DANE
INFORMACIÓN PARA TODOS

DIRECCIÓN
DE CENSOS Y DEMOGRAFIA

ESTIMACIÓN DE POBLACIÓN ÉTNICA A NIVEL SUBNACIONAL

METODOS Y RESULTADOS

Junio de 2021

El futuro es de todos
Gobierno de Colombia

CONTENIDO

Lista de tablas

Lista de gráficos

Lista de ilustraciones

1. Introducción

La creciente capacidad de almacenamiento y procesamiento de información abren la posibilidad de incrementar el conocimiento sociodemográfico, permitiendo caracterizar de manera precisa a las diferentes poblaciones, y, particularmente, a los grupos étnicos captados a través de la pregunta de autorreconocimiento, por medio de la cual se indaga a las personas si pertenecen a uno de los grupos étnicos reconocidos constitucionalmente en Colombia. Esta pregunta fue incluida, tanto en el Censo de Población y Vivienda 2018, como en diferentes encuestas de hogares desarrolladas por el Departamento Administrativo Nacional de Estadística - DANE. Lo anterior es importante en la medida en que es posible conocer y determinar las brechas que enfrentan los diferentes grupos étnicos en Colombia, en ámbitos como: salud, educación, acceso a servicios públicos y privados, mercado laboral, condiciones de pobreza, participación política, entre otros. Sin embargo, conocer estas brechas implica diversos retos, dada la existencia de sesgos en la recolección de información poblacional, específicamente dentro de los grupos étnicos, debido a que la respuesta a la pregunta que mide el autorreconocimiento étnico puede estar influenciada por diferentes factores operativos, culturales, contextuales, entre otros, que dificultan su recolección en los diferentes niveles geográficos.

Un ejemplo de lo antes planteado se evidencia en los resultados del Censo Nacional de Población y Vivienda 2018 (CNPV 2018), en el que el número de personas efectivamente censadas pasó de 41.174.853 en el 2005, a 44.164.417 en 2018; con resultados heterogéneos entre la población según autorreconocimiento étnico. Por ejemplo, la población que se auto reconoció como indígena presentó un incremento del 36,8% durante el periodo intercensal, pasando de 1.392.623 personas en 2005 a 1.905.617 en 2018; resultado explicado no solo por las mayores tasas de fecundidad de las mujeres pertenecientes a este grupo étnico, sino también a un aumento en la cobertura de esta población.

En lo que respecta a la población Negra, Afrocolombiana, Raizal y Palenquera - NARP, el CNPV 2018 contó a 2.982.224 personas que se autorreconocieron como pertenecientes a dicho grupo étnico, en contraste con el resultado del Censo General de 2005, donde a través de la misma pregunta se autorreconocieron 4.311.757 personas, con una disminución del 30,8%, lo cual no es consistente con la dinámica demográfica de la población que se había identificado en el censo anterior. Dicha disminución se explica por factores multicausales relacionados con:

- Incumplimiento de la visita en la vivienda o no contar con una entrevista satisfactoria, ya sea por falta de comunicación para la llegada del censista, ausencia del informante idóneo en el momento de la visita y/o negativa para brindar información al DANE.

- En ocasiones los censistas realizaron hasta 4 visitas a los hogares y no fue posible encontrar un informante idóneo.
- Dificultad para cubrir algunas zonas urbanas y rurales por problemas de orden público o por la negativa de los residentes a ser censados: Barranquilla, Cali, Quibdó, Policarpa y Tumaco (Nariño).
- Situaciones relacionadas con la ejecución del Censo: capacitación, contratación de personal Afro, logística.
- Las personas no se autorreconocieron en las categorías de respuesta 3. (¿Raizal del Archipiélago de San Andrés, Providencia y Santa Catalina?), 4. (Palenquero(a) de San Basilio?) y 5. (Negro(a), mulato(a), afrodescendiente, afrocolombiano(a)?).
- Los censistas no formularon la pregunta y marcaron la alternativa 6 (Ningún grupo étnico).
- Negativa de las personas a responder la pregunta (zonas urbanas).
- El autorreconocimiento es un proceso subjetivo relacionado con la formación de identidad, con procesos sociales, históricos, construcciones políticas, conceptualizaciones académicas y personales.
- El discurso de la costeñidad y el mito del mestizaje triétnico.
- La existencia del racismo y la discriminación racial.
- La negación de la etnicidad de la gente negra, afrocolombiana, raizal y palenquera en los departamentos en donde existen pueblos indígenas.
- La debilidad de las organizaciones étnicas en algunas zonas del país.

Debido a los problemas identificados en el CNPV 2018, y con el propósito de generar insumos de información adecuados para la toma de decisiones públicas en los lugares donde se ubican los grupos étnicos, el DANE diseñó estrategias para estimar, de forma indirecta, el volumen poblacional de los grupos étnicos de Colombia a nivel subnacional. Esta tarea representa uno de los principales desafíos de la etapa postcensal.

Antecedentes

Con el propósito de ampliar el contexto, a continuación, se exponen los antecedentes y características de los operativos censales y los datos obtenidos a partir del Censo Nacional de Población y Vivienda 2018 -CNPV 2018- y la Encuesta de Calidad de Vida 2018 -ECV 2018-.

CNPV 2018:

Como punto de partida para el desarrollo del CNPV 2018, se estableció un trabajo con usuarios internos y externos, el cual se realizó entre los años 2012 y 2015, periodo en el cual se definieron los aspectos relacionados con los requerimientos de información que abordaría el censo.

El CNPV 2018 cubrió todo el territorio nacional (continental e insular), obteniendo información en los diferentes niveles de división político administrativa, es decir, total nacional, departamental, municipal, cabecera y área resto municipal, incluyendo centros poblados y rural disperso. Su universo de estudio, estuvo compuesto por todas las viviendas, hogares y personas residentes habituales en el territorio colombiano, que en su área continental e insular tiene una superficie de 1.141.748 kilómetros cuadrados y se encuentra conformado por 32 departamentos y el Distrito Capital. El territorio cuenta con un total de 1.101 municipios, 20 áreas no municipalizadas y el Archipiélago de San Andrés, Providencia y Santa Catalina.

También incluyó la población residente habitual en embajadas y consulados donde Colombia actúa de conformidad con el Derecho Internacional o con las leyes colombianas. Igualmente, las personas residentes habituales en los Lugares Especiales de Alojamiento (LEA).

Las unidades estadísticas de observación estuvieron constituidas por las viviendas, los hogares, y los Lugares Especiales de Alojamiento; y las unidades de análisis por las viviendas, los hogares y las personas.

El CNPV 2018 fue un Censo de Derecho o de "Jure que implica que las personas se censan en el lugar de su residencia habitual". El Residente habitual es la persona que habita la mayor parte del tiempo en una unidad de vivienda o en un lugar especial de alojamiento, aunque en el momento de la entrevista se encuentre ausente temporalmente". Además de los residentes presentes, se consideraron como residentes habituales de un hogar las siguientes personas:

- Los miembros del hogar que en el momento de la entrevista se encuentren ausentes temporalmente por un periodo igual o inferior a seis meses, por motivos especiales, como: vacaciones, cursos de capacitación, viajes de negocio, comisiones de trabajo, entre otros.
- Los secuestrados y desaparecidos, sin tener en cuenta el tiempo de ausencia.
- Los enfermos que reciben atención en hospitales o clínicas, sin tener en cuenta el tiempo de ausencia.
- Los desplazados que residen en el hogar, sin tener en cuenta el tiempo que lleven allí.
- Las personas detenidas temporalmente en inspecciones de policía.

- Las personas privadas de su libertad y de otros derechos civiles por haber infringido la ley, y que gozan del beneficio de "casa por cárcel", sin tener en cuenta el estado en el que se encuentre su proceso.
- Los que prestan el servicio militar en la policía y duermen en sus respectivos hogares.
- Los empleados del servicio doméstico "internos", que son aquellos que duermen la mayor parte del tiempo en la vivienda donde trabajan y, por ello, se consideran miembros del hogar para el cual trabajan.
- Los pensionistas, personas que pagan por los servicios de vivienda y alimentación y, por ello, se consideran miembros del hogar al cual le pagan por estos servicios.
- Los residentes en casas fiscales.

La recolección de información del CNPV se realizó a través de internet (e-Censo) y del operativo de campo, mediante los métodos de barrido (vivienda a vivienda), de rutas, focalizado (Lugares Especiales de Alojamiento) y mixto (ruta y barrido). El censo electrónico se realizó entre el 10 de enero y el 12 de abril y el operativo de campo (puerta a puerta) se efectuó entre el 18 de abril y el 30 de octubre de 2018, aunque durante los meses de noviembre y diciembre se realizó un proceso de recuperación de personas especialmente en el departamento del Valle del Cauca, específicamente en los municipios de Cali y Buenaventura.

A continuación, se explican cada uno de los métodos de recolección:

- Electrónico: correspondió al diligenciamiento del cuestionario censal a través de la interacción con una aplicación web, directamente por la fuente primaria.
- Barrido: Recorrido sistemático de la totalidad de las unidades de observación asignadas en un área de trabajo, regresando periódicamente al centro operativo municipal a entregar información y los reportes diarios del personal de censistas en campo.
- Ruta: Método se caracteriza por la permanencia en el área de trabajo del personal de censistas en campo, hasta la finalización del recorrido. La recolección de la información se realizó a través del personal de censistas en campo, sobre un recorrido estructurado a lo largo de un río y sus afluentes o vías con sus caminos y senderos, generando un área de influencia en la cual se encuentra localizada la población.
- Mixto: Se refiere a la recolección de la información a través del personal de censistas en campo, aplicando de manera combinada tanto el método de barrido como de ruta, en municipios caracterizados por condiciones diferenciales de acceso a los territorios donde se asienta la población residente en el área rural dispersa.

Adicionalmente, hay dos métodos que fueron utilizados durante el operativo de manera excepcional debido a las características específicas de algunas poblaciones o con el fin de responder a contingencias presentadas durante el desarrollo del operativo:

- Focalizado: Se refiere a todas aquellas acciones que permiten realizar el levantamiento de la información censal de un grupo específico de población, que por características de localización y alojamiento debe abordarse de manera particular.
- Convocatoria: Estrategia de contingencia de uso excepcional, implementada exclusivamente en áreas que, por limitaciones ajenas al DANE (por razones de seguridad, orden público, etc), el personal operativo no puede acceder a las zonas para ser censadas. Para su implementación se define una ubicación específica reconocida, del área urbana o rural del municipio, donde se dispone el personal operativo necesario y los medios para la colecta censal; en este lugar se concentra la población de la comunidad no censada para brindar la información.

La planeación y el desarrollo del CNPV 2018 contaron con la participación de los grupos étnicos constitucionalmente reconocidos: indígenas, gitanos o Rrom, y comunidades negras, afrocolombianas, raizales y palenqueras. Lo anterior, de acuerdo con lo establecido en el Plan Nacional de Desarrollo 2014- 2018 "Todos por un nuevo país", y en reconocimiento de sus derechos establecidos en el marco normativo nacional e internacional, entre otros, la Constitución Política de 1991, el Convenio 169 de 1989 sobre pueblos indígenas y tribales, la Ley 21 de 1991, la Sentencia T- 576 del 4 de agosto del 2014 y los decretos 1745 de 1995, 1066 de 2015 y 1372 de 2018.

Este proceso inicia en el 2015 con los pueblos indígenas, a través de la construcción de la ruta metodológica de la consulta y concertación, realizando talleres de socialización; e incluyó la concertación del cuestionario censal y de los procesos censales. Esta labor fue adelantada con las siguientes organizaciones indígenas de la MPC: ONIC, OPIAC, AICO, CIT y Gobierno mayor; de forma independiente y autónoma con los pueblos: COFAN, Wayú y Kogüi y el Resguardo Selva Matavén.

La participación de los pueblos indígenas en la definición del enfoque étnico diferencial del cuestionario se dio mediante la implementación de la siguiente metodología:

- Realización de asambleas entre el DANE y representantes de los pueblos indígenas y espacios propios de estos grupos. Durante estas asambleas, técnicos de la DCD explicaron en detalle cada una de las preguntas del cuestionario. Posteriormente, los miembros de las comunidades indígenas, en espacios autónomos, las analizaron y retroalimentaron al DANE. Esta información fue recopilada, a nivel nacional, en la matriz denominada "retroalimentación cuestionario MPC",

la cual se encuentra ubicada en el repositorio interno de almacenamiento de la documentación del censo.
- Revisión conjunta entre el DANE, técnicos delegados por las organizaciones indígenas y expertos nacionales e internacionales de la CEPAL, de las propuestas plasmadas en la matriz generada.
- Acuerdos y protocolización de las variables y categorías del cuestionario censal a ser incluidas, lo cual se dio en el espacio de la Mesa Permanente de Concertación – MPC.

Específicamente, con los pueblos indígenas se acordaron tres preguntas (P.13 de identificación de la vivienda en un territorio étnico; P.36 de pertenencia étnica y P.37 de Lengua nativa) y ocho categorías de respuesta específicas, las cuales están incluidas en las preguntas mencionadas y en otras relacionadas con la tipología de vivienda, con la fuente de obtención del agua para preparar los alimentos y con la atención en salud de los pueblos indígenas.

Para el caso de las comunidades Negras, Afrocolombianas, Raizales y Palenqueras, la ruta de consulta y concertación fue protocolizada en el año 2016 entre el DANE y el Espacio Nacional de Consulta Previa – ENCP de medidas legislativas y administrativas de amplio alcance susceptibles de afectar a dichas comunidades, en la cual se incluyó su participación en las diferentes fases (precensal, censal y poscensal). La ruta metodológica concertada para la consulta incluyó:

- Tres reuniones con el Espacio Nacional de Consulta, reuniones con la Comisión VII, reuniones departamentales, acompañamiento técnico y fortalecimiento organizativo, también, reuniones de seguimiento a los acuerdos de consulta previa. Vale la pena resaltar que la comisión VII, llamada de comunicaciones, TIC's, censos, estadística, innovación, ciencia y tecnología, fue la designada por el Espacio nacional de consulta para adelantar las discusiones previas que se llevaron a las plenarias del ENCP.

- Reuniones con el Espacio Nacional de Consulta Previa en los cuales se abordó el contexto del Censo Nacional de Población y Vivienda, la definición y aprobación de la ruta metodológica de la consulta y concertación con las comunidades negras, afrocolombianas, raizales y palenqueras; y la presentación, revisión, análisis y ajustes del cuestionario censal y la presentación y socialización de los aspectos técnicos y los mecanismos de participación de las comunidades negras, afrocolombianas, raizales y palanqueras en los diferentes procesos censales.

- Reuniones con la comisión VII del Espacio Nacional de Consulta Previa. Con esta instancia se llevaron a cabo 9 reuniones. En los diferentes espacios se abordaron los temas relacionados con

el cuestionario censal, y los procesos censales. Se llevaron a cabo socializaciones, análisis y las discusiones a partir de los cuáles se consolidaron los acuerdos de la consulta. Los espacios se llevaron a cabo de acuerdo con el enfoque diferencial étnico en la normatividad vigente y los principios establecidos para este derecho fundamental de las comunidades. Los integrantes de esta comisión, junto con el DANE, socializaron con los representantes de las comunidades los aportes y avances del enfoque diferencial étnico en el cuestionario censal y los procesos censales, con el fin de informar y aprobar el contenido y el alcance de sus decisiones estuvo sujeto a lo acordado y definido con los integrantes del Espacio Nacional de Consulta Previa.

- Espacios Departamentales. Se realizaron 33 espacios departamentales en el marco del proceso de consulta y concertación con la población negra, afrocolombiana, raizal y palenquera, abarcando la totalidad del país. La duración de estos espacios fue de 1 a 2 días de acuerdo a la complejidad del escenario y asistieron representantes de los consejos comunitarios, las organizaciones de base y demás expresiones organizativas de las comunidades negras, afrocolombianas, raizales y palenqueras, citadas en su autonomía por los delegados de cada departamento, y con el fin de garantizar la interlocución efectiva con las comunidades y su participación en el proceso de consulta. Durante la realización de estos espacios se presentaron las generalidades del CNPV, del proceso de consulta y concertación, al tiempo que se socializaron y retroalimentaron el cuestionario y los procesos censales. Participaron en total aproximadamente 4200 líderes de las comunidades. (Ver listados de asistencia anexos). De acuerdo con los aspectos definidos en la ruta metodológica para el proceso de consulta y concertación con las comunidades Negra, Afrocolombianas, Raizales y Palenqueras, se planearon y desarrollaron treinta y dos (32) Asambleas Departamentales y una (1) distrital.

