Pulso Social Colombia

Realidades y perspectivas de las múltiples Colombias

Anexo Técnico

15 de septiembre de 2025

BID Mejorando vidas

UNIVERSIDAD EAFIT

Valor Público Centro de pensamiento e incidencia

Autores

Equipo BID:	Priscilla Gutiérrez Juárez (priscillag@iadb.org)
Equipo Universidad EAFIT	Mónica Hernández (mhernande6@eafit.edu.co) y Juan Carlos Muñoz-Mora (jmunozm1@eafit.edu.co)
Investigadores Junior:	Germán Angulo y Ana Pirela
Versión:	1.0

Índice general

Índice de figuras

1 Introducción Anexos

El presente anexo contiene la propuesta metodológica y resultados intermedios del informe de *Pulso Social* Colombia. El objetivo de la metodología utilizada fue brindar una revisión sistemática de los principales indicadores sociales de Colombia, siguiendo la metodología de Pulso Social propuesto por el Banco Inter-Americano de Desarrollo. La propuesta metodológico creado por la Universidad EAFIT y el Banco Inter-Americano de Desarrollo sigue cuatro principios rectores:

- **Transparencia**: Se pretende garantizar que los procedimientos y estrategias de análisis sean de código abierto.
- **Replicabilidad**: Estructurar protocolos de análisis de datos, siguiendo los parámetros internacionales que permitan la replicabilidad de las diferentes estrategias de análisis y visualización de datos.
- **Escalabilidad**: Permitir que los diferentes análisis pueden aumentar su incidencia en términos temporales y geográficos.
- **Sostenibilidad**: Determinar procesos factibles que garanticen la sostenibilidad del análisis en el tiempo y la posibilidad de integrar nuevos indicadores.

A partir de estos principios rectores, se inició con la consolidación de una infraestructura de datos que dieran la posibilidad de brindar análisis flexible de la situación social. Para su construcción se consultaron 32 bases de datos disponibles con variables relacionadas a la metodología de Pulso Social. Las bases tienen un **corte a diciembre 2021**, la lista completa está disponible en la libre de códigos.

Con fin de homogenizar el proceso de tratamiento se creó el *protocolo de manejo de datos del Pulso Social Colombia*, el recoge las políticas de datos abiertos y buenas prácticas del Banco Interamericano de Desarrollo (BID, 2018), el flujo de trabajo de Humans Rights Data Analysis Group (HRDAG) y las buenas prácticas para el manejo de códigos y datos en las ciencias sociales de Gentzkow & Shapiro (2014). Esto permiten una estructura que permite la replicabilidad para futuras actualizaciones y la transparencia en el manejo del datos. Todos los códigos, libro de códigos y datos fueron consignados en un repositorio localizado en la nube y de acceso abierto. El último paso de la infraestructura consiste en la creación de formas fáciles de acceso a la información (*frontend*), en particular se generaron tres formas de acceso:

- *Repositorio*: Repositorio abierto de los códigos y datos procesados durante el proceso. Los potenciales usuarios deben clonar el repositorio para utilizar los códigos de de procesamiento de datos

- *Página web*: Creación de un portal web donde se tiene acceso a toda la información del proyecto, presentaciones y demás reportes del proyecto.
- *Libraria Pulsosociacolombia*: Paquete de análisis de datos para el programa R, el cual permite el acceso a los datos básicos, análisis gráfico predeterminado y replicar los datos consignados en el informe.

La infraestructura de datos del proyecto del proyecto *Pulso Social* Colombia contiene la revisión de 404 indicadores los cuales provienen de 22 bases de datos. Esta estructura es el punto de partida para la estrategia de análisis. La figura 1 resume la infraestructura de datos de *Pulso Social* Colombia:

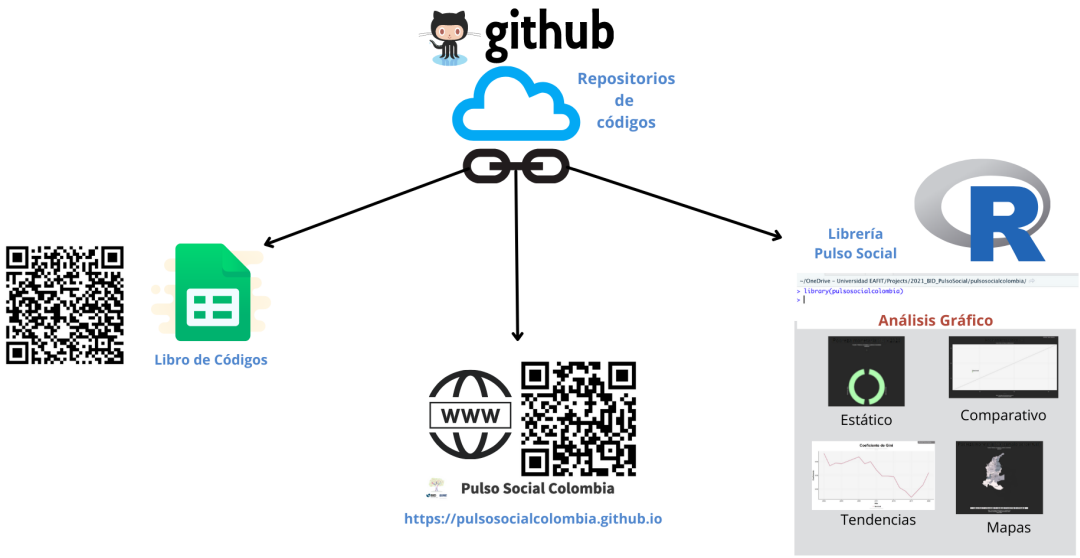

Figura 1 – Infraestructura de Datos de Pulso Social Colombia

Una vez consolidado la estructura de datos se inicia el proceso de análisis datos. El objetivo principal de la metodología *pulso social* es brindar una lectura de las oportunidades y desafíos sociales del país desde una visión territorial y diferencial. En particular, la metodología pulso social establece una perspectiva desde el ciclo de vida. La Estrategia se centra en la inversión en las personas, para desarrollar el capital humano, lo cual genera una gama de beneficios en las sociedades e implica continuidad de esta inversión a lo largo del ciclo de vida(?, ?).

Siguiendo la metodología propuesta por (?, ?), los indicadores consultados fueron agrupados en dos categorías: de contexto y resultados. Los primeros, hacen referencia a las condiciones generales del hogar, las cuales son esenciales para entender el contexto en que se elabora las políticas sociales. Esta dimensión, está compuesta a su vez por indicadores como: ingreso familiar, pobreza, desigualdad, características de la vivienda, acceso a servicios de agua potable y saneamiento, vulnerabilidad al cambio climático y covid. Por

su parte, los indicadores de resultados describen las situaciones sociales en áreas específicas a lo largo del ciclo de vida de los individuos y reflejan el contexto. Estos indicadores están agrupados en los diferentes ciclos de vida:

- *Infancia y niñez* mortalidad infantil, desnutrición crónica, asistencia escolar preescolar y primaria
- *Juventud* Asistencia escolar secundaria, fecundidad adolescente, desempleo e inactividad
- *Adultez* Desempleo, participación laboral femenina, contribución femenina al ingreso del hogar y empleo formal
- *Vejez* Pensiones, enfermedades no transmisibles y obesidad

La infraestructura de datos creada para Pulso Social Colombia, permite tener un alto número de indicadores por cada grupo de indicadores. Esto permite tener una visión completa y sistemática de cada una de las dimensiones y subdimensiones. Con el fin de facilitar el análisis a nivel departamental, se aplicó un método de reducción de dimensiones a las variables disponibles. Esto permite tener un índice que resume de manera clara, las características comunes de los indicadores que componen cada una de las dimensiones. En total, se crearon 13 índices que ofrecen una visión resumida de las condiciones sociales el país a nivel departamental.

El paso final de la estrategia de análisis fue la aplicación de un método de aprendizaje de máquina no supervisado, para identificar los patrones comunes en el territorio a partir de la lectura de los indicadores. Este análisis fue titulado "las múltiples colombias", al permitir observar las diferentes agrupaciones de las regiones a partir de una lectura sistemática de las dimensiones que componen la metodología de ciclo de vida. La siguiente figura resume la estrategia de análisis de datos empleada.

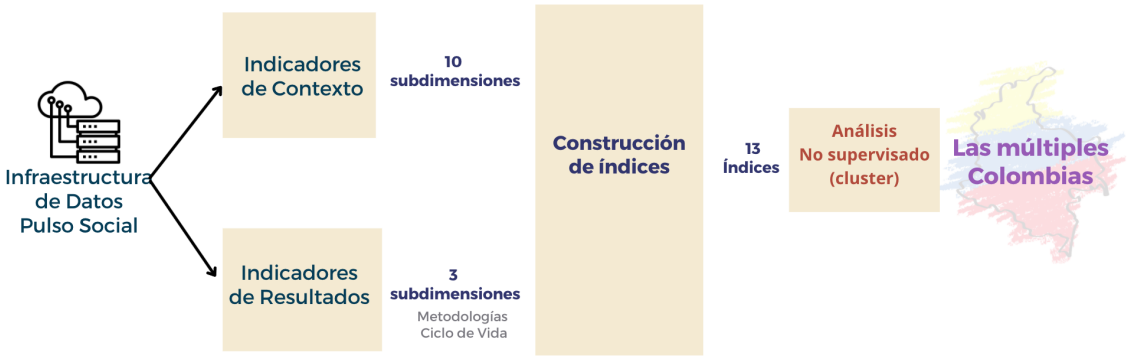

Figura 2 – Estrategia de análisis de datos

En las siguientes secciones se realiza una revisión detallada de los resultados y metodologías utilizadas en cada uno de los pasos de la metodología propuesta.

2 Protocolo de datos Pulso Social

Este documento detalla el protocolo de manejo de datos siguiendo la política de datos abiertos y buenas prácticas del Banco Interamericano de Desarrollo (BID, 2018), el flujo de trabajo de Humans Rights Data Analysis Group (HRDAG) y las buenas prácticas para el manejo de códigos y datos en las ciencias sociales de Gentzkow & Shapiro (2014).

2.1 Datos y códigos abiertos

Para garantizar la integridad, accesibilidad y replicabilidad de los datos empleados en el análisis y los resultados obtenidos, todos los datos y códigos se encuentran disponibles en el repositorio de GitHub “BID_Pulso_Social”. Los datos utilizados son escogidos y procesados de forma que se garantiza que sean exhaustivos e interoperables:

- Los datos son exhaustivos: se emplean bases de datos de fuentes oficiales desagregadas al menor nivel posible (género, grupo etario, raza/etnia, territorial).
- Los datos son interoperables: las bases de datos son procesadas y guardadas en una estructura estándar que permite combinar distintos tipos de datos.

Adicionalmente, las bases de datos con las que se elaboran los resultados y análisis finales son guardadas en formato .csv, de manera que sean reconocibles por múltiples herramientas en varios lenguajes de programación.

