IDI = CP1 + CP3*
     - *IDI = CP2 + CP3*
     - *IDI = CP1 + CP2 + CP3*

     Es fácil darse cuenta como incrementa el numero de modelos a considerar cuando se aumenta la cantidad de componentes. Así que el mejor modelo se selecciona teniendo en cuenta los criterios de AIC, BIC y Rho para el modelo espacial.
   - *Opción de interpretabilidad:* En esta opción se toman los componentes seleccionados en cada dimensión y se hace la estimación del modelo. Esta opción se divide en dos partes: la primera en donde se estima un solo modelo teniendo en cuenta todos los componentes seleccionados en cada dimensión; y la segunda en donde se hace una búsqueda exhaustiva de componentes entre todos los posibles modelos que se pueden crear (similar a la opción de precisión).

5. *Selección del modelo final:* Una vez seleccionados los mejores modelos en cada una de las opciones, se hace una comparación de las estimaciones del índice y sus errores, ya sean expresados como error cuadrático medio (MSE en inglés) o coeficiente de variación, y se elige el modelo con errores más pequeños. Esta comparación se realiza mirando el promedio y las densidades de las estimaciones y los errores.
   Vale la pena mencionar, que este proceso de estimación de modelos se hace también con modelos que consideran la autocorrelación espacial entre municipios. Así que la selección de modelos se hace igualmente entre modelos con y sin el componente espacial.

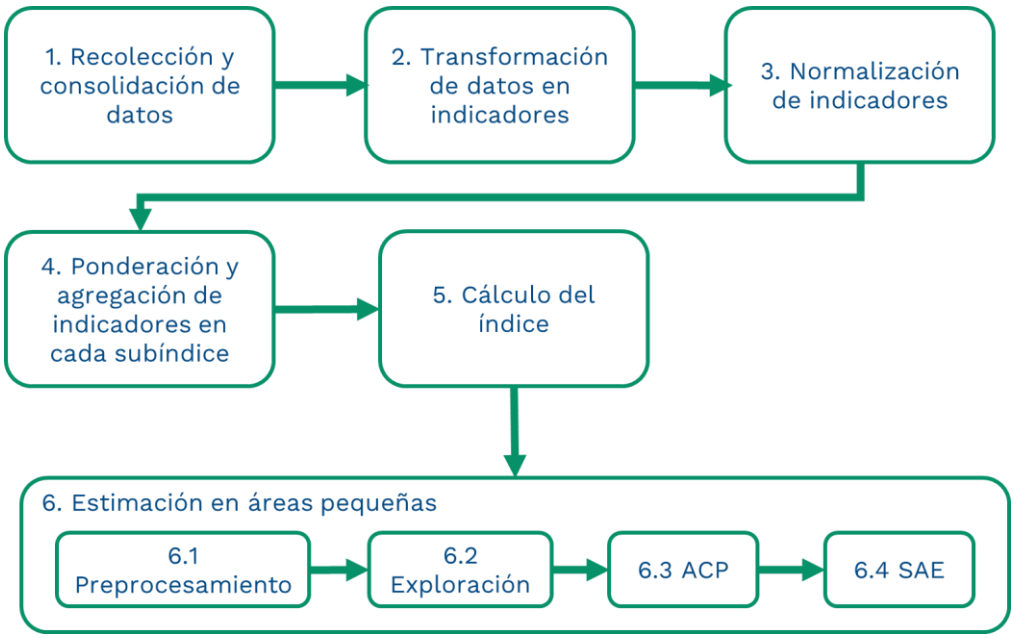
Debido a que estos pasos ya fueron realizados en el desarrollo de este proyecto, y que algunos de ellos no son necesarios para la estimación municipal del índice para el año 2020, se puede concluir que la metodología propuesta para la estimación del índice cambia de la siguiente manera (ver Figura 2):

1. *Recolección y consolidación de datos:* este paso se mantiene igual que en la metodología anterior, es decir, se hace la búsqueda de la información necesaria para hacer la estimación departamental y municipal (cuando se

puede) del índice. La base de datos más importante en este punto es la de la ECV para el año 2020, pues a partir de ella es que se calculan 16 de los 18 indicadores.