- Reuniones con equipos técnicos, centradas principalmente en la discusión sobre el cuestionario censal y las fases del CNPV 2018.

Para el caso de la consulta con las comunidades negras, afrocolombianas, raizales y palenqueras, la protocolización del cuestionario censal con enfoque diferencial tuvo como principales logros la pregunta de autorreconocimiento, las preguntas de territorialidad étnica y la incorporación de ajustes al tema de la lengua.

Con el pueblo ROM se realizaron reuniones con la Comisión Nacional de Diálogo. Este proceso concluyó con la protocolización de acuerdos relacionados con el contenido del cuestionario y con los procesos censales.

Los acuerdos sobre el cuestionario censal, se dieron fundamentalmente en lo correspondiente a su adecuación desde un enfoque diferencial, que tiene entre sus principales logros las preguntas de autorreconocimiento, territorialidad étnica y la incorporación de ajustes al tema de la lengua, así como la inclusión de las categorías de respuesta, "Vitsa" y "Kumpania" en la pregunta de pertenencia étnica, solicitadas por el pueblo Rrom.

Durante la etapa censal se implementaron las siguientes estrategias dirigidas a responder a las necesidades de los grupos étnicos de acuerdo con los diferentes métodos de recolección establecidos. En el marco de los acuerdos con los grupos étnicos, se implementaron estrategias de comunicación y sensibilización acordes con su cultura, tradiciones, necesidades comunicativas y ubicación geográfica. Adicionalmente, mediante convenios de asociación se establecieron estrategias locales y regionales concertadas.

- Para los grupos indígenas, y siguiendo los acuerdos de la Mesa Permanente de Concertación - -MPC, se instaló un Consejo Editorial con participación de cinco delegados de las organizaciones indígenas, en el cual se desarrollaron conjuntamente: 30.000 Cartillas, cuñas para emisión nacional, un spot de televisión para emisión por código cívico, 20.700 afiches (español y lenguas nativas) y 54.600 plegables en español.
- Con respecto a la comunidad Negra, Afrocolombiana, Raizal y Palenquera - NARP, y en respuesta a los acuerdos con la Comisión VII, se implementó una única estrategia de comunicación y socialización del CNPV 2018, correspondiente a una cuña radial, un spot de televisión para emisión por código cívico, un afiche en 3 referencias: español, créole y palenquero – 10.400 ejemplares, y un plegable en español – 150 mil ejemplares.
- Con la Comisión de diálogo para el pueblo Rrom (Gitano), en respuesta a los acuerdos, se produjo un afiche en dos referencias: español y Romaní, además de realizar talleres de socialización en cada Kumpania.

Por último, en la etapa de difusión, se implementaron medios de difusión y acceso a la información estadística específicos para los grupos étnicos priorizados. En respuesta a los acuerdos del proceso de consulta y concertación con grupos étnicos, se realizaron cursos y talleres con sus autoridades y líderes para la difusión de la información censal a través de capacitaciones y entrega de fichas de caracterización. Para el caso de la población indígena, a nivel de resguardos.

Para el caso de la población negra, afrocolombiana, raizal y palenquera y en el marco de los acuerdos de consulta y concertación, se acordaron las siguientes acciones:

Difundir los resultados en página web: el DANE cuenta con un micrositio para los grupos étnicos, en el cual pondrá a disposición de las comunidades negras, afrocolombianas, raizales y palenqueras, y demás interesados, la información estadística recopilada en el Censo Nacional de Población y Vivienda 2018, de dichas comunidades y la evolución de las estadísticas de los censos de los años 1993, 2005 y 2018." Respecto a la información detallada del CNPV 2018 para la población Negra, Afrocolombiana, Raizal y Palenquera, estos datos se pueden consultarse en las siguientes herramientas:

- Micrositio para la población negra, afrocolombiana, raizal y palenquera.
- Microdatos Anonimizados con los resultados del Censo Nacional de Población y Vivienda –CNPV 2018.
- Realizar estudios especializados: "el DANE realizará estudios especializados con enfoque diferencial para comunidades negras, afrocolombianas, raizales y palenqueras, en autorreconocimiento, fecundidad incluyendo adolescentes, mortalidad incluyendo mortalidad infantil, migraciones y desplazamiento, caracterización por ciclo de vida, cambios demográficos, hogar, familia, funcionamiento humano, género y urbanización. En el desarrollo de los estudios se visibilizarán las brechas existentes entre la población NARP, el total nacional y las demás comunidades étnicas incluidas en el enfoque diferencial del DANE."
- Cartilla didáctica con los resultados del CNPV 2018 para la población negra, afrocolombiana, raizal y palenquera.
- Documento de sistematización de lecciones aprendidas del CNPV 2018.

ECV 2018:

La Encuesta de Calidad de Vida (ECV) es una investigación que el DANE realiza con el objeto de recoger información sobre diferentes aspectos y dimensiones del bienestar y las condiciones de vida de los hogares, incluyendo temas como: el acceso a bienes y servicios públicos, privados o comunales, salud, educación, atención integral de niños y niñas menores de 5 años, entre otros. La consideración de estos aspectos hace posible realizar posteriores análisis a los factores que explican los diferentes niveles de vida existentes en la sociedad.

Dada la importancia de reconocer las brechas sociales y económicas con enfoque étnico y territorial, la entidad se comprometió con la profundización del marco muestral de la Encuesta Nacional de Calidad de Vida, la cual, desde finales de 2018, logra una capacidad de inferencia de fenómenos con prevalencia poblacional inferior al 10% y con errores relativos máximos del 5%, y logra una representatividad a nivel departamental.Por lo anterior, desde ese año se cuenta con resultados representativos para total nacional, departamentos, cabecera y centros poblados - rural disperso.

El formulario de la ECV se compone de once módulos que se han venido aplicando de forma permanente en los últimos años, así como de capítulos especiales de aplicación periódica que responden a necesidades de usuarios específicos.

Los capítulos permanentes son: i) Datos de la vivienda, ii) Servicios del hogar, iii) Características y composición del hogar, iv) Salud, v) Atención integral de los niños y niñas menores de 5 años, vi) Educación, vii) Fuerza de trabajo, viii)Tecnologías de información y comunicación (TIC)2, ix) Trabajo infantil3 , x) Tenencia y financiación de la vivienda que ocupa el hogar, xi) Condiciones de vida del hogar y tenencia de bienes. El capítulo de TIC se incluyó en 2012 como resultado del creciente interés en el tema y la decisión de monitorearlo a través de una operación estadística más apropiada, mientras que el de Trabajo infantil se incorporó de manera peramente desde 2014 con el propósito de abordar esta problemática en la población entre 5 y 11 años.

El objetivo general de la ECV 2018 fue obtener información para analizar y realizar comparaciones de las condiciones socioeconómicas de los hogares colombianos, las cuales posibiliten hacer seguimiento a las variables necesarias para el diseño e implementación de políticas públicas. Dentro de los objetivos específicos se encuentran:

- Actualizar la información relacionada con las condiciones socioeconómicas de la población del país.
- Obtener la información necesaria para la actualización de los indicadores sociales a nivel de viviendas, hogares y personas y para la definición de políticas que permitan diseñar y ejecutar planes sociales.
- Brindar información que permita la obtención de los indicadores de pobreza multidimensional.
- Obtener información que posibilite profundizar en un análisis con perspectiva de género.
- Facilitar el seguimiento al cumplimiento de las metas asociadas a los ODS

En el diseño del formulario o cuestionario, en el capítulo D que corresponde a características y composición del hogar que se aplica a todas las personas del hogar, se indaga por el autorreconocimiento étnico, lo que permite comparar con los resultados de otras operaciones estadísticas como el CNPV.

Metodología

Con este contexto en mente, se diseñó y ejecutó una estrategia de cuatro fases que permite estimar el volumen poblacional de los grupos étnicos a nivel nacional, departamental y municipal, con los siguientes objetivos clave:

- Realizar diagnóstico de los municipios con bajas y altas prevalencias de grupos étnicos.
- Consolidar un conjunto de aprendizaje integrando registros administrativos, encuestas a hogares y fuentes alternativas de información a nivel de grilla.
- Aplicar modelos de aprendizaje de máquinas (*machine learning*) a nivel de persona para determinar su pertenencia étnica y a nivel geográfico para estimar su participación.
- Diagnosticar los modelos aplicados y realizar cálculo de impactos.

De esta manera, se trazó una estrategia de cuatro fases para lograr dicha estimación, a través de asignación probabilística, en los niveles requeridos. La primera fase (I) consistió en determinar la proporción de personas pertenecientes a grupos étnicos a nivel nacional, utilizando la ECV 2018 dadas sus características idóneas que permiten caracterizar la población efectivamente censada sin información de autorreconocimiento étnico. La segunda fase (II) pretende asignar pertenencia étnica de la población efectivamente censada en el CNPV 2018 usando modelos predictivos de clasificación adaptados desde los métodos presentados en (Voicu, 2018), que mediante modelos de *machine learning* bayesianos determinan la pertenecía étnica de las personas usando los nombres como covariados. La tercera fase (III) consiste en utilizar la georreferenciación de la población efectivamente censada para determinar la participación de cada grupo étnico en la población omitida. Por último, en la cuarta fase (IV) se utiliza la metodología de tabla cuadrada para consolidar los resultados de las fases II y III, estimando así el volumen poblacional de los grupos étnicos a nivel municipal y departamental.

2. Fase I - Determinar la participación de personas auto reconocidas como pertenecientes a un grupo étnico a nivel nacional.

Dentro del total de personas efectivamente censadas en el **CNPV 2018**, la participación étnica presentó el siguiente comportamiento:

Tabla 1 - Participación de grupos étnicos en la población efectivamente censada CNPV 2018

Grupo Étnico	Partencia Étnica	Total	Participación
1	Índígena	1,905,617	4.3%
2	Gitano(a) o Rrom	2,649	0.0%
3	Raizal del archipiélago de San Andres, Providencia y Santa Catalina	25,515	0.1%
4	Palenquero(a) de San Basilio	6,637	0.0%
5	Negro(a), mulato(a), afrodescendiente, afrocolombiano(a)	2,950,072	6.7%
6	Ningún grupo étnico	38,678,341	87.6%
9	Sin información	595,586	1.3%
Total		**44,164,417**	**100.0%**

Sin embargo, fuentes de información como la Encuesta de Calidad de Vida 2018 (**ECV**) y la gran encuesta integrada de Hogares 2018 (**GEIH**) muestran un comportamiento diferente en el volumen de la población Negro(a), mulato(a), afrodescendiente, afrocolombiano(a) (NMAA). La siguiente la tabla muestra la estimación del volumen y sus respectivos coeficientes de variación:

Tabla 2. Comparación del volumen poblacional en ECV y GEIH

		Intervalo de confianza del 95%		
Fuente	Total	Límite inferior	Límite Superior	Coeficiente de Variación
ECV	4.637.483	4.406.182	4.868.785	2,5
GEIH	4.472.599	4.326.692	4.618.506	1,7

Los errores relativos de estas dos encuestas de hogares para el volumen de población autorreconocida como NMAA son inferiores al 5%. Este resultado permite inferir que ambas fuentes son referentes robustos del total de dicha población. A nivel departamental, las dos encuestas tienen representatividad, sin embargo, con la GEIH solo puede estimarse la población NMAA en su acumulado anual, para los departamentos de Amazonas, Arauca, Casanare, Putumayo, Guainía, Guaviare, Vaupés y Vichada, limitando así el cálculo de la participación NMAA en el total nacional. Esto explica la diferencia de 164.884 personas autorreconocidas como NMAA entre la GEIH y la ECV. Teniendo en cuenta lo anterior,

se selecciona la estimación de ECV correspondiente a 4.637.483 personas autorreconocidas como NMAA.

Ahora bien, dada la necesidad de calcular las participaciones de los grupos étnicos a niveles subnacionales, una exploración a nivel departamental de las prevalencias del autorreconocimiento en la ECV, evidencia en los errores relativos y una limitada capacidad conclusiva. La siguiente tabla muestra las participaciones de la poblacional NMAA departamental a partir de esta fuente.

Tabla 3. Proporción de autorreconocimento NMAA por Departamento en la ECV 2018

Departamento	Proporción de personas que se autorreconocen como negras, mulatas, afrodescendientes, raizales y palenqueras en la ECV			
	%	L Inf.	L Sup.	CVe
Amazonas	0.1	0.0	0.3	38.1
Antioquia	8.9	6.8	10.9	11.7
Arauca	3.0	1.8	4.1	19.4
Atlántico	2.8	2.1	3.4	12.4
Bogotá	0.5	0.1	0.8	36.7
Bolívar	34.6	31.5	37.8	4.7
Boyacá	0.3	0.1	0.6	36.5
Caldas	1.0	0.6	1.4	21.7
Caquetá	0.4	0.2	0.6	21.4
Casanare	1.1	0.5	1.7	27.8
Cauca	16.3	14.1	18.5	7.0
Cesar	5.6	4.7	6.4	8.1
Córdoba	4.0	3.3	4.8	9.7
Cundinamarca	0.6	0.3	0.9	27.8
Chocó	84.9	80.9	88.8	2.4
Guainía	0.7	0.3	1.1	30.8
Guaviare	3.9	2.7	5.2	16.0

Huila	0.3	0.1	0.5	37.0
La Guajira	12.8	10.9	14.6	7.6
Magdalena	9.1	7.9	10.3	6.7
Meta	1.6	1.0	2.2	18.2
Nariño	25.9	24.1	27.7	3.6
Norte de Santander	1.3	0.9	1.7	16.1
Putumayo	3.7	2.5	4.9	16.5
Quindío	1.5	0.9	2.1	21.7
Risaralda	0.9	0.4	1.4	29.8
San Andrés	18.2	15.8	20.6	6.7
Santander	0.8	0.4	1.2	24.1
Sucre	9.1	7.3	10.8	10.0
Tolima	0.3	0.1	0.4	28.5
Valle del Cauca	29.8	26.5	33.1	5.7
Vaupés	1.1	0.7	1.5	20.0
Vichada	0.8	0.4	1.2	26.9

De los 33 departamentos, solo Bolívar, Chocó y Nariño cuentan con un error relativo inferior al 5%, para todos los demás resulta inconveniente inferir volúmenes de esta población. En consecuencia, se determinó que era necesario plantear otras estrategias para determinar, de forma indirecta, el volumen de esta población a niveles subnacionales, lo cual representa uno de los desafíos más grandes de la etapa postcensal del CNPV 2018, tanto a nivel teórico como metodológico; también desde el punto de vista de la capacidad computacional de la entidad.

Dado que en el CNPV 2018, 595.586 personas no cuentan con información de autorreconocimiento étnico, dentro de la población efectivamente censada, se determinó que, en caso de contar con pertenecía étnica, las participaciones de los grupos étnicos podrían presentar un cambio significativo, en niveles bajos de desagregación, en el caso que el patrón de dispersión de esta población se encuentre concentrado geográficamente. Por ejemplo, si en el municipio de Tumaco en el departamento de Nariño, se asignará de forma probabilística un grupo étnico a las personas “sin información de

autorreconocimiento", la razón entre las personas en la categoría de NMAA y el total poblacional en el municipio, cambiaria de forma significativa.

Teniendo en cuenta lo anterior, se optó por utilizar la estructura étnica de la Encuesta Nacional de Calidad de Vida (ECV 2018), ya que los resultados de esta encuesta para población étnica a nivel nacional en el año 2018 cuentan con un coeficiente de variación que permite realizar inferencias sobre la participación poblacional. Los diseños muestrales de las encuestas y sus resultados demográficos se encuentran dentro del estándar internacional de calidad estadístico.