2.2 Flujo de trabajo

2.2.1 Directorios generales

El análisis empírico consiste en el procesamiento de datos y construcción de estadísticas descriptivas, organizados en dos directorios: "Dataz "Descriptives".

- "Dataçontiene el procesamiento de las bases de datos originales para construir bases estandarizadas que servirán de insumo para las estadísticas descriptivas ("Descriptives").
- "Descriptivesçontiene códigos que usan las bases estandarizadas de "Data"para elaborar bases, gráficas, tablas y demás elementos para el análisis descriptivo.

2.2.2 Procesamiento de datos

El procesamiento de una fuente de datos concreta es una tarea. Cada tarea consta de 4 carpetas: Docs, Input, Output y Src (source). Las carpetas funcionan de la siguiente forma:

- La documentación de los datos, diccionarios y otros tipos de documentos de referencia están en "Docs/"
- Los datos originales por leer estan en Ïnput/"
- Los códigos que procesan los datos originales, y construyen bases, graficas, tablas y demás, están en "Src/"
- Los resultados del procesamiento (bases, gráficas, tablas) están en .ºutput/"

Los archivos de Ïnput/"son únicamente de lectura, nunca deben sobrescribirse. En lo posible, tampoco deben cambiarse sus nombres originales o formato, de manera que el código los lea tal cual como fueron descargados de la página web o fuente, sin pasos intermedios. Esto permite replicar los resultados con mayor facilidad. Los códigos de "Src/"leen los archivos en Ïnput/"para crear bases de datos, gráficas, tablas y demás elementos que serán guardados en .ºutput/". Los códigos también pueden llamar elementos de .ºutput/": por ejemplo, tras guardar una base de datos, otro código puede llamarla para construir gráficas. Los resultados de los códigos de "Src/"son guardados en .ºutput/". La Figura 1 muestra un ejemplo del flujo de trabajo para una base de datos en particular, “Ejemplo_base_datos”. El procesamiento de esta base de datos se encuentra en el directorio “Data/Ejemplo_base_datos/” y se realiza con los códigos de “Src/” a partir de la lectura de los datos originales de “Input/”, y los resultados se guardan en “Output/”. Figura 1. Flujo de trabajo

2.2.3 Tareas autocontenidas y autodocumentadas

Al organizar cada tarea en las 4 carpetas, la tarea se convierte en autocontenida y autodocumentada:

- La tarea está autocontenida porque todas las transformaciones relacionadas con esa base de datos, desde el procesamiento del dato original (raw), hasta la construcción de otras bases, tablas y gráficas, se desarrolla y guarda en un mismo directorio: "Data/Ejemplo_base_datos/
- La tarea está autodocumentada porque todos los documentos y diccionarios relacionados con ella están en la carpeta "Docs/ todos los códigos están en "Src/". Todas las transformaciones realizadas a los datos puros están en "Src/ no se realizan cambios en hojas de cálculo de excel ni otros tipos de software que dificulten la trazabilidad en la construcción de los resultados finales.

2.2.4 Construcción de estadísticas descriptivas

El directorio “Descriptives” se encarga de leer las bases estandarizadas de “Data” y construir con ellas estadísticas descriptivas, como tablas, gráficas y demás elementos.

- En este directorio no se realiza procesamiento de datos puros, sino que se parte de los datos estandarizados, que pueden seguir transformándose para construir resultados.
- "Descriptives" también se compone de tareas, en las que se parte de una base de datos en particular y con ella se construyen estadísticas descriptivas. Por ejemplo, "Descriptives/Ejemplo_tasa_desempleo" tiene las carpetas "Input", "Output", y "Src".
- En "Input" se guardan bases de datos transformadas para producir gráficas, tablas y otros elementos.
- En "Output" se guardan resultados: gráficas, tablas y otros elementos de estadística descriptiva.
- "Src" contiene los códigos que crean bases de datos para las descriptivas, y que construyen gráficas, tablas y demás elementos

2.3 Manejo de códigos y datos

El manejo de datos tiene los objetivos de transformar los datos puros a un formato estándar, remover datos atípicos, corregir errores en las variables, generalizar los nombres de las bases de datos y de las variables y emparejar bases.

2.3.1 Nombres de carpetas y elementos

- Las carpetas se nombran con mayúscula en la primera letra y los espacios entre palabras se reemplazan con barra al piso. Por ejemplo: "Ejemplo_base_datos/".
- Las bases de datos, gráficas, tablas y otros resultados que provengan del código se nombran en minúsculas y los espacios entre palabras se reemplazan con barra al piso. Por ejemplo: "base_tasa_desempleo_municipal.rds", y "grafica_tasa_desempleo_municipal.jpeg".
- Las bases de datos, gráficas, tablas y otros que quieran hacer énfasis en los periodos separan los años con un guion. Por ejemplo: "base_tasa_desempleo_municipal_2010-2012.rds", "base_tasa_desempleo_municipal_2012-2020.rds", "grafica_tasa_desempleo_municipal_2010-2020.jpeg".
- Las bases de datos, gráficas, tablas y demás elementos construidos con el código indican qué tipo de elemento es (una base de datos, una gráfica, una tabla), una descripción de su contenido (desempleo, pobreza, educación), el nivel de desagregación (nacional, departamental, municipal), y cuando es necesario, el periodo para el que están disponibles los datos (2010-2020, 2012-2014).

2.3.2 Nombres de códigos y sintaxis

- Los códigos se nombran siguiendo un orden alfanumérico que permite identificar el orden en que se ejecutan. Por ejemplo: 01a_procesar_geih.R, 01b_procesar_ecv.R, 02a_indicadores_geih.R, 02b_indicadores_ecv.R.
- Cada código incluye una descripción corta de las acciones que realiza y la fecha en que fue editado por última vez.
- Las variables y objetos de los códigos se nombran en minúsculas con barra al piso para separar palabras. Por ejemplo: data_desempleo, tasa_desempleo.

3 Construcción de índices por dimensión

El análisis de reducción de dimensiones busca simplificar el análisis de las dimensiones de *Pulso Social*, a través de construir índices que permiten identificar las principales diferencias o brechas territoriales. Para esto, se emplea métodos de aprendizaje de máquina no supervisados que buscan aprender, desde los datos, cuáles son las asociaciones naturales a partir del comportamiento de estos. En particular, se utilizarán métodos de reducción de dimensiones que buscan transformar un gran conjunto de variables, en uno más pequeño que siga conteniendo la mayor parte de la información inicial.

Con la implementación de estos métodos se sacrifica precisión al no contar con la información completa, pero se gana simplicidad al poder contar con índices fáciles de explorar y visualizar. Existe diferentes métodos de reducción de dimensiones: Análisis de Componentes Principales (ACP), Mínimos Cuadrados Parciales, Análisis de Correspondencias y Uniform Manifold Approximation and Projection (UMAP), entre otros. Dada la naturaleza de los datos de los indicadores, todos ellos de naturaleza continua, se realizó una comparación entre los resultados de ACP y UMAP, obteniendo resultados similares (?, ?). De esta manera, y gracias a las simplicidad de su inferencia se decidió realizar en análisis con el método de Análisis de Componentes Principales usando el paquete *tidymodels* de R.

En el caso de proyecto *Pulso Colombia* el procedimiento para la realización de ACP fue el siguiente:

- **Selección de variables:** La infraestructura de datos permite recoger información de diferentes fuentes de información, con diferente disponibilidad temporal y espacial. De esta manera, se procedió a construir una base de datos que permitiría recoger la mayor parte de la información disponible para el territorio y dimensiones. Así, se tuvieron en cuenta los siguientes criterios para construir la base de datos:
  - *Temporal*: Se tomó la última información disponible por cada uno de los indicadores.
  - *Geográfico*: Dado que las unidad de análisis de las diferentes encuestas varía, se tomó la decisión a nivel departamental. En los casos en que la información departamental no estuviera disponible, se procedió a tomar la información de la capital.
  - *Disponibilidad*: Con el fin contar con la mejor información posible por cada una de las dimensiones, se priorizaron variables que tuvieran información en más del 80 % del territorio.
- **Imputación de datos:** Uno de los principales problemas de las variables sociales es que en algunos casos no está disponible para todas los territorios. Con el fin de

ofrecer una visión completa de las condiciones territoriales, se procedió a realizar un proceso de imputación que permitiera asociar el valor más probable utilizando toda la información disponible. Estos métodos permiten tener información balanceada, sin perder parte el nivel de inferencia. El método utilizado fue el siguiente:

- Imputación basado en el vecino más cercano (Nearest Neighbor, NN), el cual es un método eficaz para imputar los datos que faltan. En los casos que no existe información, el valor se sustituye por un valor obtenido a partir de casos relacionados en todo el conjunto de registros. Este método ha demostrado un alta capacidad de predicción, preservando la estructura de los datos (?, ?).
- En los casos en que la información es imputada se incluye una nota sobre ellos.
- **Estandarización de los datos:** El primer paso consistió en normalizar el rango de las variables continuas iniciales para que cada una de ellas contribuya por igual al análisis, eliminando la posibilidad de tener sesgos asociados a las magnitudes. Dados los diferentes tipos de variables, se realizaron dos tipos de estandarización consecutivos:
  - (1) Transformación orderNorm: $g(x) = \Phi^{-1}((rank(x)-,5)/(length(x)))$. Donde Φ se refiere a la CDF normal estándar, $rank(x)$ se refiere al rango de cada observación y $length(x)$ es el número de observaciones.
  - (2) Posteriormente todas las variables fueron centradas para garantizar que tengan desviación estándar 1 y media cero.
- **Construcción de la matriz de covarianzas:** Después de tener la información estandarizada, se procede a construir la matriz de covarianzas, lo que permite identificar no solo la composición de las variables totales sino también la correlación entre ellas. La idea es identificar qué información es redundante o cuales variables explican gran parte de la variación.
- **Estimación de los valores y vectores propios:** Desde el punto de vista matemático, los componente principales surgen de la estimación de los valores y vectores propios de la matriz de covarianzas. A partir de estos vectores surgen los componentes principales, los cuales son combinaciones lineales de las variables iniciales, de manera que no estén correlacionadas entre ellas y que capturen la mayor parte de la información de las variables iniciales que se exprime o comprime en los primeros componentes. Es decir, los componentes principales representan las direcciones de los datos que explican una cantidad máxima de varianza.
- **Selección de Componentes Principales:** Con el fin de contar con índices con alto nivel de explicación de la varianza total, se decidió que el valor de varianza mínima aceptada será *60 %*. Es decir, en aquellas dimensiones donde el primer componente explicara más del 60 % se eligió solo un componente; en los demás, se eligió más de un componente hasta garantizar el porcentaje mínimo aceptado.

Como resultado, se logró la construcción de 13 índices para cada una de las subdimensiones, divididos en los indicadores de contexto y resultado. A continuación se revisan los principales resultados para cada uno de las dimensiones.