2. *Transformación de datos en indicadores:* este paso se mantiene igual que en la metodología anterior. El script para hacer esta transformación ya se encuentra realizado para los años 2018 y 2019, así que para el año 2020 se puede tomar alguno de estos como referencia y hacer las modificaciones pertinentes.
3. *Normalización de indicadores:* este paso se mantiene igual que en la metodología anterior. El script de normalización es el mismo de transformación de datos.
4. *Ponderación y agregación de indicadores en cada subíndice:* en la metodología anterior este era el quinto paso y se hacía después de la estimación SAE. Ahora es necesario realizar este paso antes de la implementación de SAE porque partiendo de los resultados de la estimación directa es que se puede estimar el modelo que permite tener el índice en los municipios sin cobertura por la ECV.
5. *Calculo del índice:* una vez ponderados los indicadores y agregados por dimensión se procede a hacer la estimación de índice en los municipios con información. Si el calculo de índice se hace a nivel departamental, aquí termina la metodología. Si por el contrario el índice se calcula a nivel municipal, se debe continuar con el siguiente punto.
6. *Modelo de estimación en áreas pequeñas:* tomando como referencia la estimación del índice en los municipios con cobertura por la ECV, se procede a hacer la estimación del índice en los municipios sin información (aunque con datos en las variables auxiliares). Este proceso se hace siguiendo los pasos mencionados anteriormente de preprocesamiento, exploración, ACP y modelo SAE.

Figura 2. Metodología propuesta para el cálculo del IDI regional



Fuente: Elaboración propia

4. Resultados

A través del desarrollo metodológico descrito en la Sección 3, se obtuvieron los resultados que se presentan a continuación. Toda retroalimentación desde un punto de vista experto o de usuario por parte de la DDD es

bienvenida. Este insumo será de gran ayuda para mejorar la calidad y utilidad de los resultados obtenidos, de manera que agreguen mayor valor.

4.1. ACP opción de precisión

Como se mencionó anteriormente, se desarrollaron dos opciones para el análisis de componentes principales. En la opción de precisión se decidió retener 10 componentes tomando como criterio de selección el porcentaje de varianza explicada. En la Tabla 1 se puede ver que estos componentes representan el 90.63% de la variabilidad encontrada en las variables auxiliares.

Vale la pena mencionar que para la selección de la cantidad de componentes a retener se consideró el gráfico de sedimentación (*scree-plot* en inglés) y la regla de los valores propios mayores a uno. Sin embargo, debido a que en esta opción no existía interés en interpretar los resultados, se decidió retener tantos componentes como fueran necesarios para asegurar una buena representación de los datos, lo cual se ve reflejado en el 90% de varianza acumulada.

Tabla 1. Proporción de varianza explicada por cada componente en el ACP de precisión

Componente	Valor propio	% varianza	% varianza acumulada
CP1	6.34	33.42	33.42
CP2	2.65	13.99	47.41
CP3	1.76	9.28	56.69
CP4	1.33	7.00	63.69
CP5	1.22	6.44	70.13
CP6	1.04	5.50	75.63
CP7	0.90	4.73	80.36
CP8	0.80	4.23	84.59
CP9	0.64	3.35	87.94
CP10	**0.51**	**2.69**	**90.63**
CP11	0.44	2.30	92.93
CP12	0.32	1.71	94.64
CP13	0.29	1.52	96.16
CP14	0.25	1.29	97.45
CP15	0.20	1.05	98.50
CP16	0.15	0.80	99.30
CP17	0.11	0.58	99.89
CP18	0.02	0.11	100.00

CP19	0.00	0.00	100.00

Fuente: Elaboración propia

4.2. ACP opción de interpretabilidad

En la opción de interpretabilidad se realizaron tres análisis de componentes principales correspondientes a las dimensiones de "Economía y población", "Servicios públicos y sociales" y "Desempeño fiscal". En cada uno de ellos se seleccionaron los componentes teniendo en cuenta que tuvieran algún sentido económico y que su interpretación fuera relativamente sencilla. Como resultado se seleccionaron tres componentes en la dimensión de economía, y dos componentes en las dimensiones restantes. En la Tabla 2 se pueden ver la proporción de varianza explicada en cada análisis.

Tabla 2. Proporción de varianza explicada por cada análisis en la opción de interpretabilidad.