Tabla 4. Participación y coeficientes de variación para grupos étnicos en la ECV (2018)

Grupo Étnico	Partencia Étnica	Participación	CV
1	Indígena	4,34%	3,2
2	Gitano(a) o Rrom	0,01%	38,7
3	Raizal del archipiélago de San Andres, Providencia y Santa Catalina	0,04%	12,5
4	Palenquero(a) de San Basilio	0,02%	25,2
5	Negro(a), mulato(a), afrodescendiente, afrocolombiano(a)	9,28%	2,5
6	Ningún grupo étnico	86,31%	0,3

3. Fase II – Asignación probabilistica de la pertenencia étnica usando modelos predictivos de clasificación

La creciente capacidad de almacenamiento y procesamiento de información, abren la posibilidad de incrementar el conocimiento sociodemográfico, especialmente en las desagregaciones étnica gracias a la capacidad de realizar clasificaciones individuales de forma probabilística de personas como pertenecientes a diferentes grupos étnicos. Sin embargo, ante la ausencia de información que permita de forma directa determinar la pertenencia étnica, una de las formas más simples y eficaces es determinarla de forma indirecta empleando otro tipo de datos, como el lugar de residencia y nombres.

3.1 Metodología

Como propuesta metodológica, se adaptan los métodos presentados en (Voicu, 2018), que mediante modelos de *machine learning* bayesianos determinan la pertenecía étnica de las personas usando los

nombres como covariados. Esta asignación se da de manera probabilística usando las siguientes probabilidades condicionales:

- $\boldsymbol{P(r|a)}$, probabilidad de pertenecer a un grupo étnico dado el apellido de la persona, con $\boldsymbol{r}$ el grupo étnico y $\boldsymbol{a}$ el apellido de la persona.
- $\boldsymbol{P(n|r)}$, probabilidad de llamarse de determinada manera dado que se pertenece a un grupo étnico, con $\boldsymbol{n}$ el nombre de la persona y $\boldsymbol{r}$ el grupo étnico
- $\boldsymbol{P(t|r)}$, probabilidad de residir en determinada ubicación geográfica dado que se pertenece a un grupo étnico, con $\boldsymbol{t}$ desagregación geográfica y $\boldsymbol{r}$ el grupo étnico

Las anteriores probabilidades son los insumos para determinar la probabilidad de pertenecer a un grupo étnico, dados el apellido, el nombre y el territorio en el que reside. Lo anterior se describe de forma matemática como:

$$P(r|a,n,t) = \frac{P(r|a) * P(n|r) * P(t|r)}{\sum_{r=1}^{6} P(r|a) * P(n|r) * P(t|r)}$$

Estas probabilidades condicionales se pueden calcular con las proporciones observadas en el CNPV 2018, algunos ejemplos de proporciones para este departamento del Chocó se muestran en las siguientes tablas:

Tabla 5. Ejemplos de proporciones para el departamento del Chocó

Apellido	Grupo Étnico	P(r|a)
AGUILAR	1	0.0062
AGUILAR	4	0.0012
AGUILAR	5	0.9589
AGUILAR	6	0.0337
MOSQUERA	1	0.0017
MOSQUERA	3	0.0001
MOSQUERA	4	0.0002
MOSQUERA	5	0.9879
MOSQUERA	6	0.0100
LOZANO	1	0.0036
LOZANO	5	0.9749
LOZANO	6	0.0215

Grupo Étnico	Nombre	P(r|n)
5	BAYRON	0.9000
6	BAYRON	0.1000
1	DILMAR	0.4000
5	DILMAR	0.6000
1	JUANA	0.0252
5	JUANA	0.9627
6	JUANA	0.0121

Grupo Étnico	Municipio	P(t|r)
1	Quibdó	0.0402
2		0.0000
3		0.0004
4		0.0002
5		0.9275
6		0.0318

Se descartan del análisis los grupos étnicos Gitano o Rrom, Raizal del Archipiélago de San Andrés providencia y Santa Catalina, y Palenquero(a) de san Basilio, dado que su participación en el total nacional es de 0.006%, 0.058% y 0.015% respectivamente, la inclusión de dichas poblaciones dentro del análisis podría afectar el desempeño de los modelos predictivos. Para efectos del cálculo de probabilidad compuesta de que una persona se autor reconozca como afro, se organizan los municipios del país alrededor de regiones que incluyen municipios con patrones socioculturales y territoriales compartidos.

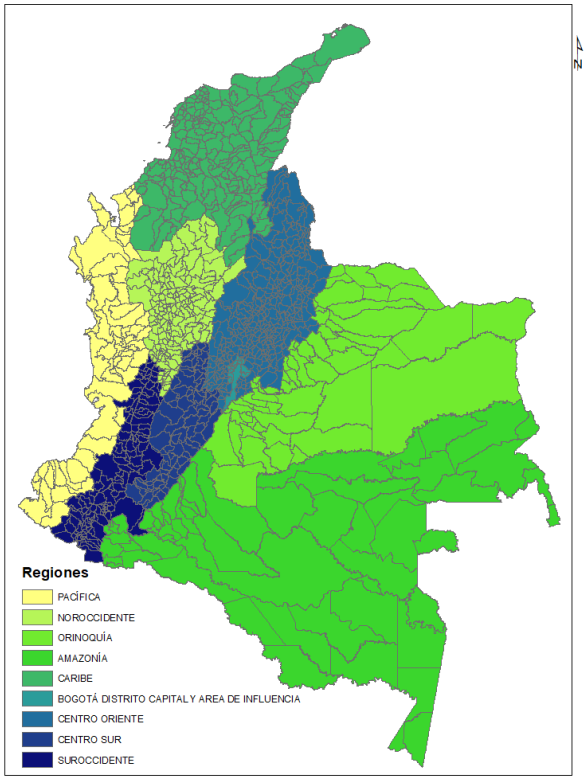
Tabla 6. Descripción de las regiones socioculturales identificadas

Etiqueta	Región	Descripción
1	PACÍFICA	La región se organiza a través de municipios que pertenecen al Andén Pacífico y que incluyen municipios de los departamentos de Chocó, Antioquia, Risaralda, Valle del Cauca, Cauca y Nariño. Es una región con alta prevalencia de población afro, entre ella, población que ha ocupado históricamente zonas de las cuencas del pacífico y que a través de prácticas tradicionales y organizativas se organiza a través de consejos comunitarios en territorios colectivos de comunidades negras.
2	NOROCCIDENTE	Región que incluye municipios que pertenecen a la zona influencia de la colonización antioqueña. Parte de su población afro ha estado en la zona desde la época de la colonia y otra ha migrado de la región pacífica.
3	ORINOQUÍA	La región articula los departamentos que se ubican en los llanos orientales de Colombia. Adicionalmente cuenta con baja prevalencia de población Afro que generalmente ha llegado a las zonas por procesos migratorios recientes.
4	AMAZONÍA	Departamentos localizados en la selva húmeda tropical de la cuenca del río Amazonas, con baja prevalencia de la población afro que llegó a las zonas a través de procesos migratorios recientes.
5	CARIBE	Departamentos que comprenden a los departamentos del norte del país con influencia del Caribe, con presencia significativa de población Afro, debido al tráfico trasatlántico de población africana esclavizada en épocas coloniales a través del puerto de Cartagena.
6	DISTRITO CAPITAL Y ÁREA DE INFLUENCIA	Comprende a Bogotá Distrito Capital, y a los municipios que son su zona de influencia por constituirse en municipios dormitorio de la capital. Cuenta con una población afro significativa principalmente en Bogotá y Soacha, aunque cuenten con porcentajes de participación relativamente bajos por al alto

		volumen de población que no se reconoce dentro de un grupo étnico.
7	CENTRO ORIENTE	Departamentos que cubren la zona centro-norte de la cordillera oriental, con procesos prolongados de mestizaje entre población indígena y de descendencia española, y baja prevalencia de población Afro, cuya presencia obedece a procesos migratorios recientes.
8	CENTRO SUR	Departamentos que corresponden al denominado Tolima grande ubicado sobre las cordillera central y oriental, y el valle del río Magdalena, con procesos prolongados de mestizaje entre indígenas y descendientes de españoles, y baja prevalencia de población afro.
9	SUROCCIDENTE	Municipios del suroccidente del país ubicados en zona andina y el valle del río Cauca, con alta presencia de población indígena y afro.

La ilustración 1 muestra la conformación de las de las regiones descritas anteriormente.

Ilustración 1. Conformación de las regiones socioculturales identificadas



Con el fin de mitigar el sesgo de escritura presente en los nombres y apellidos, debido a que son susceptibles a presentar errores de escritura, se hace relevante la implementación de una función que convierta una cadena de caracteres a su fonema y lograr identificar el nombre por el sonido que produce al ser pronunciado, el objetivo básico es codificar de la misma forma los nombres con la misma pronunciación. Para lograr este objetivo se desarrolló una función que recibe una cadena de caracteres como parámetro, realiza la conversión de cada carácter y retorna su representación fonética. Algunos ejemplos son:

Tabla 7. Ejemplos resultados de la función fonética

ENTRADA	SALIDA
YOHANA	YN
JOANA	YN
JENNY	YNI
YENY	YNI
GONZALES	GNZLZ
GONZALEZ	GNZLZ

Posteriormente se realiza el cálculo de las probabilidades condicionales usando la regionalización final con los nombre y apellidos en sus formas fonéticas, por último, se procede a realizar el ajuste de los modelos empleando los **43,534,030** registros de la población efectivamente censada que cuenta con información en la pregunta de autorreconocimiento étnico y no pertenece a las etnias Rom, Raizal del Archipiélago de San Andrés providencia y Santa Catalina, y Palenquero(a) de san Basilio.

3.2 Resultados

Con la finalidad de asignar pertenecía étnica a las 595.586 personas que no cuentan con información, se tiene como insumo un conjunto de datos que cuenta con 43.534.030 filas y 24 variables que representan las probabilidades observadas de pertenecer a un grupo étnico dado algunas de las tres características indirectas de nombre, apellido y lugar de residencia. Luego se ajusta un modelo de aprendizaje de máquinas para clasificar estas personas en una de las 3 categorías objeto del análisis, 1 para indígenas, 5 para NMAA y 6 para ningún grupo étnico.

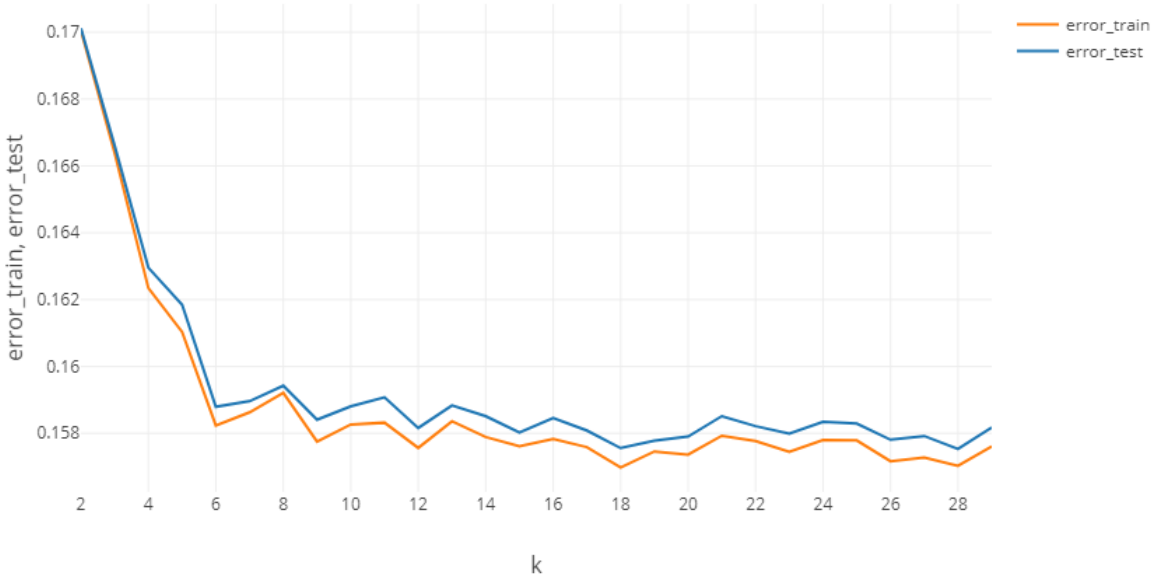
EL modelo es el *Random Forest* desarrollado por (Breiman, 2001) que se ajusta en la plataforma *Databricks* en lenguaje de programación Python, para dicho análisis se emplean las siguientes librerías:

- RandomForestClassifier,
- MulticlassMetrics,
- MulticlassClassificationEvaluator

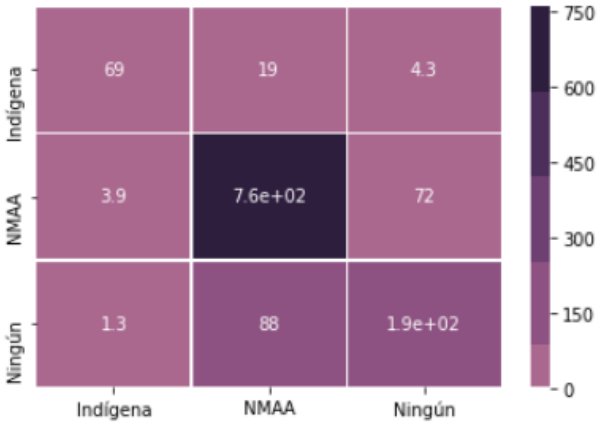
Para el ajuste del modelo, se realiza una partición del insumo en un conjunto de aprendizaje y un conjunto de prueba, con una distribución 80-20, es decir, 80% aprendizaje y 20% prueba, para cada una de las 9 regiones construidas.

A continuación, se presentan los resultados de las validaciones cruzadas sobre los errores de predicción, seguido se selecciona el número de árboles que hacen que el error de entrenamiento y prueba sea mínimo; por último, se realizan las predicciones con el parámetro seleccionado y se evalúan las matrices de confusión en miles de personas para todas las regiones sobre los conjuntos de entrenamiento y prueba:

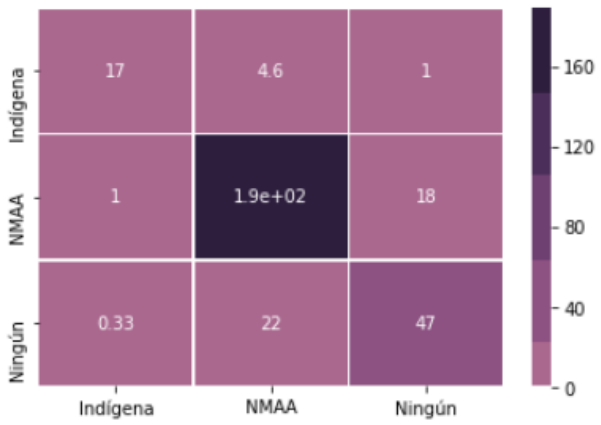
Gráfica 1. Resultados fase II - Región Pacífica (1)