3.1 Indicadores de contexto

3.1.1 Crecimiento económico y productivo

Esta dimensión está compuesta por 6 variables que contienen información, principalmente, sobre valor agregado, producción bruta, establecimientos, índice de diversidad económica, personal promedio por empresa y promedio de intensidad lumínica. Cada una de las variables fueron centradas y se realizó el análisis de Componentes principales. Los siguientes gráficos muestran las correlaciones de cada uno de las variables con los componentes (dimensiones), resultado del análisis de ACP:

Cuadro 1 – Estadísticas descriptivas

Statistic	N	Mean	St. Dev.	Min	Max
Índice de diversidad económica	32	18.490	7.883	10.404	36.497
Establecimientos	25	285.360	547.950	4	2,317
Personal promedio por empresa	25	88.482	43.708	19.000	191.357
Producción bruta	25	13,105,635,657.000	17,663,721,351.000	5,616,680	55,572,141,464
Valor agregado	25	4,678,038,365.000	6,697,888,992.000	2,840,397	21,868,265,750

Las principales estadísticas descriptivas indican que, en promedio, el índice de diversidad económica de los departamentos es 16.6 y tienen una producción bruta de más de 10 billones de pesos. Así mismo, el valor agregado, producción bruta, los establecimientos y el personal promedio por empresa tienen un peso negativo sobre la primera dimensión y positivo sobre la segunda, mientras que la diversidad económica tiene un peso positivo sobre la primera y la intensidad lumínica un peso negativo sobre la segunda. Adicionalmente, cabe destacar que el primer componente explica en más de 60 % la varianza del índice.

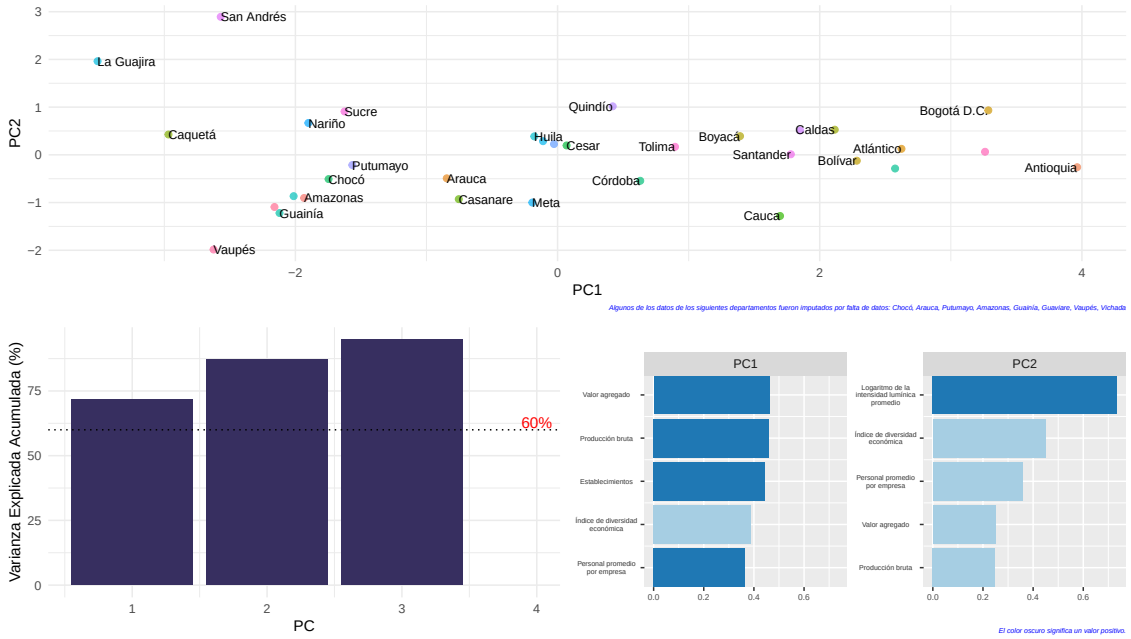

Figura 3 – Resultados de Análisis de Componente Principal - Crecimiento

3.1.2 Acceso a servicio de agua potable y saneamiento

Esta dimensión está compuesta por 5 variables que contienen información sobre acceso a servicio de agua potable y saneamiento, principalmente en cuanto a fuentes de agua como acueducto público, pozo bomba, acueducto comunal o carrotanque y condiciones de saneamiento dentro de la vivienda. Cada una de las variables fueron centradas y se realizó el análisis de Componentes principales. Los siguientes gráficos muestran las correlaciones de cada uno de las variables con los componentes (dimensiones), resultado del análisis de ACP:

Cuadro 2 – Estadísticas descriptivas

Statistic	N	Mean	St. Dev.	Min	Max
Fuente de agua de calidad alta	33	76.166	23.185	0.911	98.981
Fuente de agua de calidad baja	33	15.488	20.193	0.019	88.950
Fuente de agua de calidad media	33	8.346	10.501	0.386	43.552
Saneamiento dentro de la vivienda	33	72.987	22.591	13.364	99.024

Las principales estadísticas descriptivas indican que, en promedio, más de la mitad de los departamentos tienen como alternativa para abastecimiento de agua potable el acueducto público, 10.7 % tienen acueducto comunal, 6.8 % pozo bomba y 0.8 % carrotanque. Así mismo, las fuentes de agua como acueducto público, comunal o carrotanque y el saneamiento dentro tienen peso positivo sobre el índice, mientras que el uso de pozo bomba tiene un peso negativo en el primer componente. En cuanto a la segunda dimensión, todas las variables, con excepción del saneamiento dentro, tienen efecto negativo sobre

el índice. Cabe destacar que el primer componente no logra explicar ni en un 50 % la varianza del índice, pero el acumulado entre las dos sí lo hace.

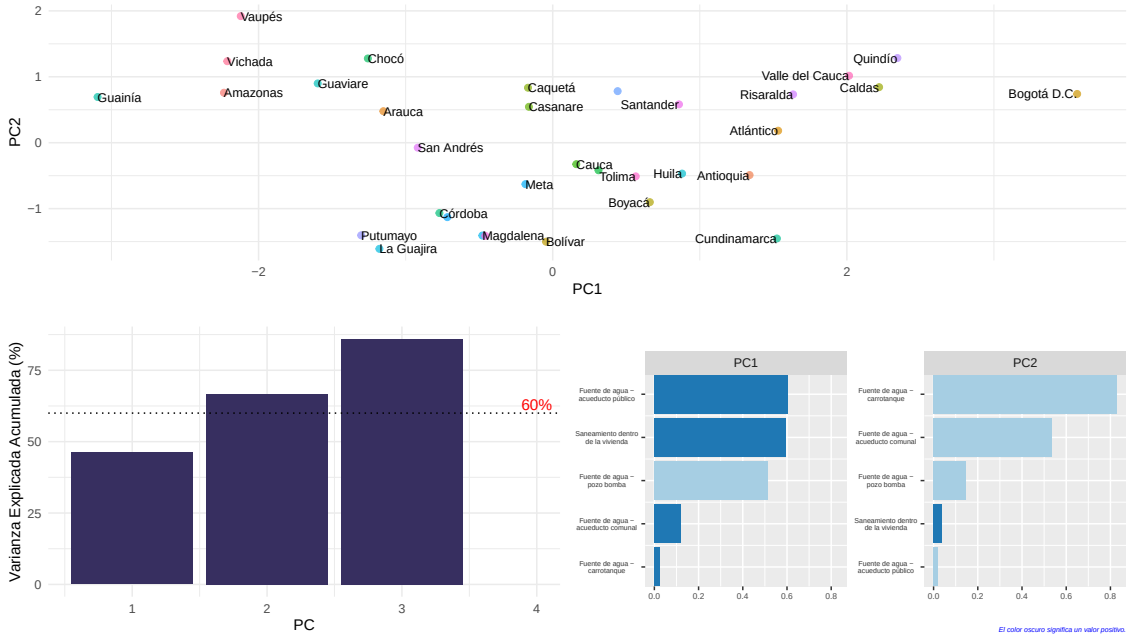

Figura 4 – Resultados de Análisis de Componente Principal - Agua Potable y Saneamiento

3.1.3 Covid

Esta dimensión está compuesta por variables que contienen información sobre casos reportados y fallecidos por Covid-19. Cada una de las variables fueron centradas y se realizó el análisis de Componentes principales. Los siguientes gráficos muestran las correlaciones de cada uno de las variables con los componentes (dimensiones), resultado del análisis de ACP:

Cuadro 3 – Estadísticas descriptivas

Statistic	N	Mean	St. Dev.	Min	Max
Casos totales de Covid - 19	33	818.672	606.036	0.036	2,219.207
Porcentaje de fallecidos de Covid - 19	33	5.277	17.044	0.000	100.000
Fallecidos por Covid - 19	33	20.772	15.148	0.000	54.499

Las principales estadísticas descriptivas indican que, en promedio por departamento se han reportado 818.672 casos, donde Bogotá D.C ha sido la ciudad con más casos. Por otro lado, se han reportado, en promedio 20.772 fallecidos por Covid, lo que corresponde a un porcentaje de fallecidos de 5 %. Todas las variables tienen un peso negativo sobre el índice y la primera dimensión explica en más de un 60 % la varianza del índice.

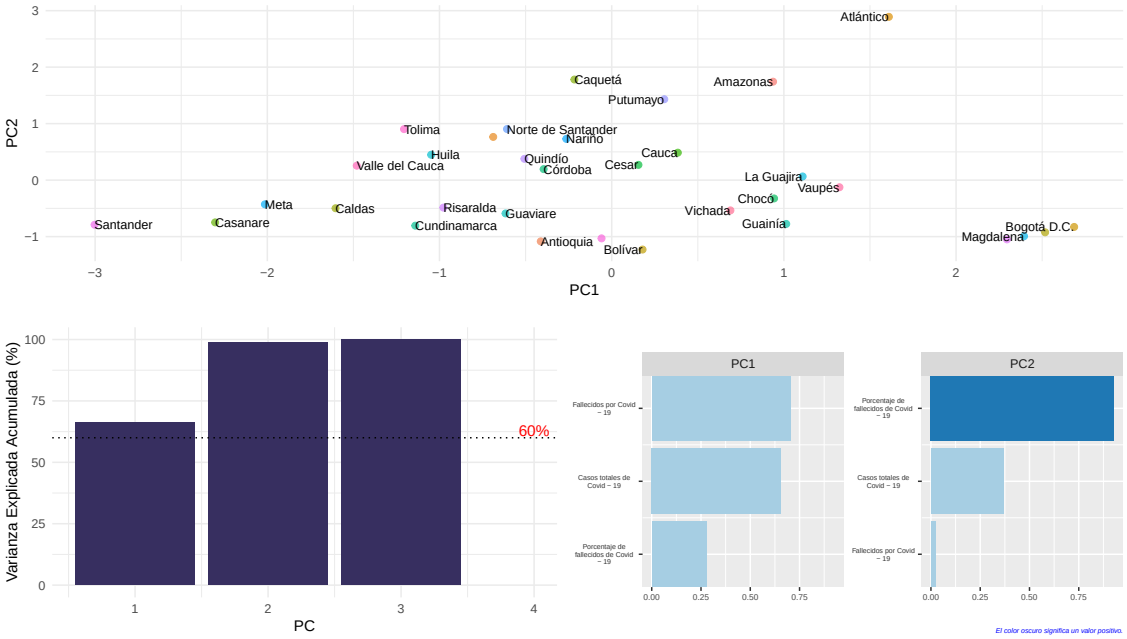

Figura 5 – Resultados de Análisis de Componente Principal - Covid

3.1.4 Cambio climático

Esta dimensión contiene información sobre la vulnerabilidad climática de los departamentos. El siguiente gráfico muestra la distribución a nivel nacional del índice por cuartiles. Los resultados indican que, en promedio por departamento, el índice de vulnerabilidad climática es 7 %, interpretado como el porcentaje del área considerada como vulnerable al cambio climático. Así mismo, los departamentos con mayor vulnerabilidad climática son Guaviare, Tolima, Quindío, Cundinamarca, Boyacá, Bogotá D.C, Santander y Sucre.