	Economía y población			Servicios públicos y sociales			Desempeño fiscal		
Componente	**Valor propio**	**% varianza**	**% varianza acumulada**	**Valor propio**	**% varianza**	**% varianza acumulada**	**Valor propio**	**% varianza**	**% varianza acumulada**
CP1	2.37	33.86	33.86	3.26	54.45	54.45	3.13	52.18	52.18
CP2	1.38	19.79	53.65	**1.37**	**22.83**	**77.28**	**1.31**	**21.88**	**74.06**
CP3	**0.97**	**13.92**	**67.58**	0.91	15.11	92.39	0.87	14.59	88.64
CP4	0.81	11.58	79.16	0.33	5.42	97.81	0.45	7.57	96.22
CP5	0.67	9.64	88.80	0.13	2.19	100.00	0.20	3.40	99.62
CP6	0.59	8.41	97.21	0.00	0.00	100.00	0.02	0.38	100.00
CP7	0.20	2.79	100.00						

Fuente: Elaboración propia

4.3. Comparación de estimaciones entre opciones

Como se mencionó en la sección 3, se realizó la estimación municipal del índice utilizando tres tipos de modelos: 1) tomando las coordenadas de las proyecciones de las 10 primeras componentes y haciendo la selección del mejor modelo; 2) tomando las coordenadas de las proyecciones de 7 componentes y haciendo la estimación directamente; y 3) tomando las coordenadas de las proyecciones de 7 componentes y haciendo la selección del mejor modelo. Esto para los tipos de modelos: espacial y no espacial.

La decisión final respecto al mejor modelo se hizo tomando como referencia la estimación directa en los municipios con cobertura en la ECV, y comparando el promedio del error cuadrático medio y el coeficiente de variación. En la Tabla 3 se puede ver que el mejor modelo es el Tipo 2 espacial (modelo con las 7 componentes principales), porque es el que menor error cuadrático y coeficiente de variación promedio presenta. Es importante notar que las diferencias en estimación de los modelos considerados no son grandes pues en algunos casos se empiezan a ver diferencias en el cuarto decimal.

Tabla 3. Comparación de resultados entre modelos

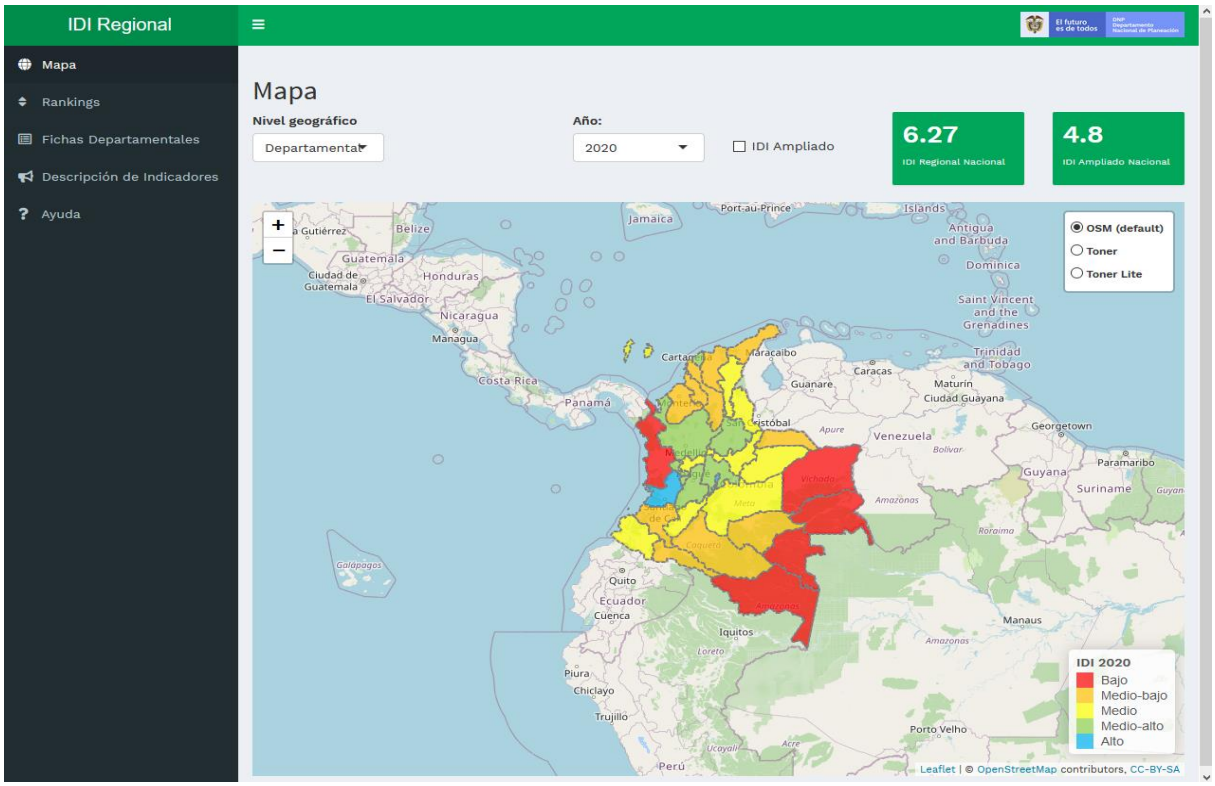
Modelo	Tipo de modelo	Estimación promedio	Error cuadrático medio	Coeficiente de variación
Estimación directa	N/A	3.2381	0.2383	7.9130
Tipo 1	No espacial	3.2261	0.0525	0.0731
Tipo 2	No espacial	3.2255	0.0517	0.0726
Tipo 3	No espacial	3.2255	0.0517	0.0726
Tipo 1	Espacial	3.2258	0.0514	0.0725
Tipo 2	**Espacial**	**3.2257**	**0.0506**	**0.0720**
Tipo 3	Espacial	3.2230	0.0511	0.0723