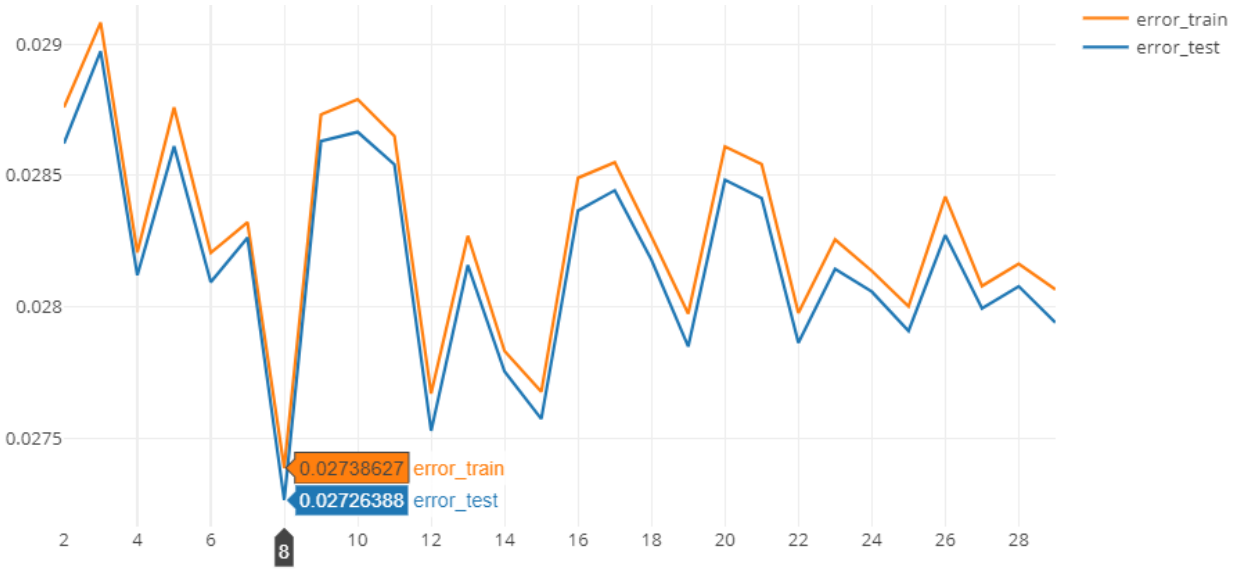
Conjunto de entrenamiento



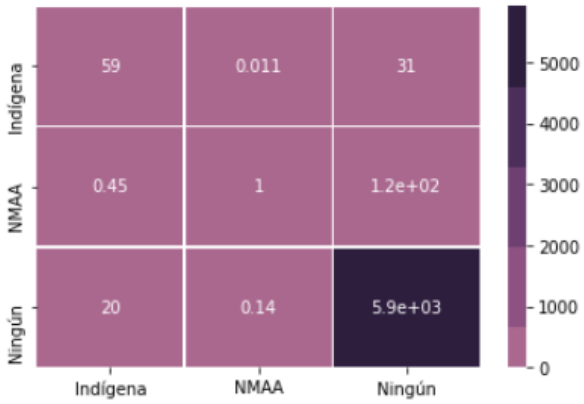
Conjunto de prueba



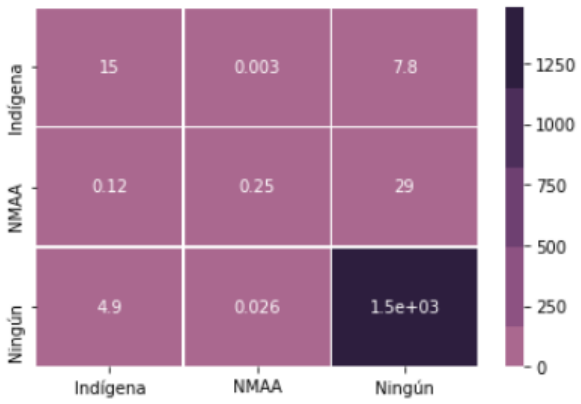
Gráfica 2. Resultados fase II - Región Noroccidente (2)



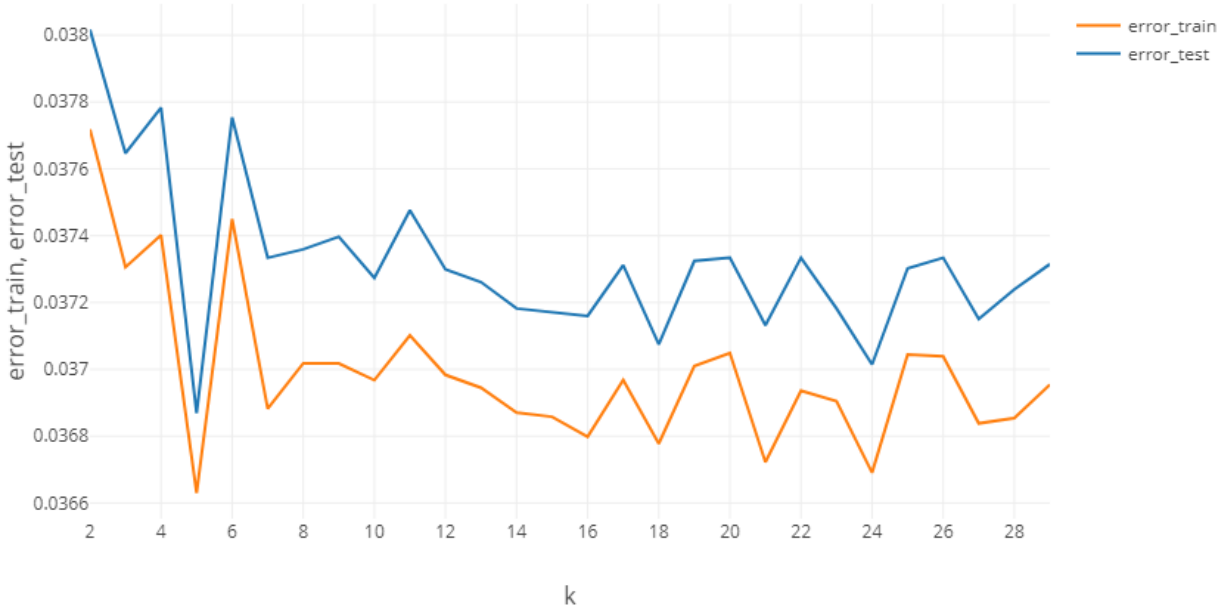
Conjunto de entrenamiento



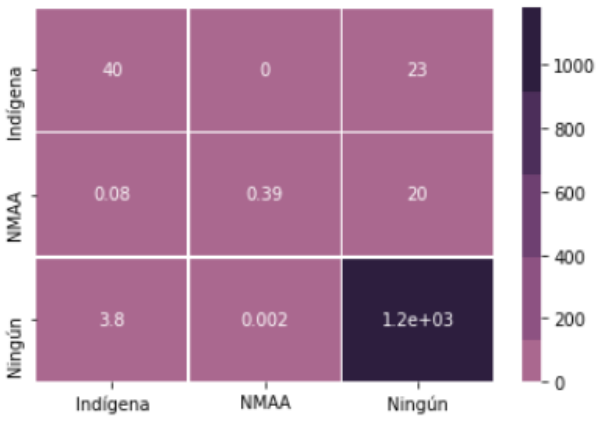
Conjunto de prueba



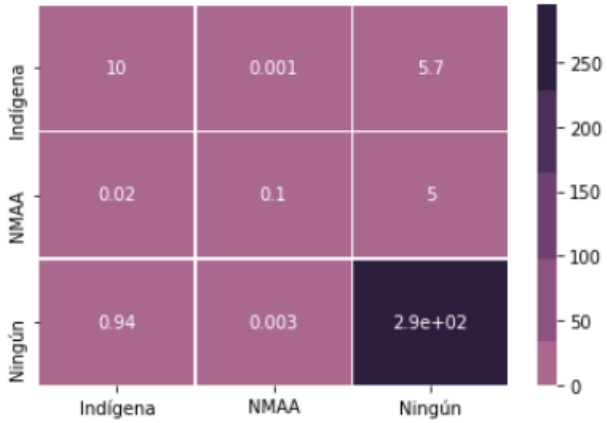
Gráfica 3. Resultados fase II - Región Orinoquía (3)



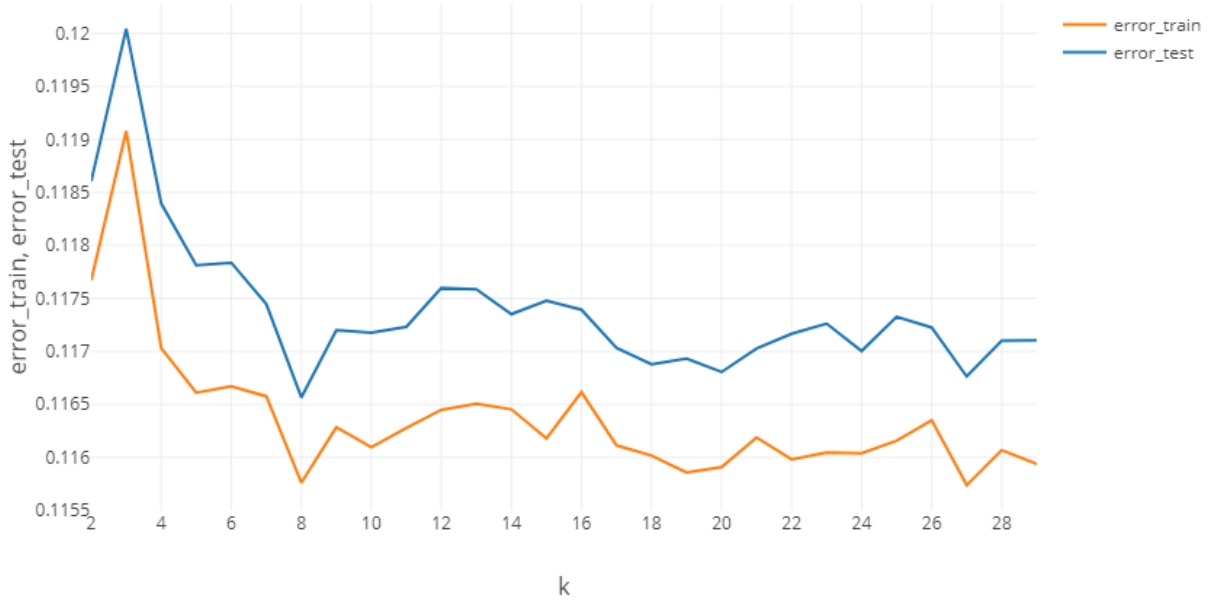
Conjunto de entrenamiento



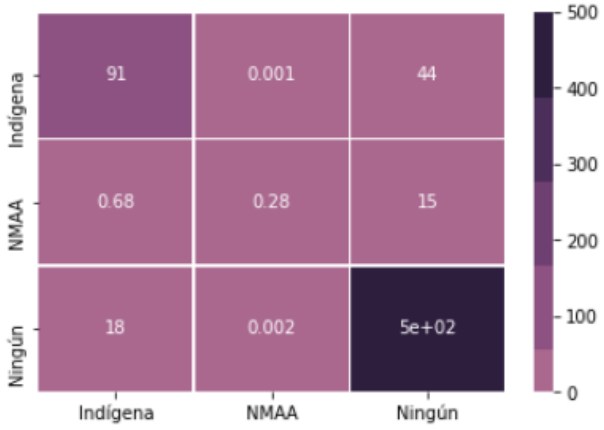
Conjunto de prueba



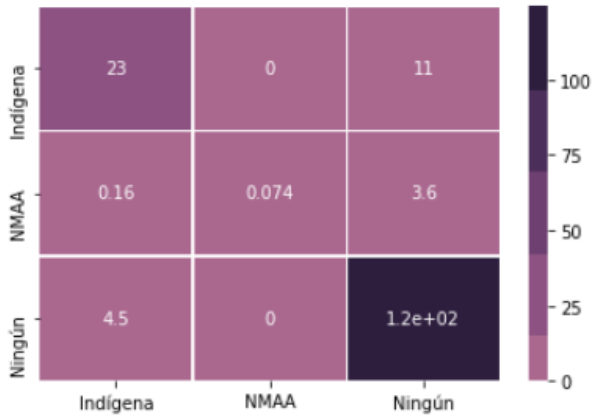
Gráfica 4. Resultados fase II - Región Amazonía (4)



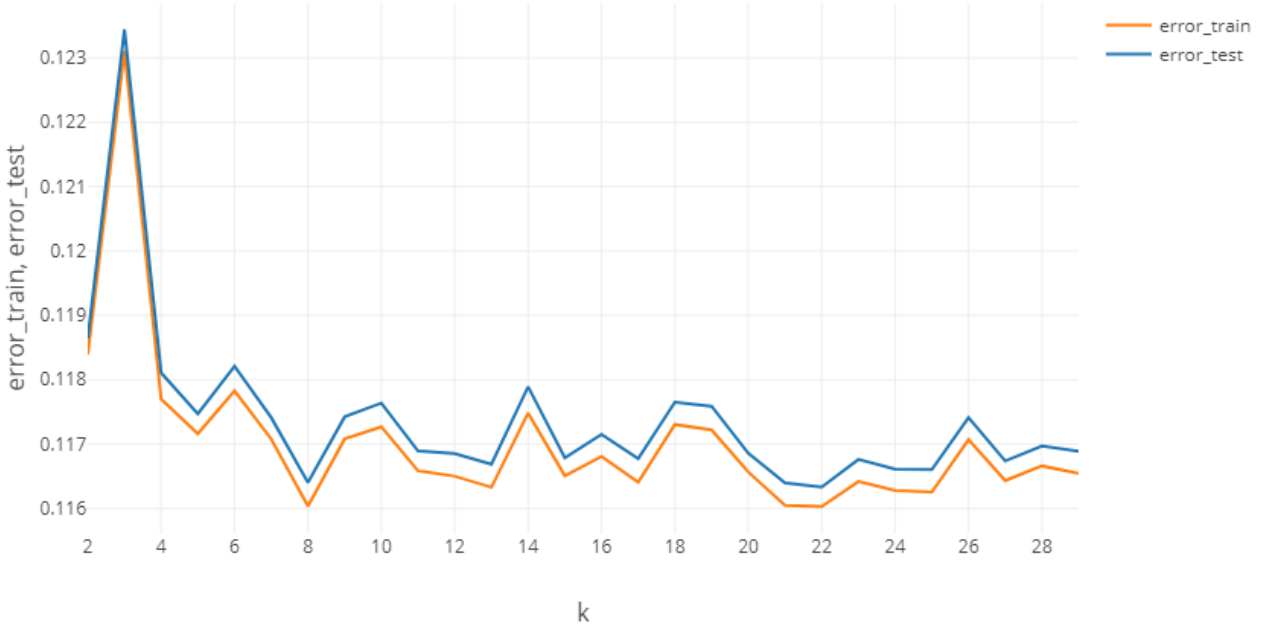
Conjunto de entrenamiento



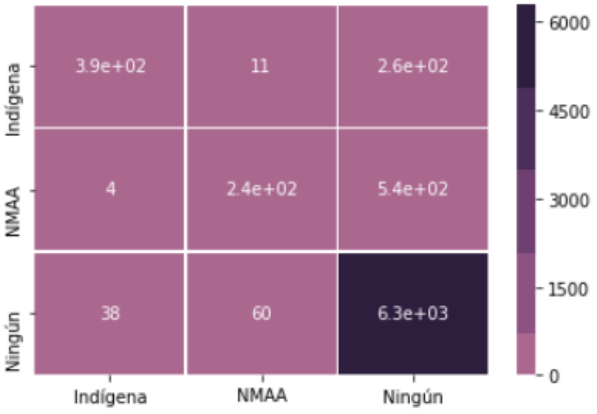
Conjunto de prueba



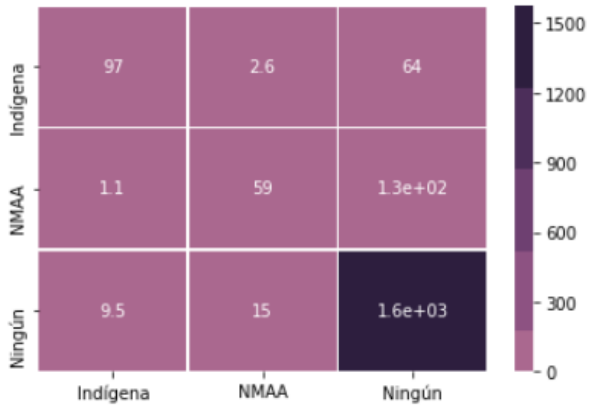
Gráfica 5. Resultados fase II - Región Caribe (5)