Cuadro 4 – Estadísticas descriptivas

Statistic	N	Mean	St. Dev.	Min	Max
Indice de Vulnerabilidad Climatica	33	7.013	20.043	0.0001	98.557

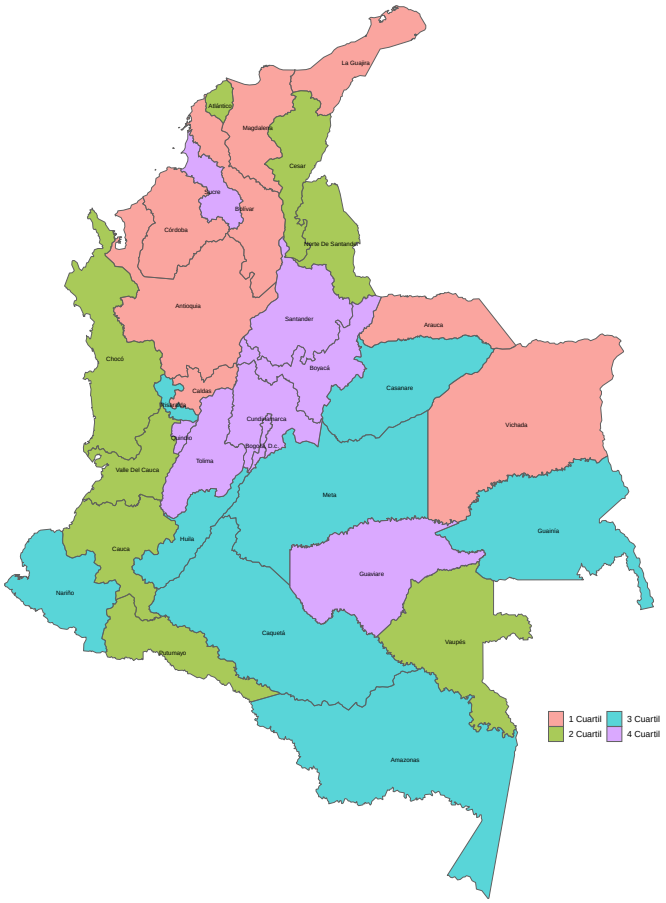

Figura 6 – Resultados de Análisis de Componente Principal - Cambio Climático

3.1.5 Pobreza

Esta dimensión está compuesta por 5 variables que contienen información sobre la pobreza en los departamentos, principalmente en cuanto a las personas con Necesidades Básicas Insatisfechas, personas en condiciones de pobreza monetaria y pobreza extrema. Cada una de las variables fueron centradas y se realizó el análisis de Componentes principales. Los siguientes gráficos muestran las correlaciones de cada uno de las variables con los componentes (dimensiones), resultado del análisis de ACP:

Cuadro 5 – Estadísticas descriptivas

Statistic	N	Mean	St. Dev.	Min	Max
IPM (					

Las principales estadísticas descriptivas indican que, en promedio por departamento, el

porcentaje de personas con NBI es de 24.2 %, en condiciones de pobreza monetaria es 46 % y en condición de pobreza extrema es 17.4 %. Así mismo, las tres variables tienen un peso positivo sobre la primera dimensión del índice y solo el porcentaje de personas con NBI lo tiene en el segundo componente. Cabe destacar que la primera dimensión explica en más de un 70 % la varianza del índice.

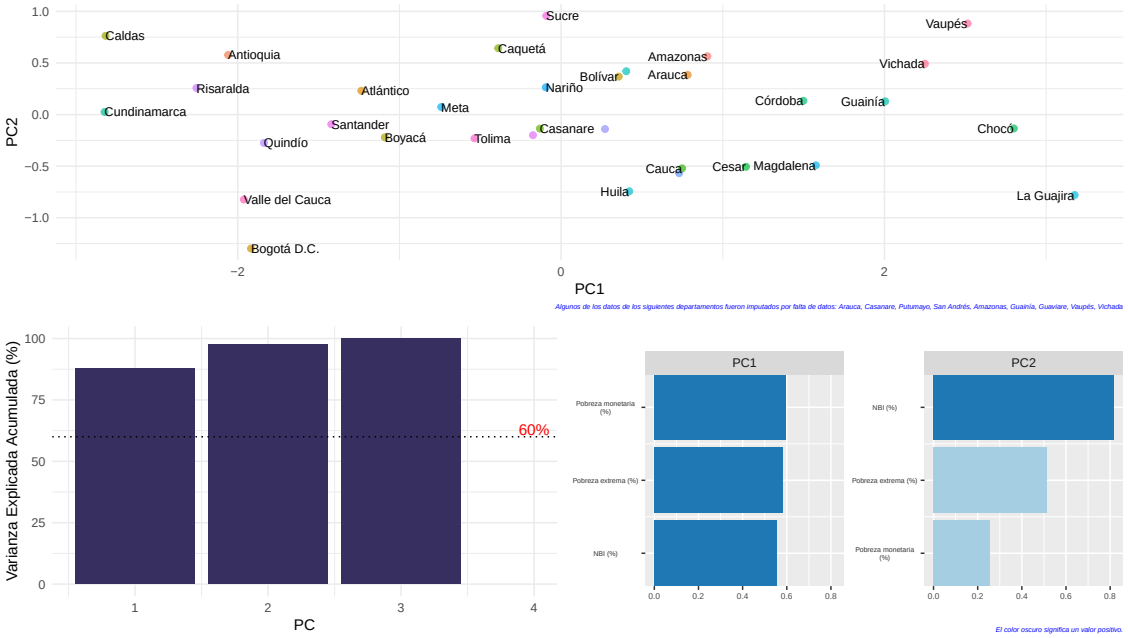

Figura 7 – Resultados de Análisis de Componente Principal - Pobreza

3.1.6 Características de las Viviendas

Esta dimensión está compuesta por variables relacionadas con el material de las paredes de los hogares, material del piso y el porcentaje de personas que viven en condición de hacinamiento crítico. Cada una de las variables fueron centradas y se realizó el análisis de Componentes principales. Los siguientes gráficos muestran las correlaciones de cada uno de las variables con los componentes (dimensiones), resultado del análisis de ACP:

Cuadro 6 – Estadísticas descriptivas

Statistic	N	Mean	St. Dev.	Min	Max
Material de la pared de calidad alta	33	75.859	21.966	14.317	99.022
Material de la pared de calidad baja	33	4.648	6.480	0.166	31.610
Material de la pared de calidad media	33	19.421	19.145	0.322	69.622
Material del piso de calidad alta	33	50.471	23.602	5.354	91.332
Material del piso de calidad media-baja	33	49.529	23.602	8.668	94.646
Hacinamiento crítico (					

Las principales estadísticas descriptivas indican que, en promedio, el porcentaje de hogares que cuentan con material de paredes en ladrillo, bahareque no revocado, bahareque revocado y madera es 74 %, 2.5 %, 2.7 % y 15.5 % respectivamente. Mientras que el

porcentaje de hogares que tienen como material del piso madera, marmol y baldosa es 0.9 %, 0.23 % y 46.6 % respectivamente. Por último, el porcentaje de personas que viven en condición de hacinamiento crítico es 10.4 %. En el primer componente, el hacinamiento crítico tiene un peso negativo sobre el índice y el material de las paredes en ladrillo, así como el material del piso en baldosa o marmol tienen peso positivo sobre el índice. Cabe destacar que el primer componente no logra explicar ni en un 50 % la varianza del índice, pero el acumulado entre las dos sí lo hace.

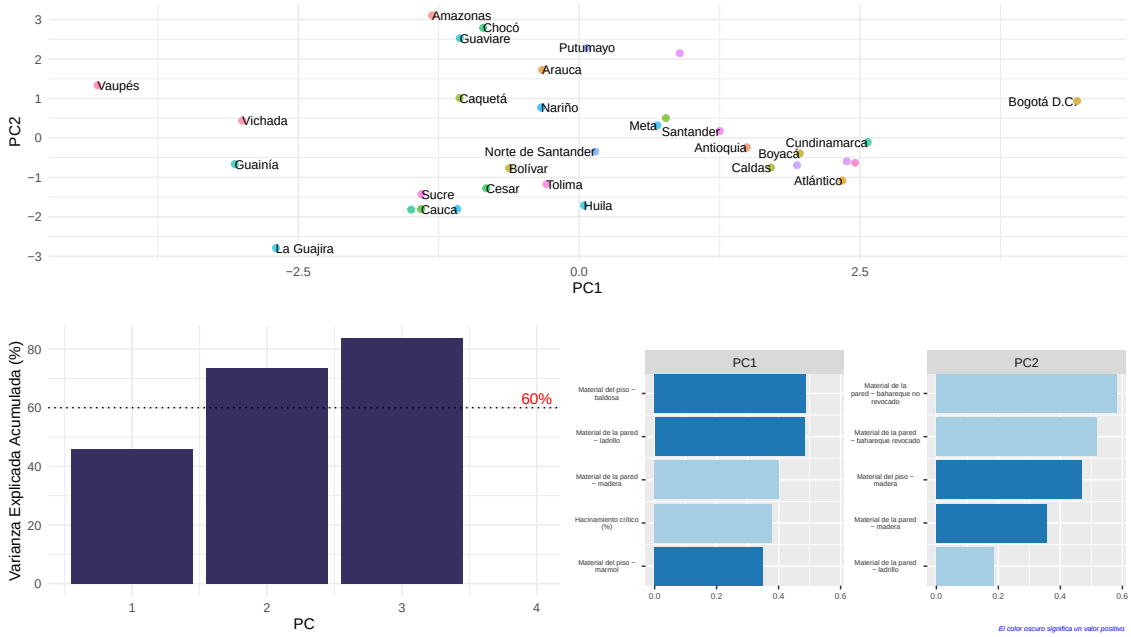

Figura 8 – Resultados de Análisis de Componente Principal - Vivienda

3.1.7 Desigualdad

Esta dimensión está compuesta por 1 variable que corresponde al Coeficiente de GINI de desigualdad. En promedio por departamento, este coeficiente es de 0.506 y el siguiente gráfico muestra la distribución del GINI por el territorio nacional y por cuartiles, donde los departamentos del último cuartil, es decir, con mayores niveles de desigualdad, son La Guajira, Cesar, Antioquia, Chocó y Huila.