Fuente: Elaboración propia

5. Herramienta de visualización

Una vez desarrollada la metodología y aplicada para los años 2018 y 2019, se procedió a crear un tablero de control con el objetivo de visualizar los resultados obtenidos. Esta herramienta se compone de cinco pestañas en donde el usuario puede ver diferentes tipos de contenido (ver Figura 3).

Figura 3. Captura de pantalla del tablero de control



Fuente: Elaboración propia

Como las estimaciones para el año 2020 no están disponibles aún, se decidió dejar un par de scripts en R que sirven como plantilla para realizar todo el proceso de estimación municipal del índice y preparar la información para el tablero de control. En el documento de *Manual de Usuario* de la herramienta se menciona el paso a seguir para hacer esta actualización de resultados. Además, en dicho manual se encuentran los detalles y descripción de las características principales de la herramienta. De esta manera se garantiza que el equipo de la DDD pueda hacer la actualización y presentación de resultados para los tres años de manera sencilla y eficiente.

6. Conclusiones y recomendaciones

A partir de la metodología desarrollada y de los resultados obtenidos para cumplir los objetivos de este proyecto, planteados en el plan de trabajo, se presentan a continuación las principales conclusiones obtenidas por el equipo de la UCD y las principales recomendaciones para un mejor uso y aprovechamiento del proyecto.

1. La metodología propuesta en el desarrollo de este proyecto permite realizar la estimación municipal del IDI regional a través de componentes principales en modelos SAE.
2. El uso de Análisis de Componentes Principales permite aprovechar toda la información auxiliar existente en cada municipio.
3. El tablero de control permite a los usuarios visualizar los resultados obtenidos en los últimos tres años y a su vez descargar conjuntos de datos a nivel departamental y municipal.

4. Con la finalización de este proyecto, el equipo de la DDD tiene los insumos necesarios para realizar y visualizar las estimaciones del índice para el año 2020.
5. Se recomienda actualizar el Análisis de Componentes Principales con alguna periodicidad o si se implementan cambios en la metodología de cálculo del indicador.

7. Socialización

Los resultados de este proyecto se socializaron con la Dirección de Desarrollo Digital.

8. Contacto

Si tiene alguna duda, comentario o sugerencia sobre este proyecto, o si le gustaría conversar con la Unidad de Científicos de Datos sobre la posibilidad de una nueva fase para el mismo, puede comunicarse con nosotros a través del correo electrónico ucd@dnp.gov.co.

Anexos

Anexo 1 Variables auxiliares

Tabla 4. Variables auxiliares

Etiqueta	Variable	Dimensión
X1	Densidad poblacional	Economía y población
X2	Porcentaje población rural	Economía y población
X3	Distancia lineal a la capital del departamento - km (Kilómetros)	Economía y población
X4	Distancia lineal a Bogotá - km (Kilómetros)	Economía y población
X5	Actividades primarias	Economía y población
X6	Actividades terciarias	Economía y población
X7	Peso relativo municipal en el valor agregado departamental	Economía y población
X8	Cobertura acueducto	Servicios públicos y sociales
X9	Cobertura alcantarillado	Servicios públicos y sociales
X10	Cobertura de aseo	Servicios públicos y sociales
X11	Cobertura de energía eléctrica rural	Servicios públicos y sociales
X12	% Afiliados al régimen subsidiado	Servicios públicos y sociales
X13	% Afiliados al régimen contributivo	Servicios públicos y sociales
X14	Indicador de desempeño fiscal	Desempeño institucional
X15	Magnitud de la deuda	Desempeño institucional
X16	Dependencia de las transferencias	Desempeño institucional
X17	Dependencia de los recursos propios	Desempeño institucional
X18	Magnitud de la inversión	Desempeño institucional
X19	Capacidad de ahorro	Desempeño institucional

Fuente: DDD - DNP