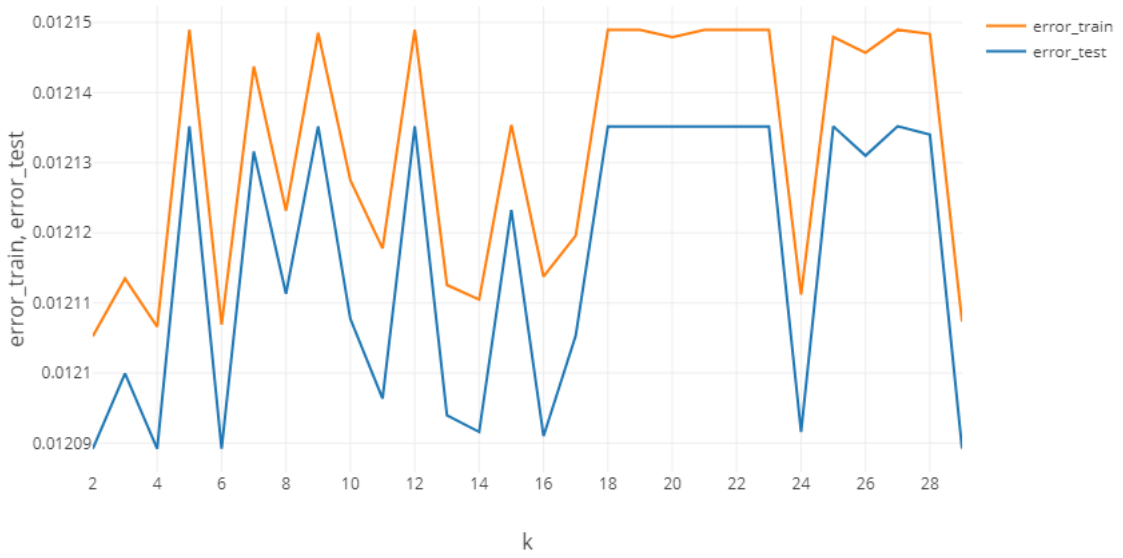
Conjunto de entrenamiento



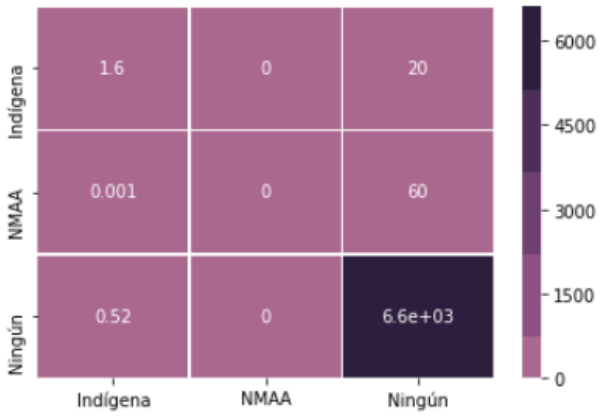
Conjunto de prueba



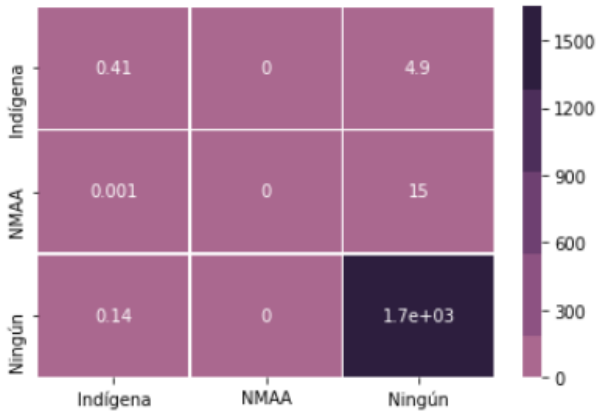
Gráfica 6. Resultados fase II - Región Bogotá (6)



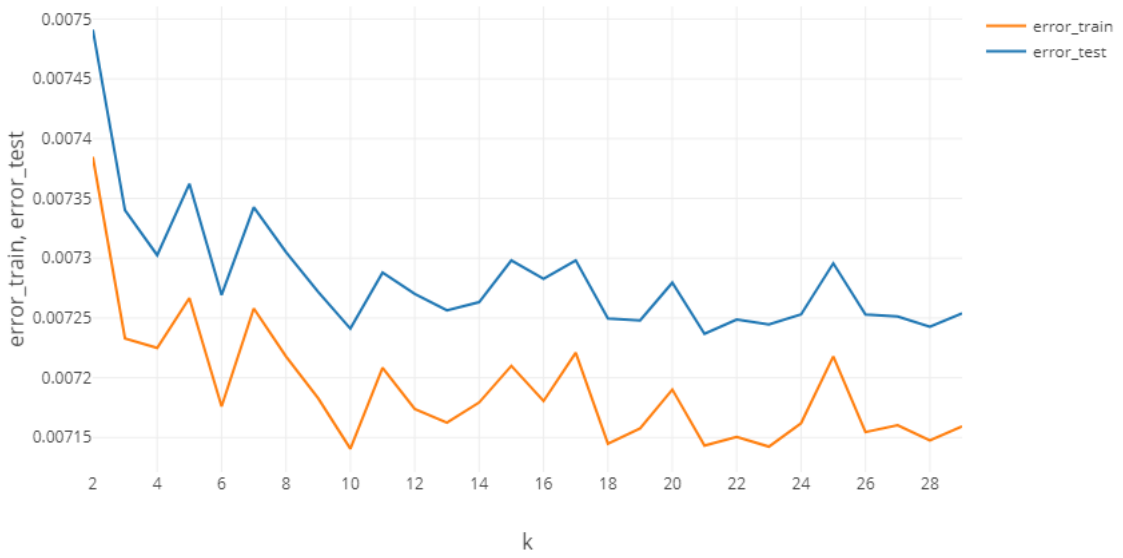
Conjunto de entrenamiento



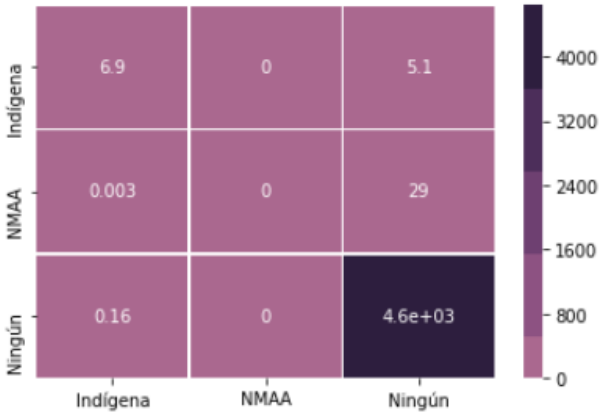
Conjunto de prueba



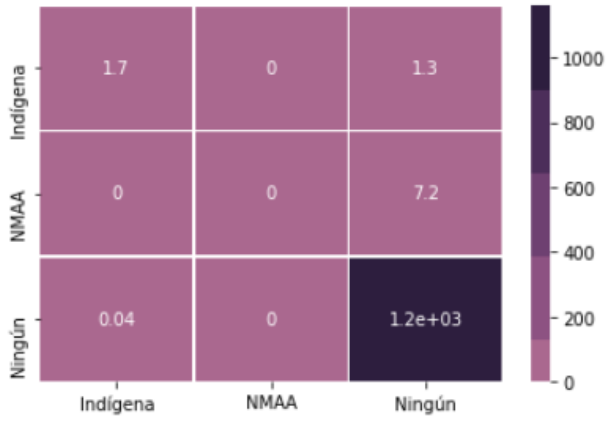
Gráfica 7. Resultados fase II - Región Centro Oriente (7)



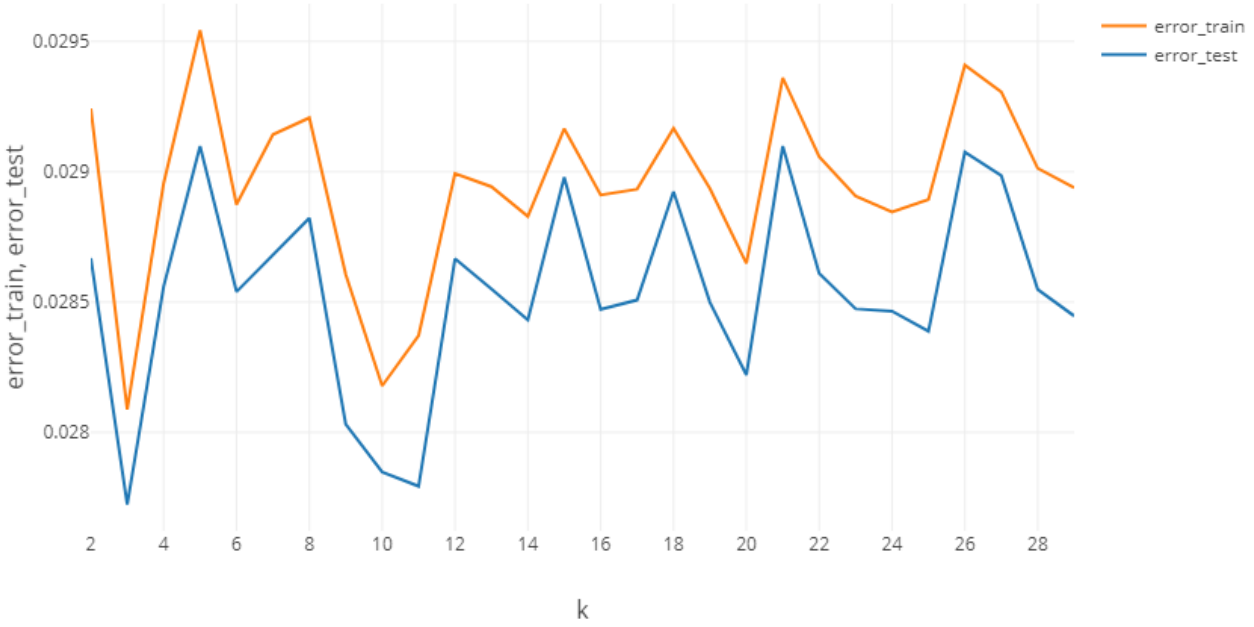
Conjunto de entrenamiento



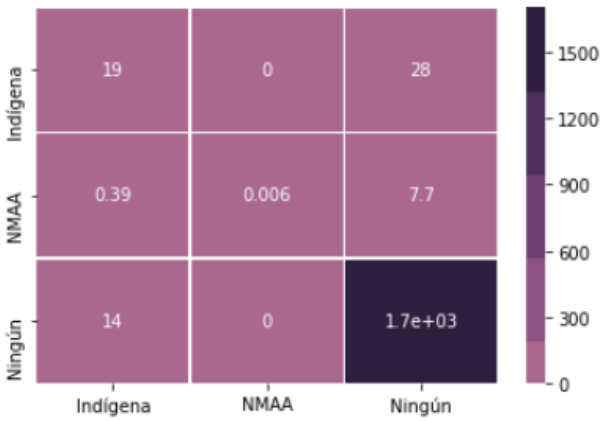
Conjunto de prueba



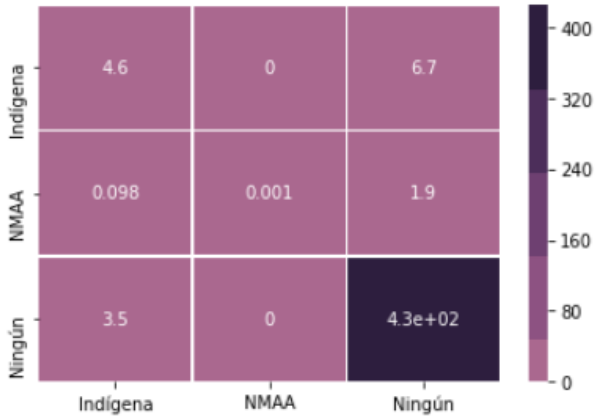
Gráfica 8. Resultados fase II - Región Centro Sur (8)



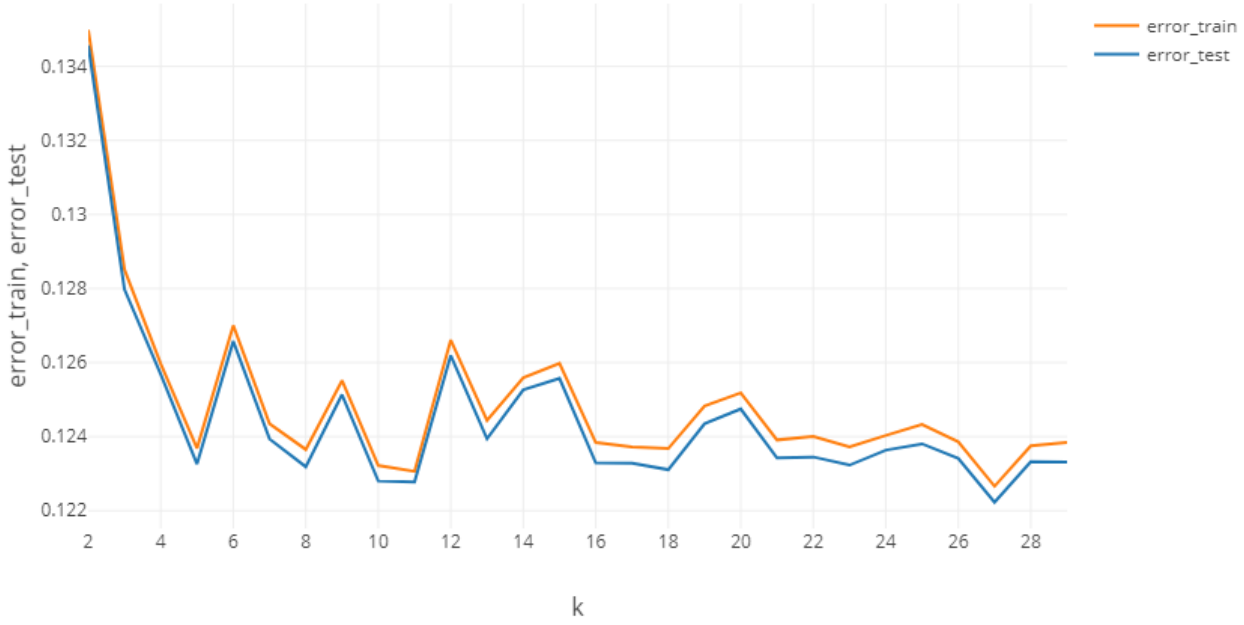
Conjunto de entrenamiento



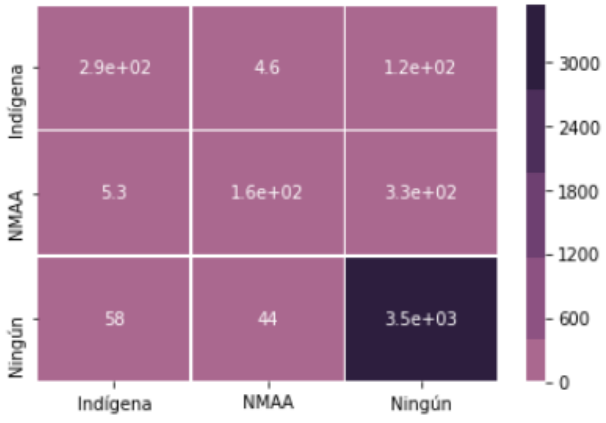
Conjunto de prueba



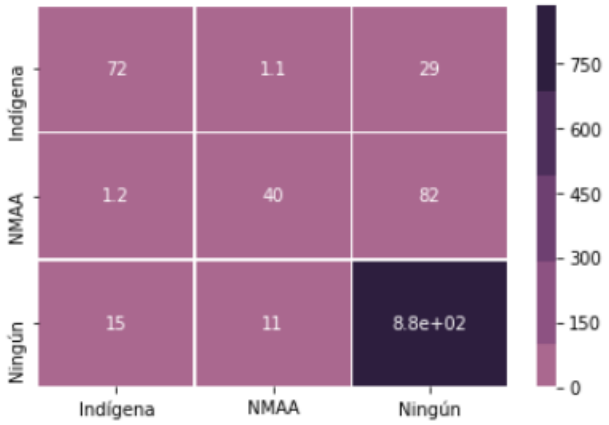
Gráfica 9. Resultados fase II - Región Surocidente (9)



Conjunto de entrenamiento



Conjunto de prueba



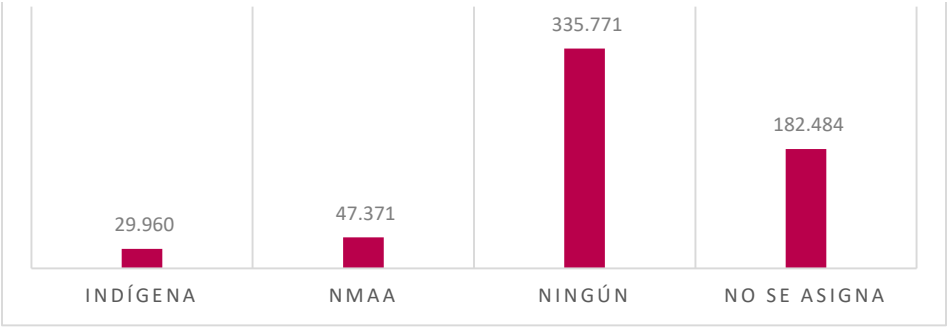
El ajuste de los modelos empleando la regionalización final es satisfactorio, sin embargo, las regiones de Bogotá y Centro sur no alcanzan un poder predictivo aceptable, por lo tanto, no se recomienda el uso de dichos modelos para predecir el autorrecnocimiento en estas dos regiones.