Cuadro 7 – Estadísticas descriptivas

Statistic	N	Mean	St. Dev.	Min	Max
Indice de Gini	24	0.485	0.031	0.420	0.554

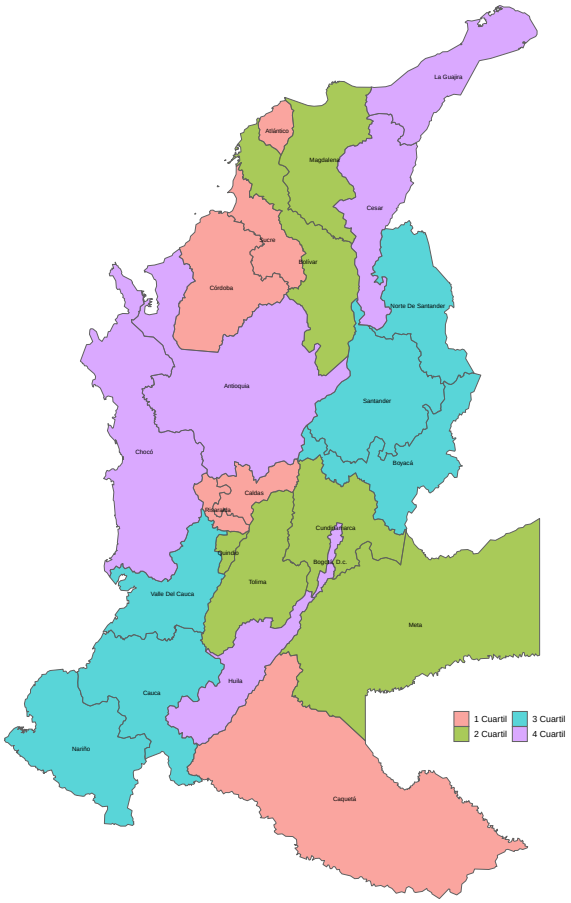

Figura 9 – Resultados de Análisis de Componente Principal - Desigualdad

3.1.8 Estructura demográfica

Esta dimensión está compuesta por 4 variables relacionadas con la población de los departamentos por grupo etario: porcentaje de personas adultas, jóvenes, infantes y ancianas. Cada una de las variables fueron centradas y se realizó el análisis de Componentes principales. Los siguientes gráficos muestran las correlaciones de cada uno de las variables con los componentes (dimensiones), resultado del análisis de ACP:

Cuadro 8 – Estadísticas descriptivas

Statistic	N	Mean	St. Dev.	Min	Max
Población por grupo etario - Adultez	33	42.501	5.096	26.484	50.805
Población por grupo etario - Infancia	33	23.161	4.822	16.171	36.467
Población por grupo etario - Juventud	33	24.589	3.178	19.866	33.109
Población por grupo etario - Vejez	33	9.749	3.242	3.940	16.029

Las principales estadísticas descriptivas indican que, en promedio, la estructura demo-

gráfica de los departamentos se clasifica así: 42 % personas adultas, 23 % infantes, 24 % jóvenes y 9.7 % ancianos. En el primer componente, la población de infantes y jóvenes tiene un peso negativo sobre el índice, mientras que la población anciana y adulta tiene un peso positivo. Por otro lado, en el segundo componente la población anciana y los infantes tienen un peso positivo sobre el índice y la población adulta y joven tienen un peso negativo. Se destaca además que la primera dimensión explica en casi 100 % la varianza del índice.

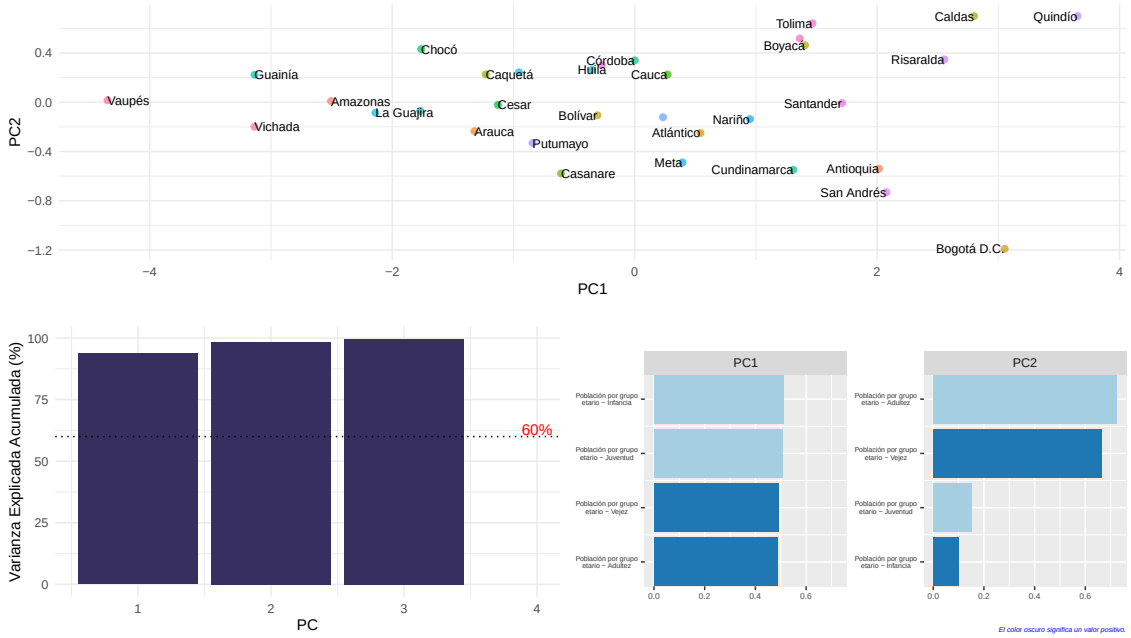

Figura 10 – Resultados de Análisis de Componente Principal - Estructura Demográfica

3.1.9 Ingreso Familiar

Esta dimensión está compuesta por 1 variable principal relacionada con el ingreso de las personas ocupadas, pero a su vez, se descompone por percentiles y se clasifica entre minorías y no minorías. Cada una de las variables fueron centradas y se realizó el análisis de Componentes principales. Los siguientes gráficos muestran las correlaciones de cada uno de las variables con los componentes (dimensiones), resultado del análisis de ACP:

Cuadro 9 – Estadísticas descriptivas

Statistic	N	Mean	St. Dev.	Min	Max
Percentil 25 del ingreso laboral (Millones de pesos)	9	2,437,192.000	714,892.000	1,529,167	3,435,556
Percentil 50 del ingreso laboral (Millones de pesos)	9	3,526,189.000	397,064.700	2,766,877	3,911,659
Percentil 75 del ingreso laboral (Millones de pesos)	9	4,962,833.000	758,660.600	3,911,965	6,125,671

En general, el ingreso laboral promedio por departamento corresponde a un aproximado de 2, 3 y 5 millones COP, para los percentiles 25, 50 y 75 en el caso de las minorías y,

es igual a 1.5,2 y 3 millones COP, para los percentiles 25, 50 y 75 en el caso de las no minorías. Todos los grupos de ingresos tienen un peso positivo sobre el índice para el primer componente, al igual que para el segundo, con excepción del percentil 25 y 75 de las minorias. Se destaca además que la primera dimensión explica en casi 75 % la varianza del índice.

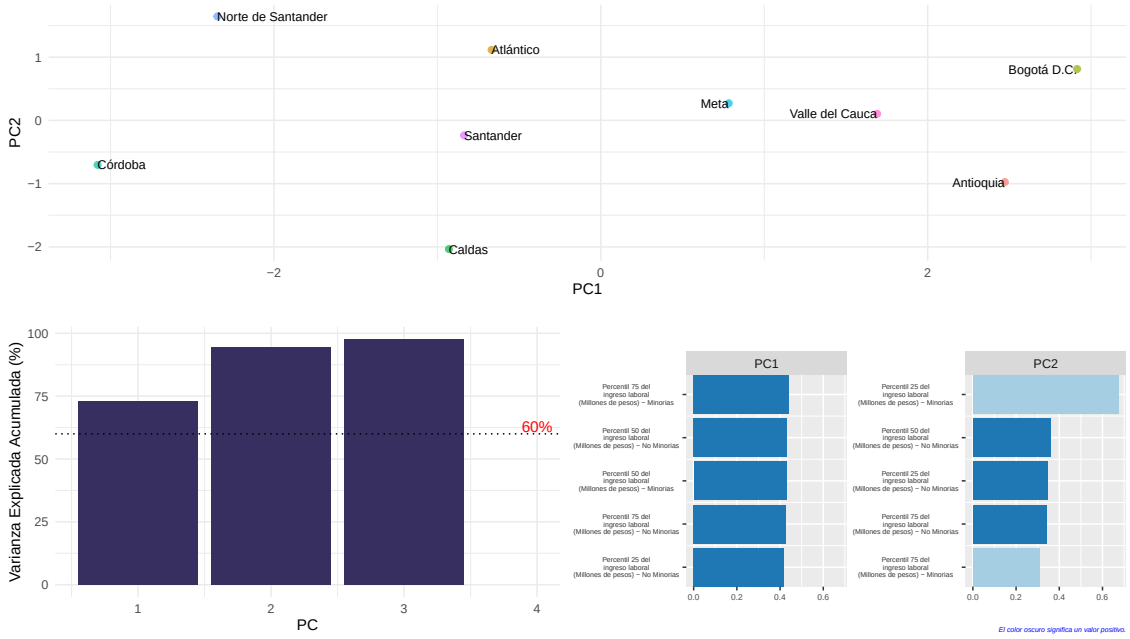

Figura 11 – Resultados de Análisis de Componente Principal - Ingreso Familiar

3.2 Indicadores de Resultados

3.2.1 Infancia y Niñez

Esta dimensión está compuesta por 12 variables principales relacionadas con la infancia y niñez, en términos de tamaño de clase en primaria, asistencia escolar, tasa de mortalidad infantil y nacidos vivos. Cada una de las variables fueron centradas y se realizó el análisis de Componentes principales. Los siguientes gráficos muestran las correlaciones de cada uno de las variables con los componentes (dimensiones), resultado del análisis de ACP:

Cuadro 10 – Estadísticas descriptivas

Statistic	N	Mean	St. Dev.	Min	Max
Tamaño clase en primaria	33	19.170	4.346	10.678	28.111
Asistencia escolar de menores de 5 años (					

Las estadísticas principales indican que, en promedio por departamento, el número de estudiantes por docente en básica primaria es de 19, mientras que el porcentaje de personas menores de 5 años que asisten al colegio es de 19.5 % y 92 % para las personas de 7 a 12 años que asisten a la primaria. Por otro lado, el porcentaje de niños nacidos vivos en 5 años en establecimientos de salud es de 93 % y la tasa de mortalidad infantil es 11 %. En el primer componente, solo la asistencia escolar y el porcentaje de nacidos vivos tienen un peso positivo sobre el índice y en el segundo componente todas tienen peso positivo. Se destaca además que la primera dimensión no alcanza a explicar 50 % de la varianza del índice y el acumulado de las dos dimensiones solo logra explicar aproximadamente el 60 % de esta varianza.