El ajuste en la región Pacífica es el que presenta mejor comportamiento, dado que la distribución de los grupos étnicos en dicho territorio permite un gran poder predictivo, como se puede observar en sus

matrices de confusión, en donde la diagonal de la matriz acumula la mayor cantidad de individuos.

Para las demás regiones, el desempeño es aceptable, sin embargo, las diagonales de las matrices de confusión no acumulan la mayor cantidad de individuos, siendo la categoría de NMAA la que presenta la mayor movilidad hacia la categoría de ningún grupo étnico. Este comportamiento, es explicado por la variabilidad que presentan los nombres NMAA y su similitud con los nombres de las personas que se autorreconocen en ningún grupo étnico. A pesar de este comportamiento, la predicción realizada por los modelos permite estimar la pertenencia étnica de forma satisfactoria. El resultado final de las predicciones se muestra en el gráfico 10, donde, de las 595.586 personas que no cuentan con información en la pregunta de pertenecía étnica, el 5% son asignados a la categoría indígena, el 8% a NMAA, el 56,4% a Ningún grupo étnico y el 30.6% no es posible asignar grupo étnica con esta metodología.

Gráfica 10: Resultados finales Fase II



Los cinco municipios que mayor cantidad de indígenas y NMAA reciben producto de la imputación se muestran en las s tabla 8 y tabla 9:

Tabla 8: Municipios con mayor ajuste Indígena

Departamento	Municipio	Indígena
VICHADA	CUMARIBO	2,060
CHOCO	ALTO BAUDO	2,003
LA GUAJIRA	DIBULLA	1,940
CHOCO	BOJAYA	1,397
CHOCO	LLORO	1,172

Tabla 9: Municipios con mayor ajuste NMAA

Departamento	Municipio	NMAA
NARIÑO	TUMACO	6,000
CHOCO	RIOSUCIO	4,763
VALLE DEL CAUCA	BUENAVENTURA	4,169
NARIÑO	BARBACOAS	2,634
CHOCO	QUIBDO	2,105

4. Fase III – Utilización de la georreferenciación de la población efectivamente censada para determinar la participación de cada grupo étnico en la población omitida

En los últimos años, el creciente volumen de información recolectada en registros administrativos, redes sociales, imágenes satelitales y variables geoespaciales, son recursos muchas veces inexplorados y por ende subutilizados en demografía. La principal limitación para el aprovechamiento de esta información radicaba en los problemas de calidad y oportunidad de los conjuntos de datos, y en los problemas de viabilidad de su uso por carencias en la capacidad de cómputo, procesamiento y almacenamiento.

En (Stevens, 2015) se presentan tres casos (Camboya, Kenia y Vietnam) en los cuales fue posible mapear la distribución espacial de la población censada de todo un país a niveles de desagregación muy finas, gracias a la integración de censos oficiales y de información auxiliar que incluye variables geoespaciales captadas remotamente por imágenes satelitales. El uso de dichas variables se sustenta en la alta correlación entre la distribución espacial de dichas variables y la presencia de población humana en el territorio.

La metodología propuesta en Stevens (2015) emplea técnicas de aprendizaje de máquinas a niveles oficiales de desagregación geográfica. El modelo es entrenado al menor nivel de desagregación que disponga la operación censal, y "aprende" de los covariados espaciales y las variables censales, para luego predecir ponderaciones de densidad de población sobre una grilla o cuadrícula de 100 x 100 metros. Este modelado de ponderaciones es usado como superficie para llevar a cabo la redistribución simétrica de los conteos censales al nivel de dicha grilla.

4.1 Metodología: cálculo de omisión en resguardos indígenas

El Censo Nacional de Población y Vivienda (CNPV) realizado en 2018 tuvo entre sus objetivos contar y caracterizar a las personas residentes en Colombia, así como las viviendas y los hogares del territorio nacional. Los desafíos logísticos, temáticos y metodológicos que se desprenden de la realización de un censo son diversos, y el impacto de sus resultados son significativos. Una vez realizado el censo, surge la necesidad de estimar el volumen de población que por una u otra circunstancia no se censó. Para realizar la estimación de población omitida en el CNPV 2018, el DANE implementó el método directo[1], gracias al empadronamiento del censo por documento de identidad. Este método consiste en la implementación de técnicas de captura-recaptura, usando una encuesta de post-enumeración. En este caso se usó la Encuesta de Calidad de Vida de 2018 y las variables de identificación de cada persona mejoradas mediante el emparejamiento con los registros administrativos de la Registraduría Nacional. Este cálculo arrojó como resultado un 8,5% de omisión a nivel nacional.

Sin embargo, la técnica implementada para el cálculo de omisión nacional no permite hacer desagregaciones geográficas, debido a que la recolección de información en el censo se realizó en un periodo extendido que viola el supuesto de no migración interna. Para medir la omisión a niveles sub-nacionales, fue necesario adoptar metodologías en las cuales los efectos de la migración interna durante el periodo extendido no sesguen los resultados.

Con el fin de mitigar las fuentes de variabilidad más comunes de una operación censal, la estrategia implementada incorporó los siguientes procedimientos:

1 Ver (CEPAL, 2014)

- Estimación de personas en viviendas ocupadas con todas las personas ausentes (Promedios ponderados con granularidad geográfica).

- Estimación de personas por sub-enumeración dentro de los hogares (ajuste del tamaño promedio del hogar a nivel municipal).

- Estimación de personas en zonas no visitadas o incompletas (Modelo jerárquico bayesiano en áreas censadas por rutas, y promedios ponderados con granularidad geográfica para áreas censadas mediante el método de barrido).

La implementación de estos procedimientos para los tres escenarios de omisión censal arrojó resultados de población omitida a nivel de unidad de cobertura urbana y rural. Los detalles de este proceso pueden ser consultados en el documento Evaluación de cobertura nacional y subnacional CNPV 2018, en DANE (2020).

Las unidades de cobertura son unidades geográficas definidas y delimitadas por el DANE con fines de planeación, desarrollo y control operativo del censo, y no equivalen a niveles de desagregación territorial conocidos como barrios o comunas en zonas urbanas, o como veredas o predios en zonas rurales. Sin embargo, cada unidad operativa está contenida completamente en una clase de un municipio específico. Esto quiere decir que una unidad de cobertura nunca cubre parte de una cabecera municipal y del área resto de un mismo municipio, o parte de dos municipios diferentes. Esta exclusividad de la unidad de cobertura en una clase de un único municipio permitió estimar la población ajustada a todos los niveles subnacionales de la división político-administrativa del país (departamento, municipio, clase) por adición de la población ajustada en las unidades operativas contenidas en cada uno de esos niveles.

Sin embargo, las unidades operativas no cumplen la condición de exclusividad territorial con respecto a los resguardos indígenas. Para este nivel de desagregación se presentaban dos tipos de unidades de cobertura:

- *Situación A*- Unidades de cobertura completamente contenidos en un solo resguardo

- *Situación B*- Unidades de cobertura con una parte dentro de un resguardo, y otra parte fuera de él.

La superposición de unidades de cobertura censales con los polígonos de los resguardos indígenas, generan tres escenarios distintos, dados por las posibles combinaciones de las dos situaciones anteriores:

- *Escenario 1*. A-A…: resguardo conformado por una o varias unidades de cobertura contenidas completamente dentro del resguardo.

- *Escenario 2*. A-B…: resguardo conformado por alguna(s) unidad(es) de cobertura contenida(s) completamente en su interior, y fragmentos de unidades de cobertura que también cubren áreas fuera del resguardo.

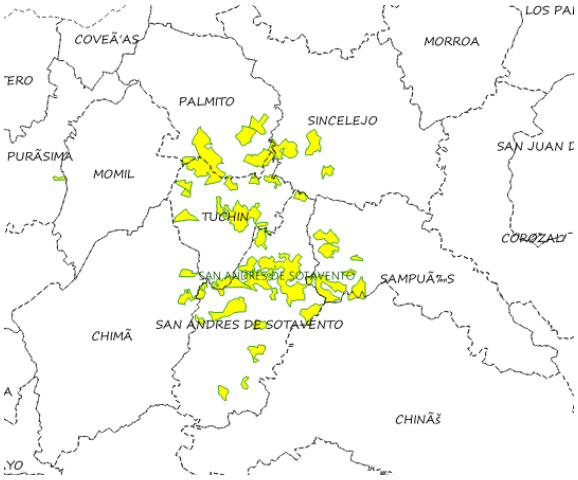
- *Escenario 3*. B-B: resguardo conformado por uno o varios fragmentos de unidades de cobertura que también cubren áreas fuera del resguardo.

Esta diversidad de escenarios se deriva del gran número de particularidades que dichos territorios presentan con respecto a sus límites geográficos, y al tamaño de su territorio. De acuerdo con estos criterios, los resguardos constituyen un conjunto de escenarios completamente heterogéneo.

Con respecto al tamaño de su territorio, existen resguardos con áreas muy pequeñas como el resguardo Anaba (Código DANE 1807) en Ortega –Tolima, que cuenta con 9,9 hectáreas, mientras que existen otros resguardos que abarca varios municipios, como el resguardo Alta y Media Guajira, cuya extensión excede las 930.000 has. y tiene jurisdicción en cinco municipios.

Con respecto a los límites geográficos de los resguardos, los resguardos no están siempre circunscritos a un solo municipio, y en algunos casos ni siquiera a un solo departamento. Por otra parte, el territorio de un resguardo no siempre constituye un área continua, algunos están conformados a partir de muchos predios discontinuos, como es el caso del resguardo San Andrés de Sotavento en el departamento de Córdoba, cuyo mapa se presenta a continuación.

Ilustración 2. Mapa resguardo San Andrés de Sotavento



Fuente: DANE, Marco geográfico CNPV 2018

En el caso de los resguardos del escenario 1, el cálculo de población ajustada por omisión podría resultar del mismo procedimiento utilizado para los niveles de desagregación DIVIPOLA, es decir, por adición de la población ajustada de las unidades de cobertura que lo componen. En el caso de los resguardos descritos en los escenarios 2 y 3, que están compuestos total o parcialmente por uno o varios fragmentos incompletos de unidades de cobertura, no es posible utilizar este mismo procedimiento.

Una primera opción para estimar la población ajustada de un fragmento de unidad operativa, es hacerlo a partir de la proporción de su área en la unidad de cobertura completa. Sin embargo, este procedimiento parte de un supuesto falso, según el cual la población censada y la población omitida de un área operativa se distribuye de manera homogénea en su territorio.

Este supuesto no es consistente con el patrón de distribución de la población en los resguardos, que es muy heterogénea. En algunos casos, las viviendas se concentran en un puerto ribereño, a lo largo del río, o en un claro del bosque, mientras buena parte de su territorio se encuentra despoblado. En otros casos, existe una distribución más homogénea y dispersa de la población. El patrón de asentamiento está relacionado tanto con las características naturales del territorio, como a su densidad poblacional, y en este segundo aspecto la heterogeneidad que presentan estas poblaciones es también enorme.

Por ende, se debe proponer e implementar estrategias diferenciadas para establecer un ajuste por cobertura dentro de los resguardos, que cumpla las siguientes condiciones:

1) Que sea consistente con la estimación de población ajustada por omisión a nivel de unidad operativa.

2) Que tenga en cuenta la distribución no homogénea de la población censada y omitida al interior de las unidades de cobertura.

Teniendo en cuenta lo anterior, esta sección tendrá como objetivo explicar las metodologías aplicadas en el cálculo de la omisión censal del CNPV en grupos étnicos.

4.1.1 Cálculo de omisión censal en resguardos indígenas implementando aprendizaje de máquinas.

Debido a la diversidad de estadísticas oficiales disponibles, y a sus diferencias en cuanto a estimadores, niveles de desagregación, periodicidad, etc., es necesario documentar cómo se adapta el método presentado en **(Stevens, 2015)** a la estimación de la omisión de población en resguardos en el caso colombiano.

Para desarrollar la metodología, se consolidaron tres conjuntos de datos como insumos: un conjunto de aprendizaje (conjunto de datos que ajustan los parámetros del modelo), un conjunto de prueba (datos para diagnóstico del modelo) y un conjunto de predicción (una vez ajustado y realizado el diagnóstico se predice sobre el nivel de interés). Los tres conjuntos cuentan con información de las mismas variables.

Conjunto de aprendizaje y prueba

El conjunto de datos de aprendizaje y prueba se encuentra desagregado a nivel de unidad de cobertura, que es la unidad geográfica más pequeña del marco operativo, a la cual se encuentran desagregados

los resultados del CNPV 2018[2], y, por otra parte, se restringe a las unidades de cobertura del área resto municipal (centros poblados y área rural dispersa), que es donde se ubican los resguardos indígenas. Como variable dependiente se toma la población ajustada por omisión, es decir, el número de personas efectivamente cesadas en la unidad de cobertura más el ajuste por omisión en dicha unidad (PER_AJUS).

Las variables explicativas utilizadas en la metodología provienen de tres tipos de fuente:

1) Censo Nacional de Población y Vivienda 2018 (CNPV 2018) y su ajuste por omisión a nivel de unidad de cobertura.

2) Registros administrativos: base de datos de instituciones educativas del Ministerio de Educación Nacional (2019); base de datos de instituciones prestadoras de salud (IPS) del Ministerio de Salud (2019); cartografía base del IGAC (2019).

3) Google Earth Engine: plataforma web propiedad de Google, diseñada para llevar a cabo análisis científicos y visualización de conjuntos de datos geoespaciales. Pueden consultarse a través de un API (interfaz de programación de aplicaciones) en lenguajes Javascript y Python (Google, 2020). Descargas realizadas entre el 10 y el 20 de diciembre de 2019.

Las variables explicativas que aporta el Censo Nacional de Población y Vivienda 2018, a nivel de unidad de cobertura son: total de hogares (hogares), total de unidades de vivienda (TOTAL_U_VI), número de personas omitidas por subenumeración de hogares (S), número de personas omitidas en viviendas ocupadas con todas las personas ausentes (A), número de personas omitidas en zonas no visitadas o incompletas (Z).

Las variables explicativas que aportan los registros administrativos son: densidad de construcciones normalizada (DEN_CONSTR), densidad de instituciones educativas normalizada (DEN_COLEG), densidad de instituciones prestadoras de servicios de salud normalizada (DEN_IPS), y distancia a centros poblados (DIST_CP).

[2] Información bajo reserva estadística. Toda la información recolectada en para censos y encuestas de los procesos estadísticos del DANE está protegida por la Ley 79 de 1993 o Ley de Reserva Estadística. Los datos suministrados al DANE a través de censos, encuestas u operaciones estadísticas "no podrán darse a conocer al público ni a las entidades u organismos oficiales, ni a las autoridades públicas, sino únicamente en resúmenes numéricos"

Las variables que aporta Google Earth Engine son: Luces nocturnas (LUCES_NOCT), elevación en metros (ELEVACION), índice de vegetación normalizado (NDVI), pendiente en grados (PENDIENTE), porcentaje de superficie de agua (PORCENT_AG).

En la siguiente tabla se listan los covariados espaciales utilizados, su definición y modo de cálculo.

Tabla 10. Covariados espaciales

Variable	**Definición**
Densidad de construcciones normalizada (DEN_CONSTR)	Densidad por hectárea de puntos de construcciones, provenientes de la capa CONSTRUCCION_P de la cartografía base 1:100.000 del IGAC (actualizada a 2019).
Densidad de instituciones educativas normalizada (DEN_COLEG)	Densidad por hectárea de puntos de instituciones educativas, provenientes de la base de datos del Ministerio de Educación Nacional (2019).
Densidad de instituciones prestadoras de servicios de salud normalizada (DEN_IPS)	Densidad por hectárea de puntos de instituciones prestadoras de salud (IPS), provenientes de la base de datos del Ministerio de Salud (2019).
Distancia a centros poblados (DIST_CP)	Distancia euclidiana (medida en línea recta) desde el centroide de cada celda de la grilla hasta el centro poblado o cabecera municipal más cercana.
Luces nocturnas (LUCES_NOCT)	Reflectancia promedio en cada celda de la grilla, obtenida a partir de un raster de 470 metros de resolución por píxel de imágenes VIIRS DNB, que a su vez es resultado del promedio de imágenes mensuales del año 2018.
Elevación en metros (ELEVACION)	Altura sobre el nivel del mar promedio en cada celda de la grilla, se obtiene a partir del DEM (modelo digital de elevación) SRTM con resolución de 30 metros/píxel.
Índice de vegetación normalizado (NDVI)	Estimación de la intensidad de radiación de ciertas bandas del espectro electromagnético que la vegetación emite o refleja. Se calculó el índice

	promedio en cada celda de la grilla, a partir de las bandas B4 y B5 (rojo e infrarrojo cercano) de imágenes del sensor LANDSAT 8 del año 2018, con resolución de 30 metros/píxel, resultado a su vez del promedio de imágenes del año 2018. Esta medida indica la cantidad, calidad y desarrollo de la vegetación en la superficie de la grilla. El índice adquiere valores entre -1,0 y 1,0
Pendiente en grados (PENDIENTE),	Pendiente o inclinación del terreno en grados promedio en cada celda de la grilla, calculado a partir del DEM (modelo digital de elevación) SRTM, con resolución de 30 m/píxel.
Porcentaje de superficie de agua (PORCENT_AG)	Porcentaje de superficies de agua con respecto al área total de la celda de la grilla (1 km2), obtenido a partir del conteo de los píxeles correspondientes a agua y multiplicando este valor por 900 m2 (área del píxel de 30 m x30 m).