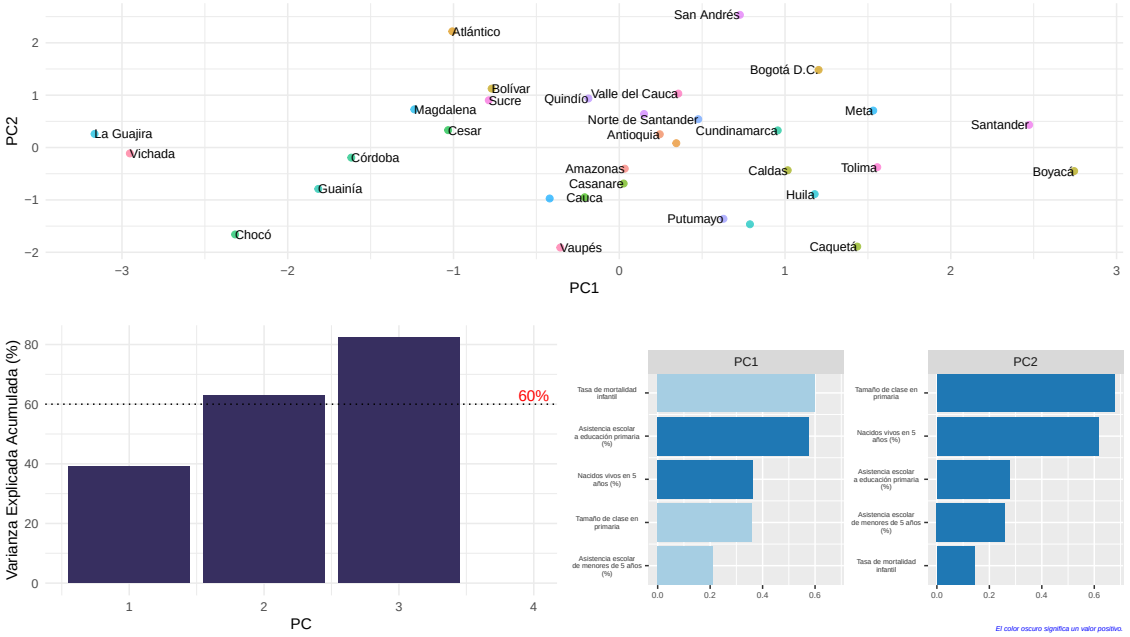

Figura 12 – Resultados de Análisis de Componente Principal - Infancia

3.2.2 Juventud

Esta dimensión está compuesta por 20 variables principales relacionadas con los jóvenes, en términos de tamaño de clase en secundaria y media, asistencia escolar de 13-17 años, diferencias en la terminación de educación secundaria entre indígenas y no indígenas, jóvenes de 18 a 28 años que han asistido a una institución de educación superior, edad promedio del primer matrimonio, mujeres con educación secundaria o superior y tasa de fecundidad de mujeres jóvenes. Cada una de las variables fueron centradas y se realizó el análisis de Componentes principales. Los siguientes gráficos muestran las correlaciones de cada uno de las variables con los componentes (dimensiones), resultado del análisis de ACP:

Cuadro 11 – Estadísticas descriptivas

Statistic	N	Mean	St. Dev.	Min	Max
Tamaño clase en secundaria y media	33	20.103	2.710	13.946	24.074
Asistencia escolar de 13 a 17 años (					

Las estadísticas principales indican que, en promedio por departamento, el número de estudiantes por docente en secundaria y media es de 19, mientras que el porcentaje de personas entre 13 y 17 años que asisten al colegio es de 70.4 % y 22.7 % de los jóvenes entre 18 y 28 años han asistido a una institución de educación superior. Por otro lado, la diferencia en la terminación de educación secundaria entre indígenas y no indígenas es de 21.5 %. Otros indicadores muestran que la edad promedio del primer matrimonio de

los jóvenes es 20 años y el porcentaje de mujeres con educación secundaria o superior es de 78 %, mientras que la tasa de fecundidad de mujeres jóvenes es de 63 %. Solo esta última variable tiene un peso negativo sobre el índice y se destaca además que la primera dimensión no alcanza a explicar 50 % de la varianza del índice y el acumulado de las dos dimensiones solo logra explicar aproximadamente el 60 % de esta varianza.

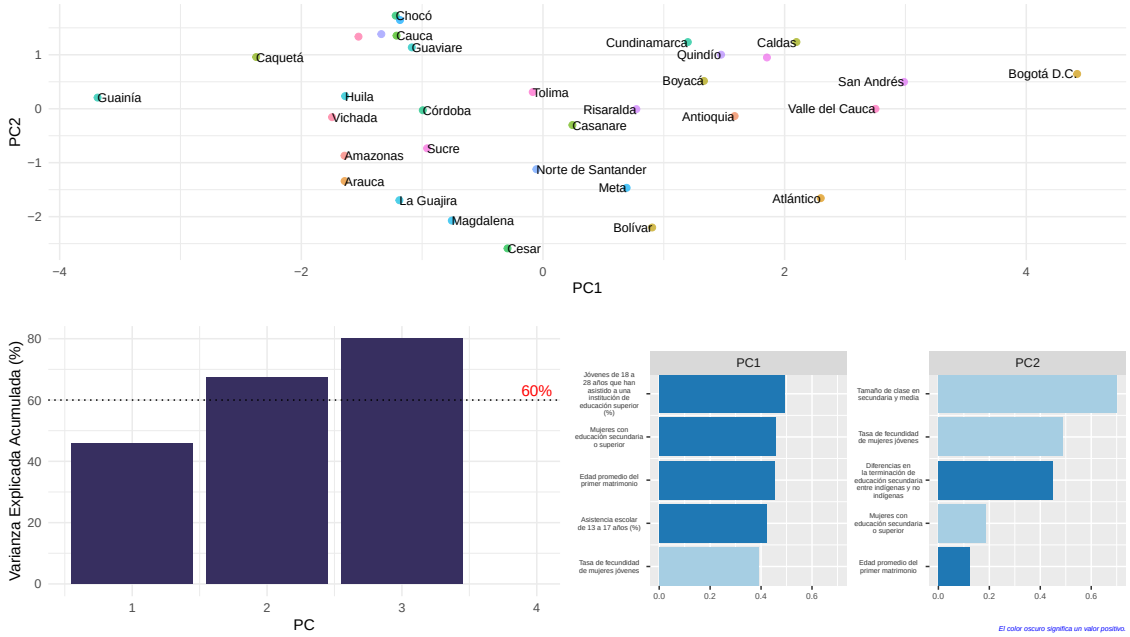

Figura 13 – Resultados de Análisis de Componente Principal - Juventud

3.2.3 Adultez

Esta dimensión está compuesta por 16 variables principales relacionadas con la población adulta, en términos de diferencias en los años de educación promedio entre indígenas y no indígenas, tasa de desempleo, lesiones autoinfligidas intencionalmente (suicidios) y agresiones (homicidios). Cada una de las variables fueron centradas y se realizó el análisis de Componentes principales. Los siguientes gráficos muestran las correlaciones de cada uno de las variables con los componentes (dimensiones), resultado del análisis de ACP:

Cuadro 12 – Estadísticas descriptivas

Statistic	N	Mean	St. Dev.	Min	Max
Diferencias en los años de educación promedio entre indígenas y no indígenas	31	-0.620	1.421	-3.434	2.336
Afiliados al régimen contributivo	33	746,173.800	1,350,349.000	4,992	6,541,090
Afiliados al régimen subsidado	33	726,320.500	587,901.600	15,861	2,446,172
Tasa de desempleo (					

Las estadísticas principales indican que, en promedio por departamento, la diferencia en los años de educación promedio entre indígenas y no indígenas es de 1 %, la tasa

de desempleo en hombres es de 11.9 % y 20.3 % para mujeres. La tasa de suicidios es de 4.8 % para hombres y 1.4 % para mujeres y la tasa de homicidios es de 21.2 % para los hombres y 1.7 % para mujeres. Todas las variables tienen un peso positivo sobre el índice en el primer componente y solo la tasa de homicidios, para hombres y mujeres, tiene un peso negativo sobre el índice en el segundo componente. Por último, la primera dimensión no alcanza a explicar el 40 % de la varianza del índice y el acumulado de las dos dimensiones no logra explicar aproximadamente ni el 60 % de esta varianza.

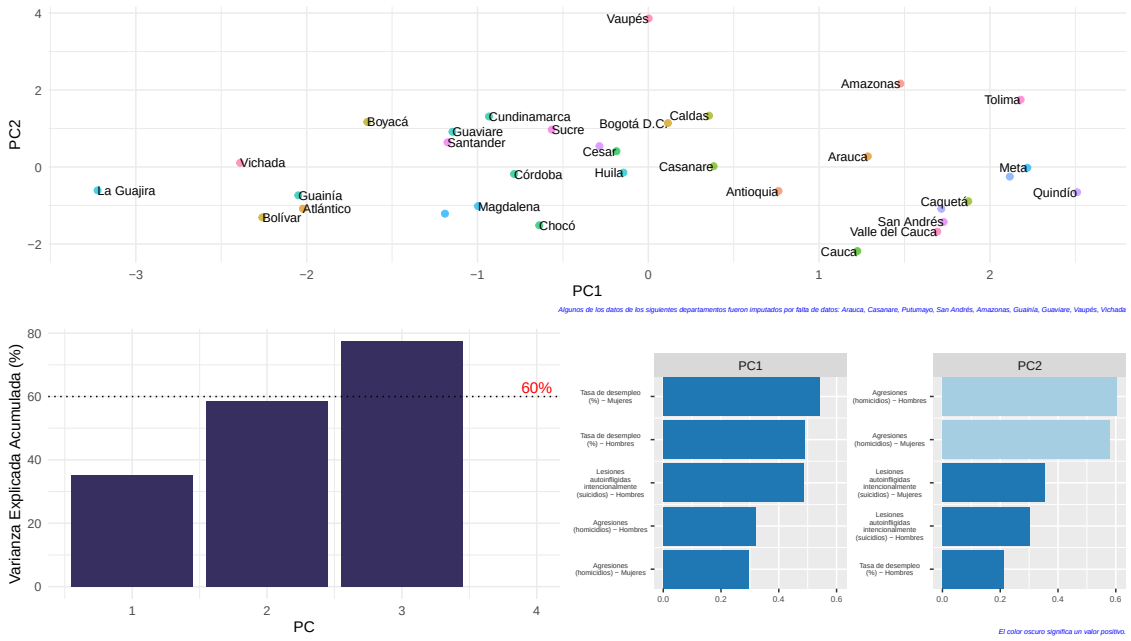

Figura 14 – Resultados de Análisis de Componente Principal - Adultez

3.2.4 Vejez

Esta dimensión está compuesta por 6 variables principales relacionadas con la población anciana, en términos de defunciones por Enfermedades no Transmisibles (ENT) como cáncer, cardiovasculares, diabetes, respiratorias y otras y enfermedades prematuras. Cada una de las variables fueron centradas y se realizó el análisis de Componentes principales. Los siguientes gráficos muestran las correlaciones de cada uno de las variables con los componentes (dimensiones), resultado del análisis de ACP:

Cuadro 13 – Estadísticas descriptivas

Statistic	N	Mean	St. Dev.	Min	Max
Tasa de pensionados	14	28.950	8.202	16.500	39.500
Razón H/M con pensión	14	1.451	0.135	1.240	1.670
Defunciones por ENT - Diabetes	33	1,508.242	2,038.054	10	8,375
Defunciones por ENT - Cancer	33	2,771.879	3,227.062	21	12,404
Defunciones por ENT - Cardiovascular	33	1,351.061	1,589.016	18	6,551
Defunciones por ENT - Otros	33	599.848	762.286	4	3,461
Defunciones por ENT - Respiratorias	33	262.697	272.641	2	1,030
Defunciones por ENT - Diabetes	14	25.093	7.133	14.700	33.800
Tasa de pensionados - Hombres	14	36.379	10.695	19.300	50.700

Las estadísticas principales indican que, en promedio por departamento, el número de defunciones por enfermedades no transmisibles (muertes por cada 100,000 habitantes) como cáncer, enfermedades cardiovasculares, diabetes, respiratorias y otras es 78.4, 139.3, 18.6, 56.6, 26.3, respectivamente. Mientras que el número de defunciones prematuras por cada 100,000 habitantes es de 187.6. Todas las variables tienen un peso positivo sobre el índice en el primer componente y las defunciones por otras ENT y por diabetes tienen un peso negativo sobre el índice en el segundo componente. Por último, el primer componente no alcanza a explicar el 60 % de la varianza del índice, pero el acumulado de las dos componentes sí e incluso es explica casi el 75 %.

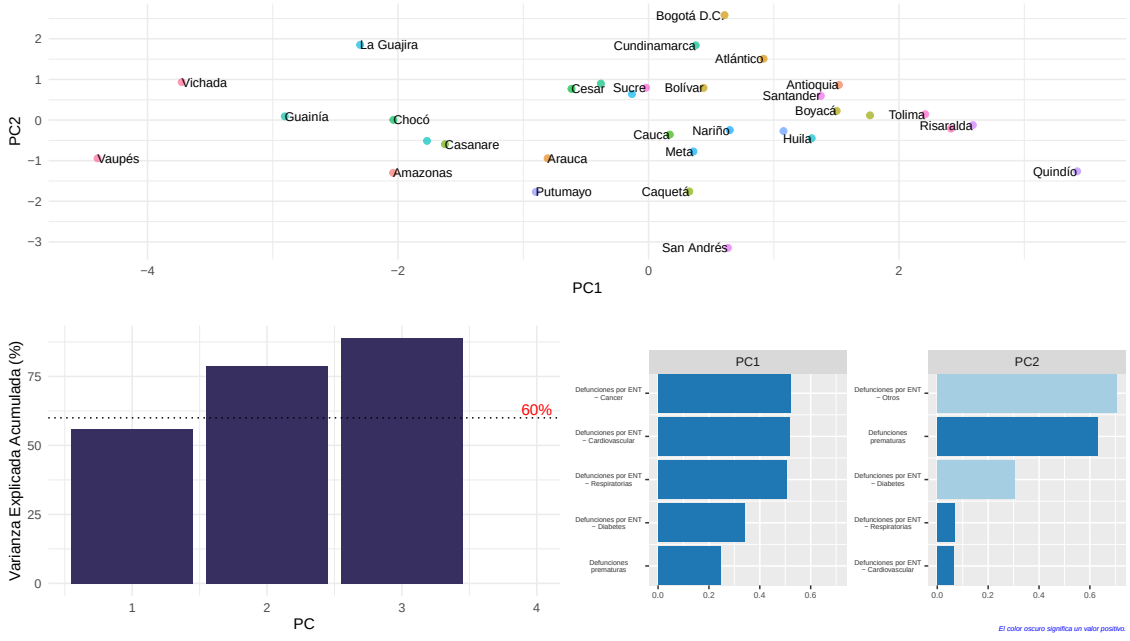

Figura 15 – Resultados de Análisis de Componente Principal - Vejez

4 Construyendo las múltiples Colombia: resultados y validación del análisis de clústers

Una vez construido una aproximación simple para cada una de las sub-dimensiones se realizó el último paso del análisis: identificar las realidades territoriales a partir del análisis sistemático de la información disponible. En este sentido, se realizó un análisis de aprendizaje de máquina no supervisado para dividir las observaciones en grupos que cumplan las siguientes principios(?, ?):

- **Principio de cohesión**: Se busca que cada grupo tenga una alta homogeneidad al interior en las variables utilizada. Es decir, en términos del presente proyecto, se busca que las regiones o departamentos sean altamente parecidas en las 13 dimensiones que componen la metodología *Pulso Social*.
- **Principio de separación**: Es necesario garantizar que la diferencia entre los grupos sea claramente definida. Es decir, que cada uno de los grupos tengan condiciones diferentes entre ellos.

De esta manera, el sentido de la aplicación de la metodología de conglemerados en el proyecto de *Pulso Social* es poder brindar una lectura de las brechas que existen en el territorio colombiano a partir de una análisis sistemático de las dimensiones que componen la metodología. A continuación, se presentan las estadísticas descriptivas de los indicadores por cada una de las dimensiones incluidas en el análisis, resultado del análisis de los componentes principales:

Cuadro 14 – Estadísticas descriptivas de los índices empleados para el análisis de conglomerados

El primer paso para el análisis de conglomerados, es la estimación de la *matrix de distancia o disimilitud* la cual estable una medida de (dis)similitud entre cada par de observaciones. Existen varios métodos para estimar la distancia, dado los resultados similares entre los diferentes métodos se eligió el método estándar a partir de la estimación con distancias euclidianas[1]:

[1] Entre los métodos alternativos para la estimación de la matriz de distancia se destacan: distancia Manhattan, la distancia de correlación pearson, entre otros

$$d_{ecu}(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

La siguiente figura presenta los resultados del análisis de la aplicación de la matriz. Los resultados sugieren la presencia de varios grupos de departamentos con condiciones similares. En primer lugar, regiones como Antioquia, Valle del Cauca, Risaralda, Caldas y Quindio tienen una alta similitud. En contraste, en segundo lugar, otros departamentos como Vaupés, la Guajira, entre otros son diferentes a los iniciales pero similares entre ellos. Esta es la primera aproximación a los posibles cluster que surgirán del análisis.

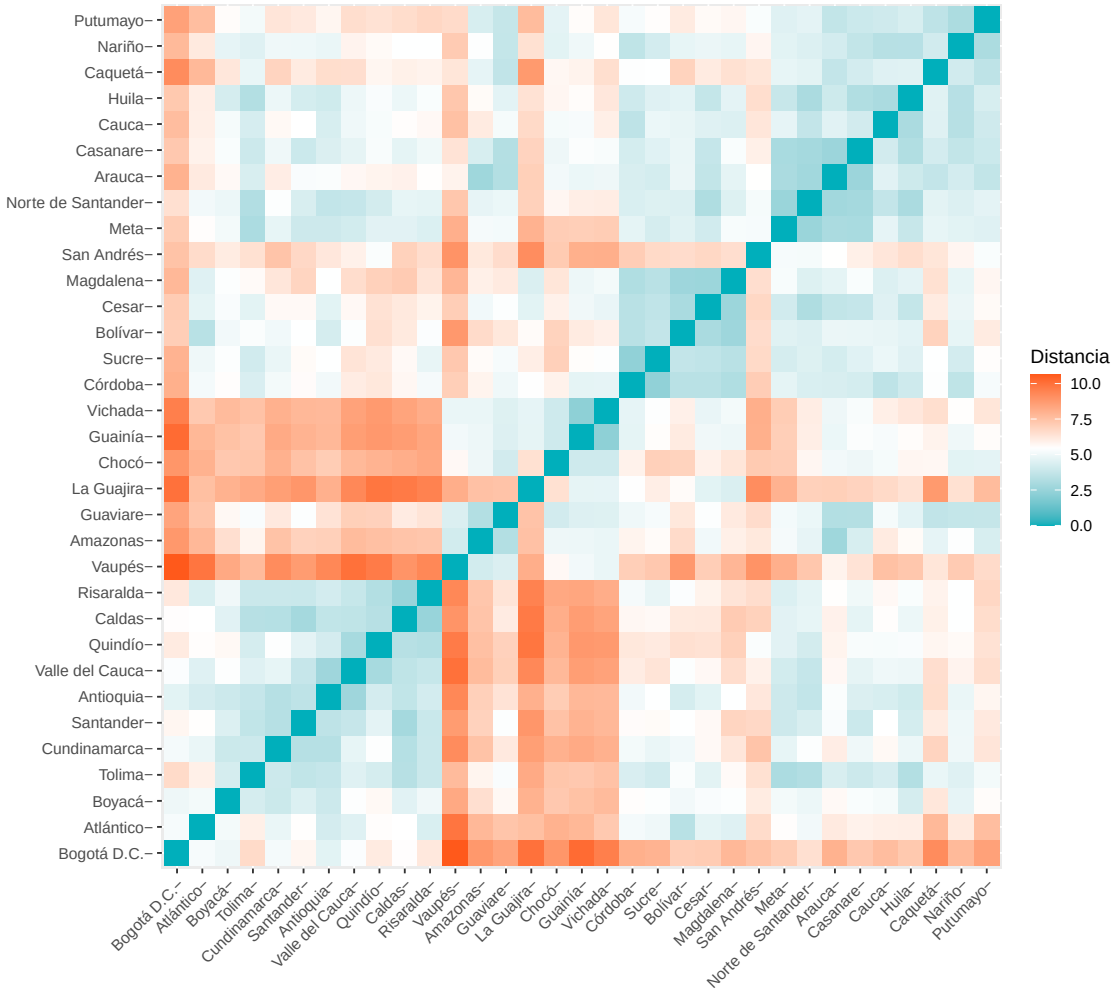

Figura 16 – Matriz de distancia euclideana

El método de cluster utilizado es el método de cluster jerárquico aglomerativo. Este método también conocido como AGNES (Agglomerative Nesting) funciona de la siguiente forma: cada objeto se considera inicialmente como un clúster de un solo elemento, luego de manera iterada se va combinando en una cluster mayor. Este proceso se repite hasta que todos las observaciones pertenezcan a una sola raíz(?, ?). Esta lógica permite una representación visual conocido como *dendograma*.

average	single	complete	ward
0.51	0.37	0.72	0.83

Cuadro 15 – Coeficiente de Aglomeración por diferentes métodos de enlace

4.0.1 Seleccionando el método de enlace

Dado que existen diferentes métodos de disimilitud para .enlazar"los diferentes cluster, se inicio un análisis de comparación a partir del *coeficiente de aglomeración* el cual indica el poder de explicación de la estructura de agrupación que se ha obtenido mediante el enlace establecido. Entre más cercano a 1 se tiene una mejor agrupación(?, ?). La siguiente tabla presenta los resultados, los cuales permiten concluir que el enlace *ward* es más apropiado.