Conjunto de predicción

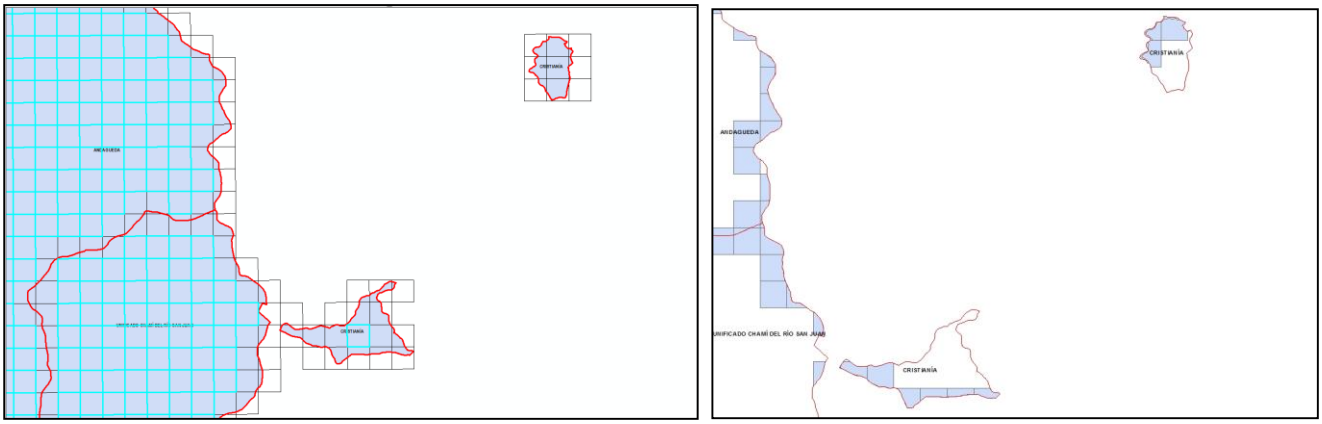
El conjunto de datos de predicción se encuentra a nivel de grilla de 1km x 1km. Estas dimensiones son una adaptación de la propuesta original de (Stevens, 2015) donde se utiliza una grilla de 100 m. por 100 m, más fina y precisa, pero más demandante en cuanto a capacidad de procesamiento y alamacenamiento, lo cual excede las posibilidades de cómputo del DANE al momento de la implementación de este ejercicio.

La grilla cuenta con un conteo de personas por celda, que corresponde a la población censada en las unidades censales georreferenciadas dentro del área de cada celda de grilla. Este conjunto de datos se tiene para todo el país, por ello, es posible calcular el peso de cada celda con respecto a la unidad de cobertura donde se encuentra contenida. A continuación, las variables del CNPV 2018 son distribuidas según el peso de la celda.

Las variables geoespaciales, por su parte, son estadísticas resumen de toda el área de la celda, ubicada sobre su centroide. De esta manera, se garantiza que el conjunto de predicción cuente con las mismos

covariados que el conjunto de aprendizaje. Sin embargo, dada la capacidad de cómputo disponible para realizar el cálculo de los covariados espaciales al nivel de celda de grilla, solo se cuenta con información de dichos covariados para las celdas y trozos de celda que se traslapan con los polígonos de los resguardos indígenas. Un ejemplo para dichas celdas se ve en las gráfica 11.

Gráfica 11. Ejemplo de celdas y trozos de celda



Los polígonos de la imagen encerrados por una línea roja son límites de resguardos. Como puede observarse, algunas celdas se encuentran completamente contenidas dentro del resguardo (imagen izquierda), y algunas celdas solo tienen una parte de ellas al interior del resguardo (imagen derecha). Por ello se cuenta con valores calculados de los covariados espaciales para aquellas celdas o trozos de celda que estén completamente contenidas dentro de los límites de los resguardos.

Estimación

Se aplican dos modelos de aprendizaje supervisado, Random Forest Regression y Gradient-boosted tree regression (GBTR). También se genera una estimación mediante distribución de la población ajustada ponderando por la población georreferenciada en cada una de las celdas.

Para determinar la configuración optima de los hiperparámetros en los dos primeros modelos aplicados, no fue posible realizar una validación cruzada con base en criterios estadísticos, relacionados con la minimización del error cuadrático medio, bajo distintas configuraciones del hiperparámetro, debido a

que, en el momento de realizar el ejercicio, no se contaba con la capacidad de cómputo necesaria para dicho proceso.

Por lo tanto, se realizó una asignación empírica de los hiperparámetros, que combinó la reducción de la raíz del error cuadrático medio, el aumento del coeficiente de determinación y la capacidad de procesamiento disponible.

Con el modelo de Random Forest regression**,** descrito en (Breiman, 2001) se realiza cierta cantidad de regresiones ajustada por dos hiperparámetros: el número de árboles y la profundidad del modelo. Cada regresión arroja una predicción, y el resultado final del modelo corresponde al promedio de dichas predicciones. En este caso, se usaron 45 árboles y una profundidad de 15. El tiempo de ejecución fue de 9 minutos y 53 segundos.

Por su parte, en el modelo de GBTR descrito en (Friedman, 2002) el resultado final es la predicción generada en la iteración con menor error cuadrático medio, producto de la combinación de las predicciones realizadas en las iteraciones previas. Los hiperparámetros que utiliza son el número máximo de iteraciones, y la profundidad. Para la presente estimación, se definió un máximo de 40 iteraciones y una profundidad de 10. El tiempo de ejecución fue de 43 minutos y 53 segundos.

Transformaciones de la variable dependiente

La variable dependiente es el conteo de personas por celda de la grilla. La variable respuesta en la propuesta de (Stevens, 2015) es el logaritmo natural de la densidad de población, es decir una transformación de la variable población, con la cual se calcula a posteriori el conteo de personas, usando los resultados del censo al nivel de desagregación más pequeño disponible. Sin embargo, ante el pobre desempeño de dicha transformación en las primeras adaptaciones de la metodología, se emplea como alternativa el logaritmo natural del conteo de personas. Para las variables explicativas de los modelos, se cuenta con información geoespacial y oficial disponible al nivel de desagregación de grilla.

Con la configuración de hiperparámetros asignada, el comportamiento de las métricas de desempeño de los modelos se muestra en la tabla11.

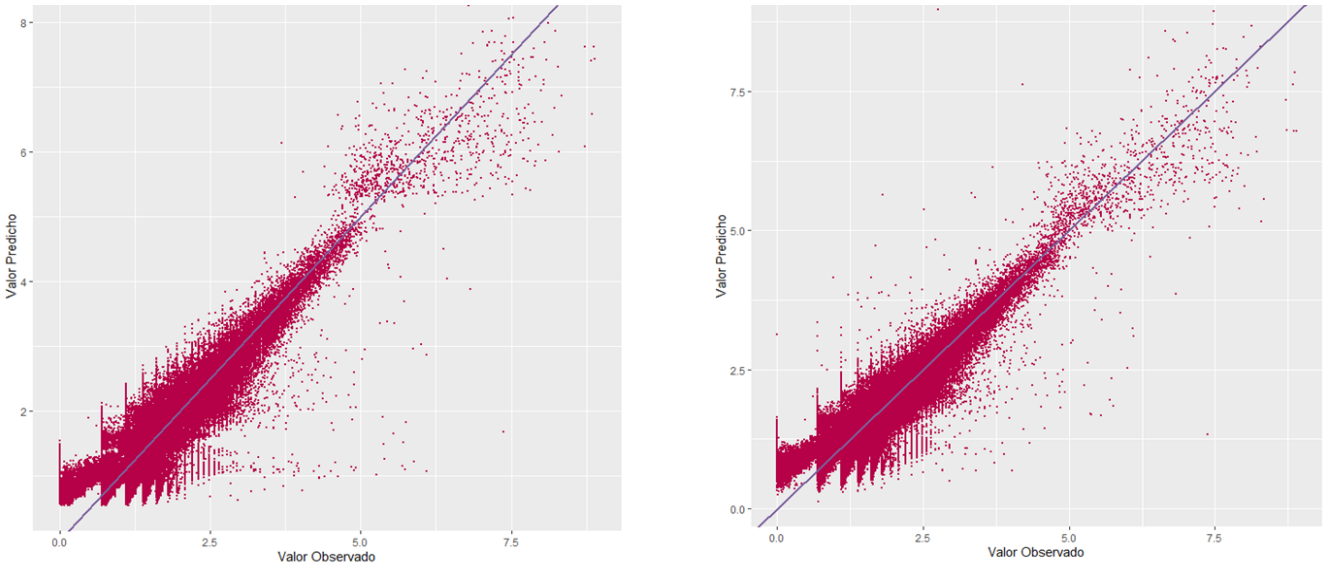
Tabla 11. ***Diagnóstico de los modelos***

Diagnóstico		
	Random Forest	GBTR
RMSE	0,457891	0,453902
R^2	0,704725	0,709847

Donde RMSE es la raíz cuadrada del error cuadrático medio que representa el promedio de las distancias al cuadrado entre los valores estimados por los modelos y los observados, y R^2 que representa el coeficiente de determinación, interpretado como el porcentaje de variabilidad de los valores observados explicado por las predicciones de los modelos.

Aunque las diferencias en las métricas entre los dos modelos se presentan después de la tercera cifra decimal, el modelo GBTR presenta un mejor desempeño que el Random Forest, ya que cuenta con un menor error cuadrático medio y un mayor coeficiente de determinación. Las gráfica 12 evidencian el comportamiento de las predicciones respecto a lo observado.

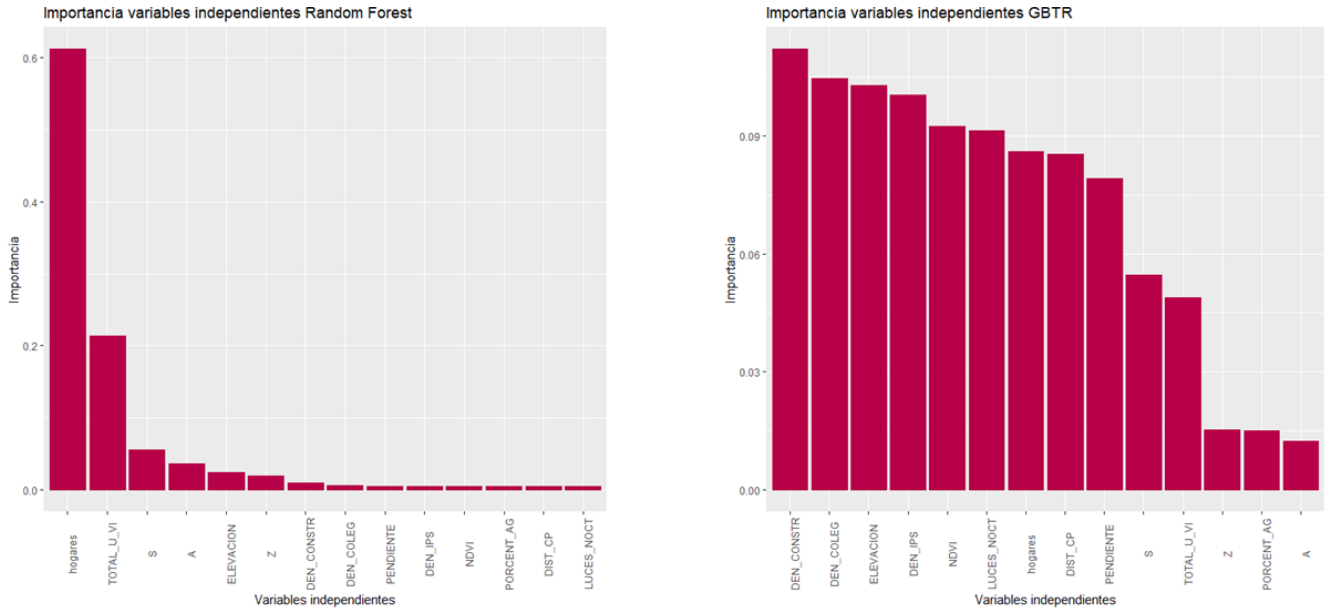
Gráfica 12. Gráficas modelos Random Forest y GBTR



Se observan comportamientos similares en ambos modelos, sin embrago, en el Random Forest pueden observarse dos grupos de resultados, aquellos con logaritmo natural mayor a 5 y los menores a 5, y refleja una mayor variabilidad que las predicciones del GBTR. Al observar la contribución de cada una

de las variables a los modelos, se corrobora el mejor desempeño del modelo GBTR, al contar con un número mayor de variables que contribuyen de modo significativo a las predicciones del modelo, especialmente aquellas de naturaleza geoespacial. En contraste, RandomForest explica el 65% del coeficiente de determinación con la variable censal de número de hogares. La gráfica 13 muestra el aporte de cada una de las variables a los modelos:

Gráfica 13. Aporte de las variables a los modelos



Aun contando con información limitada, los modelos aplicados y su desempeño son satisfactorios, y la distribución de los datos describe un comportamiento de ojiva similar al alcanzado en (Stevens, 2015), con los conteos originales. Las gráfica 14 muestran dicho comportamiento para el conjunto de prueba.

Gráfica 14. Comportamiento de los modelos de Random Forest y GBTR

4.2 Resultados.

4.2.1. Resultados de las estimaciones realizadas en resguardos indígenas

Dadas las limitaciones en los insumos utilizados, al no contar con los valores de los covariados en el total de grillas del área resto municipal, se emplean mecanismos alternativos para determinar el ponderador de cada grilla y trozo de grilla:

Alternativa 1: En cada una de las Áreas Operativas se calcula un conjunto de pesos para las grillas que pertenecen a dicha área, esto es:

$$\boldsymbol{P_l}^{*} = \begin{cases} \boldsymbol{P_l}^{\boldsymbol{m}}; \ \boldsymbol{l} \in \boldsymbol{R} \\ \boldsymbol{P_l}^{\boldsymbol{o}}; \ \boldsymbol{l} \notin \boldsymbol{R} \end{cases}$$

$$\boldsymbol{w_l} = \frac{\boldsymbol{P_l}^{*}}{\sum_{\boldsymbol{l}} \boldsymbol{P_l}^{*}}$$

Donde

- $P_l{}^m$ es la **población predicha por el modelo** para la *l-ésima* grilla en cada Área Operativa.

- $P_l{}^o$ es la **población observada** para la *l-ésima* grilla en cada Área Operativa.
- R es el conjunto de **subíndices que pertenecen a un resguardo** en cada Área Operativa.
- l es el **subíndice** para cada una de las grillas.

Alternativa 2: Es similar al caso anterior, pero se tiene que $\boldsymbol{P_l{}^*}$, viene dado por:

$$\boldsymbol{P_l{}^* = \begin{cases} P_l{}^o + (P_l{}^m - P_l{}^o) \cdot \dfrac{P_l{}^m}{P_l{}^o + P_l{}^m}; & l \in R \\ P_l{}^o \quad ; & l \notin R \end{cases}}$$

De esta forma,

- Si $P_l{}^m > P_l{}^o$ entonces $P_l{}^* > P_l{}^o$.
- Si $P_l{}^m < P_l{}^o$ entonces $P_l{}^* < P_l{}^o$.