4.0.2 Determinando el número de clústers

Después de aplicar el método de enlace más adecuado, es necesario determinar el número de clúster o grupos óptimo. Aunque no existe un criterio único y concluyente sobre su determinación, el método más utilizado es el método del codo, *(elbow method)*, el cual determina el número óptimo de cluster a partir de encontrar el número que minimiza la variación total:

$$argmin\left(\sum_{k=1}^{k} W(C_k)\right)$$

Donde C_k es el k-esimo cluster y $W(C_k)$ es la variación intra del cluster. La suma cuadrada total dentro del clúster (wss) mide la similitud de la agrupación, la cual se desea que sea lo más pequeña posible. Los resultados se presentan en la siguiente gráfica.

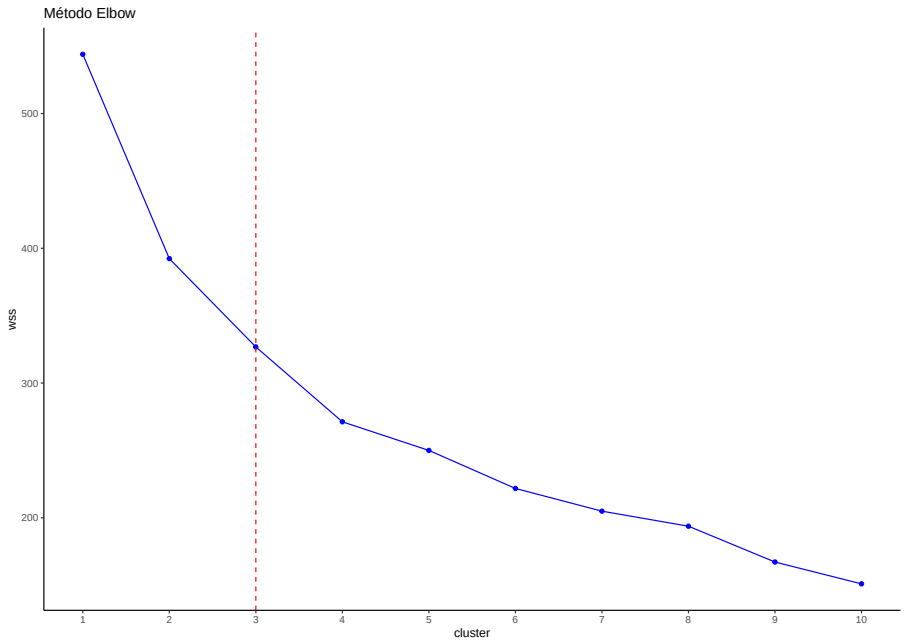

Figura 17 – Método del codo para estimar el número óptimo de clusters

Después de una discusión técnica y con base a los resultados encontrados se concluye que el número óptimo de clusters es *tres (3)*.

4.0.3 Dendograma y representación de los clusters

Una vez definido el método de enlace el número óptimo de clusters se procede con la última parte del análisis que es presentación gráfica de los resultados.

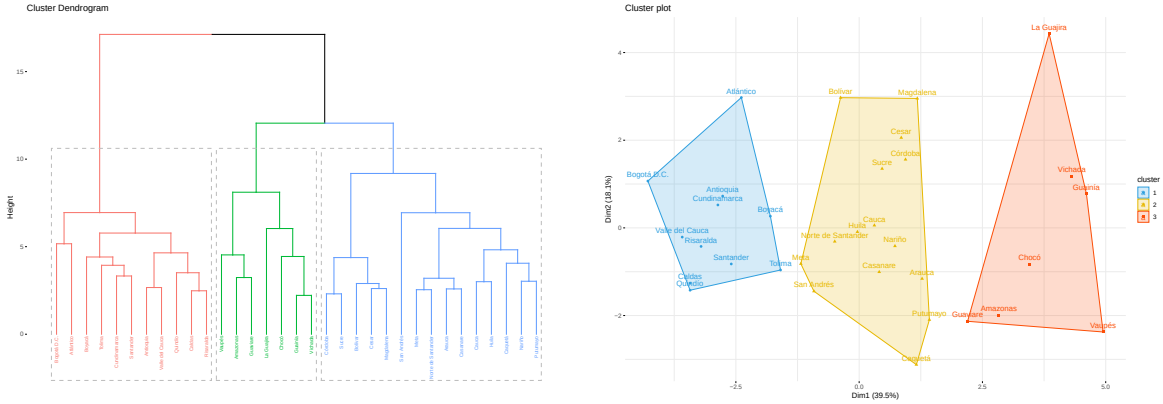

Figura 18 – Representación gráfica de los resultados cluster de *Pulso Colombia*

La composición de los cluster es la siguiente:

- **Cluster 1 - Colombia del Futuro:** Bogotá D.C.,Boyacá, Caldas, Cundinamarca,Antioquia,Quindío,Risaralda, Santander, Tolima, Valle del Cauca y Atlántico
- **Cluster 2 - La Colombia en transición:** Bolívar,Caquetá,Cauca, Cesar, Córdoba,Huila, Magdalena,Meta, Nariño, Norte de Santander Sucre, Arauca, Casanare,Putumayo, San Andrés
- **Cluster 2 - La Colombia rezagada:** Chocó, La Guajira, Amazonas, Guainía, Guaviare,Vaupés,Vichada

Dados sus características y para facilitar el análisis, los grupos fueron nombrados. La representación espacial de la composición de los clusters se define a continuación.

Figura 19 – Represetación Filogénica de la agrupación por cluster de *Pulso Colombia*

El análisis detallado se encuentra en el informe

4.0.4 Contribución de las variables a la agrupación

Finalmente, se presenta la composición de las variables por cada tipo de cluster. En análisis se realiza al interior del documento.

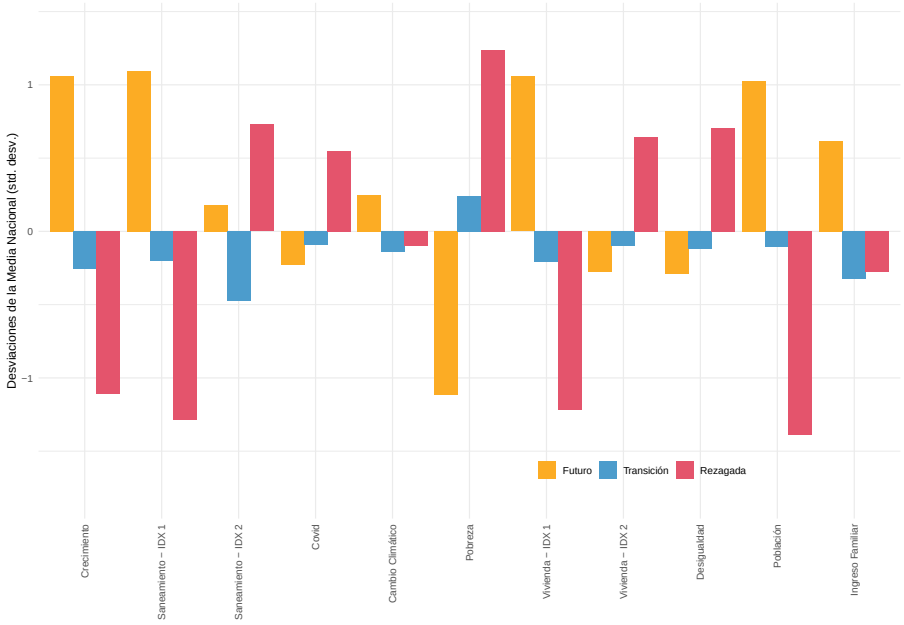

Figura 20 – Medias de los indicadores de contexto por tipo de cluster

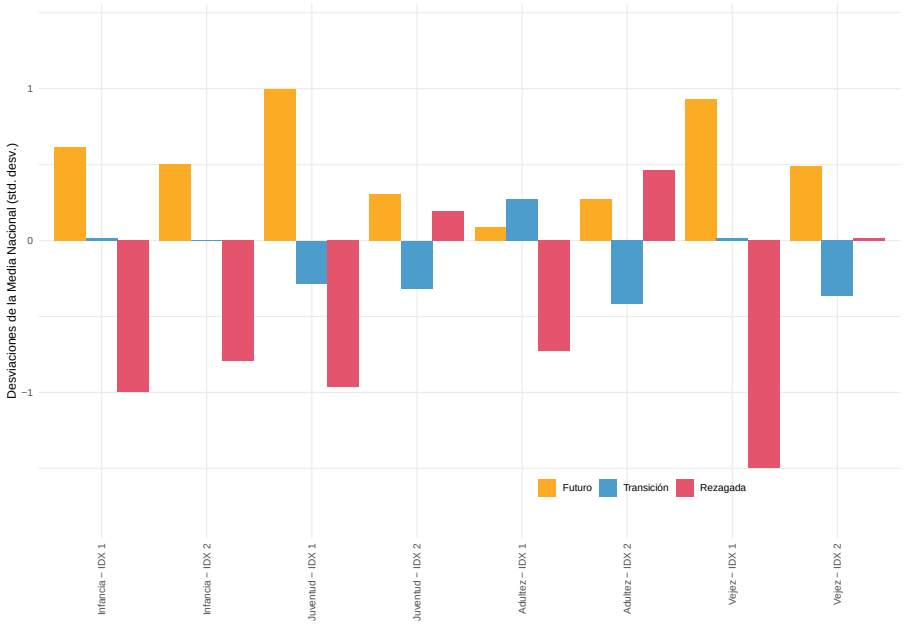

Figura 21 – Medias de los indicadores de resultados por tipo de cluster

Referências bibliográficas