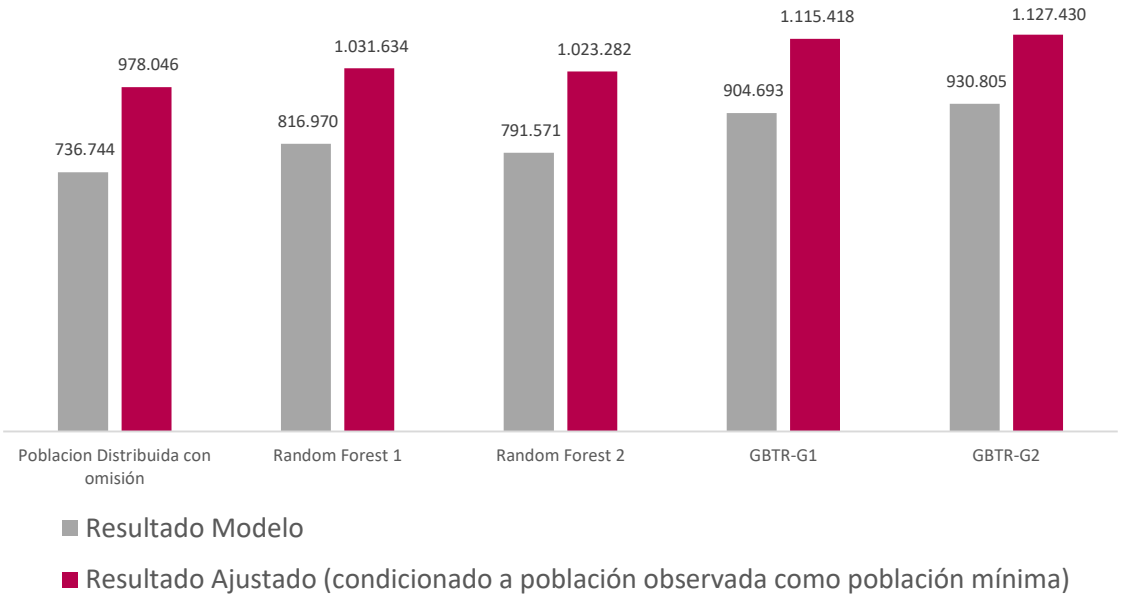
Como resultado de las dos alternativas descritas para determinar el ponderador de cada grilla y trozo de grilla, se generaron dos escenarios distintos para cada modelo. Por esta razón, cada modelo arroja dos alternativas de resultado diferentes.

Finalmente, se construyó un quinto escenario de estimación de la omisión en los resguardos, mediante un procedimiento simple: distribuir la omisión de cada unidad de cobertura en las celdas contenidas en ella, de manera proporcional a la población censada y georreferenciada en cada celda, y luego, sumar las poblaciones ajustadas por omisión de las celdas que se intersecan con cada resguardo para obtener la estimación del resguardo. Este escenario distribuye espacialmente la población omitida de manera proporcional a la distribución de la población censada.

El supuesto subyacente al procedimiento anterior es consistente con la omisión generada tanto por sub-enumeración de personas como por las de viviendas ocupadas con personas ausentes; sin embargo, puede presentar sesgos importantes que subestiman la omisión en zonas no visitadas o incompletas.

En la gráfica 15, se presentan los resultados de las cinco estimaciones generadas: dos por el modelo de Random Forest, dos por el modelo de GBTR, y uno por la distribución de la omisión proporcional a la población censada.

Gráfica 15. Resultados de los modelos y resultado ajustado

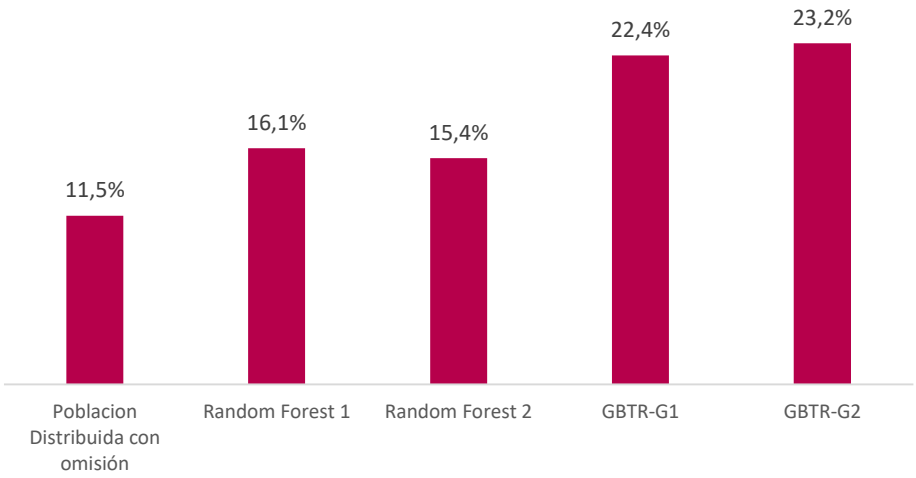


En la gráfica anterior, las barras grises corresponden al resultado "crudo" de la estimación, mientras que las barras rojas reflejan el resultado ajustado, es decir, condicionado a que en ningún caso la población ajustada por omisión puede ser inferior a la población censada. La diferencia entre uno y otro resultado ocurre porque, en algunos resguardos, los modelos predicen menos población de la que efectivamente se censó. El resultado ajustado de los modelos asigna a estos casos el nivel de la población censada, asumiendo un escenario de cero omisiones.

Este resultado solo tiene en cuenta resguardos del marco censal con polígono de límites definido. No incluye resguardos coloniales, para los cuales hubo un tratamiento separado que será descrito más adelante. La omisión poblacional total del CNPV 2018 en resguardos con polígono definido, para cada

una de las estimaciones realizadas, condicionadas a que la población censada fuera la población mínima estimada, se muestra en la gráfica 16.

Gráfica 16. Resultado de los modelos



4.2.1.1 Escogencia final del modelo.

Contando con el resultado de los cinco modelos de estimación de la omisión en resguardos indígenas, se realizó una evaluación empírica de sus resultados, comparándolos e identificando su nivel de consistencia con información de referencia proveniente de fuentes disponibles. Estos referentes son:

- Población censada en el CNPV 2018
- Población certificada en el resguardo en el año 2018
- Población estimada por la cartografía social que sirvió de insumo para el diseño operativo del CNPV 2018 en estos territorios
- Autocensos indígenas disponibles del año 2016, entregados por el Ministerio del Interior al DANE.

Las comparaciones se hicieron para cada resguardo, y en los casos en que el resguardo tuviera jurisdicción en más de un municipio, para cada fragmento de resguardo por municipio. El resultado de

todos los modelos en cada fragmento se comparó con los referentes anteriores, obteniendo 4 posibles resultados.

i) Similar al referente: cuando el valor de la estimación se encuentra entre el 80% y el 120% del valor del referente.
ii) Inferior al referente: cuando el valor de la estimación es menor que el 80% del referente.
iii) Superior al referente: cuando el valor de la estimación es mayor al 120% del referente
iv) Sin referente: no existe el referente disponible para comparar.

La evaluación pretendía identificar en qué medida el resultado de cada modelo era consistente con los referentes disponibles, ya fuera porque confluía con ellos o porque difería de un modo esperable según los sesgos de estos últimos.

De las combinaciones resultantes de comparar el modelo con los referentes, se identificaron 40 temáticamente consistentes. Esta selección de combinaciones tuvo en cuenta los sesgos propios de dichas fuentes, particularmente las de la cartografía social y los autocensos indígenas que tienden a sobreestimar la población de estos territorios en comparación con el CNPV 2018, teniendo en cuenta que este último utiliza el concepto de residente habitual para censar a la población: la cartografía social por su carácter declarativo, ya que dependía de la memoria y percepción del líder indígena que informaba el dato poblacional, y los autocensos por la inclusión de personas que no son residentes habituales de los resguardos. En la tabla 12 se listan cuáles combinaciones se consideraron más consistentes temáticamente.

Tabla 12. Combinaciones entre resultados del modelo de cálculo de omisión en resguardos, y referentes de población en resguardos disponibles, consideradas consistentes.

Referentes de Evaluación disponible			
CNPV 2018	**Certificación 2018**	**Cartografía Social**	**Autocensos indígenas**
Similar	Similar	Similar	Superior
Similar	Superior	Inferior	Superior

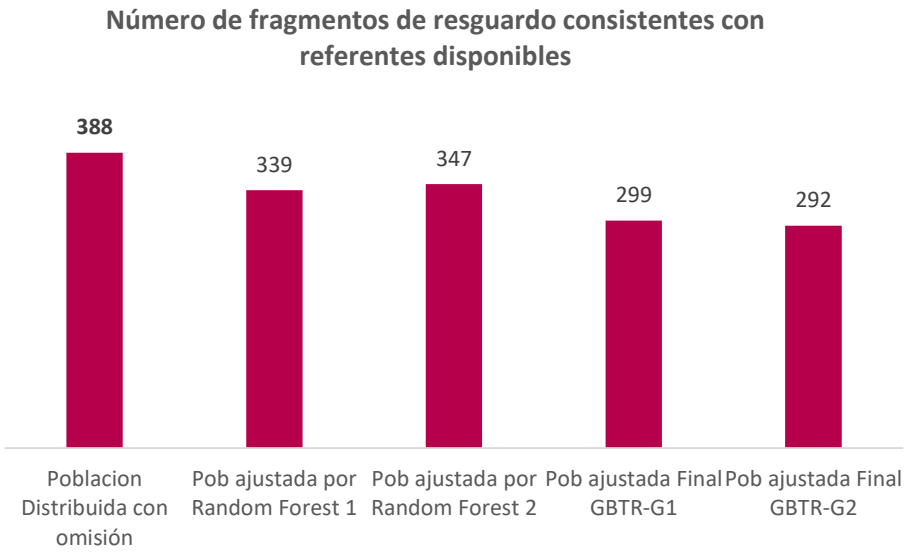
Referentes de Evaluación disponible			
CNPV 2018	**Certificación 2018**	**Cartografía Social**	**Autocensos indígenas**
Similar	Inferior	Inferior	Superior
Similar	Inferior	Superior	Sin Referente
Similar	Similar	Superior	Superior
Similar	Similar	Superior	Sin Referente
Similar	Similar	Inferior	Sin Referente
Similar	Similar	Superior	Inferior
Similar	Inferior	Inferior	Sin Referente
Superior	Inferior	Similar	Sin Referente
Superior	Inferior	Inferior	Sin Referente
Similar	Similar	Sin Referente	Sin Referente
Similar	Similar	Similar	Sin Referente
Similar	Superior	Inferior	Sin Referente
Similar	Inferior	Similar	Superior
Similar	Similar	Sin Referente	Inferior
Similar	Sin Referente	Inferior	Sin Referente
Superior	Superior	Inferior	Sin Referente
Superior	Superior	Inferior	Superior
Similar	Similar	Superior	Similar
Superior	Superior	Similar	Sin Referente

Referentes de Evaluación disponible			
CNPV 2018	**Certificación 2018**	**Cartografía Social**	**Autocensos indígenas**
Similar	Similar	Similar	Similar
Similar	Superior	Inferior	Inferior
Similar	Inferior	Similar	Sin Referente
Similar	Similar	Similar	Inferior
Similar	Similar	Sin Referente	Similar
Similar	Inferior	Inferior	Similar
Similar	Similar	Inferior	Similar
Similar	Similar	Inferior	Superior
Similar	Inferior	Superior	Similar
Similar	Inferior	Superior	Inferior
Similar	Similar	Inferior	Inferior
Similar	Inferior	Inferior	Inferior
Superior	Inferior	Similar	Similar
Similar	Inferior	Sin Referente	Similar
Superior	Inferior	Inferior	Inferior
Similar	Inferior	Similar	Similar
Superior	Superior	Similar	Similar
Similar	Inferior	Similar	Inferior
Sin Referente	Inferior	Inferior	Sin Referente

La evaluación final de los modelos consistió en obtener el número de resguardos (para resguardos en un solo municipio) o fragmentos de resguardo (para resguardos con jurisdicción en más de un municipio) consistentes con los referentes disponibles para cada modelo de cálculo de omisión. El que obtuviera más fragmentos de resguardo consistentes se consideró como el modelo con mejor desempeño.

En este caso, la estimación con mayor consistencia con los referentes disponibles fue la distribución proporcional de la omisión por población censada. Los modelos de Random Forest y GBTR, tienden a sobreestimar más la población indígena en los resguardos, si se les compara con los referentes disponibles.

Gráfica 17 Evaluación de resultados contra referentes disponibles



Sin embargo, debe tenerse en cuenta que el supuesto de distribución de la omisión proporcional a la población censada, no se cumple cuando la omisión de un resguardo se debió a problemas de cobertura, es decir, donde hubo zonas sin visitar. Este sesgo de la estimación se corroboró en la poca consistencia temática que presenta su resultado en resguardos o fragmentos de resguardo con población censada igual a cero o de un solo dígito, es decir, en aquellos donde el efecto de la no cobertura es mayor.

En cambio, en estos fragmentos de resguardo los resultados del modelo GBTR 2 son los que presentan mayor consistencia temática, derivado del peso explicativo que les confiera a las variables geoespaciales para predecir la población en estos territorios, sin depender del resultado censal.

Teniendo en cuenta estas observaciones, se decidió adoptar la población ajustada por distribución de la omisión proporcional a la omisión censada, en la mayoría de resguardos y/o fragmentos de resguardo del país, y aplicar los resultados del modelo GBTR 2 para los casos en que la población censada estuviera entre cero y nueve, es decir, población censada de un solo dígito.

Resguardos coloniales

Para aplicar cualquiera de los 5 métodos antes descritos, es indispensable contar con los límites de cada resguardo. Dichos límites están dados por la cartografía oficial, aportada por la entidad competente, que para el caso de los resguardos indígenas es la Agencia Nacional de Tierras y el IGAC.

En el caso de los resguardos coloniales y republicanos (52 resguardos), no existe una cartografía oficial, por tratarse de territorios constituidos con anterioridad a la creación de la moderna institucionalidad agraria. Algunos de ellos están soportados por cédulas reales de origen colonial. Son resguardos en espera de surtir el proceso de clarificación estipulados en la normatividad (Ley 160 de 1994, Artículo 85; Decreto 1071 de 2015, título 7 capítulo 6), para que queden definidos sus límites exactos.

Al carecer de unos límites territoriales definidos e incorporados a la cartografía oficial, no había posibilidad de aplicar ninguna de las técnicas antes descritas para estimar la población ajustada por omisión. Con respecto a la realidad territorial de estos resguardos, solo se conoce su jurisdicción en el área resto municipal de algunos municipios.

Dadas la realidad *sui generis* que constituyen los resguardos de origen colonial y republicano, y reconociendo que a pesar de no contar con un territorio bien delimitado sus poblaciones sí pudieron tener omisión de población en el CNPV 2018, en estos resguardos se aplicó el porcentaje de omisión calculado en el área resto municipal donde se encuentra.

Estimación de la población indígena en resguardos

Los métodos hasta ahora descritos permiten estimar la población total ajustada por omisión en los resguardos indígenas. Sin embargo, para efectos de la certificación de población indígena en resguardos, y de las proyecciones de población indígena en estos territorios, resulta indispensable estimar qué parte de la población ajustada por omisión es indígena. No siempre la población censada en un resguardo se autorreconoce como indígena en su totalidad, suele ocurrir que, personas de otro

grupo étnico o sin pertenencia étnica, se han asentado en estos territorios, como resultado de uniones familiares interétnicas, o debido a otras realidades históricas.

El criterio general aplicado para diferenciar la población indígena de la no indígena, fue utilizar la proporción observada en la población censada de cada resguardo o fragmento de resguardo, y aplicarlo a la población ajustada por omisión. Al aplicar este procedimiento, se parte del supuesto que no se presentó un sesgo significativo por pertenencia étnica en la omisión censal de los resguardos, o lo que es igual, que en estos territorios la proporción de población indígena es similar en lo censado y en lo omitido. Este criterio se utilizó en el 95,4% de los casos, de acuerdo con la siguiente fórmula general

$$PI_a = PT_a * \frac{PI_c}{PT_c}$$

Donde:

PIa es la población indígena ajustada por omisión en el resguardo o fragmento

PTa es la población total ajustada por omisión en el resguardo o fragmento

PIc es la población indígena censada en el resguardo o fragmento

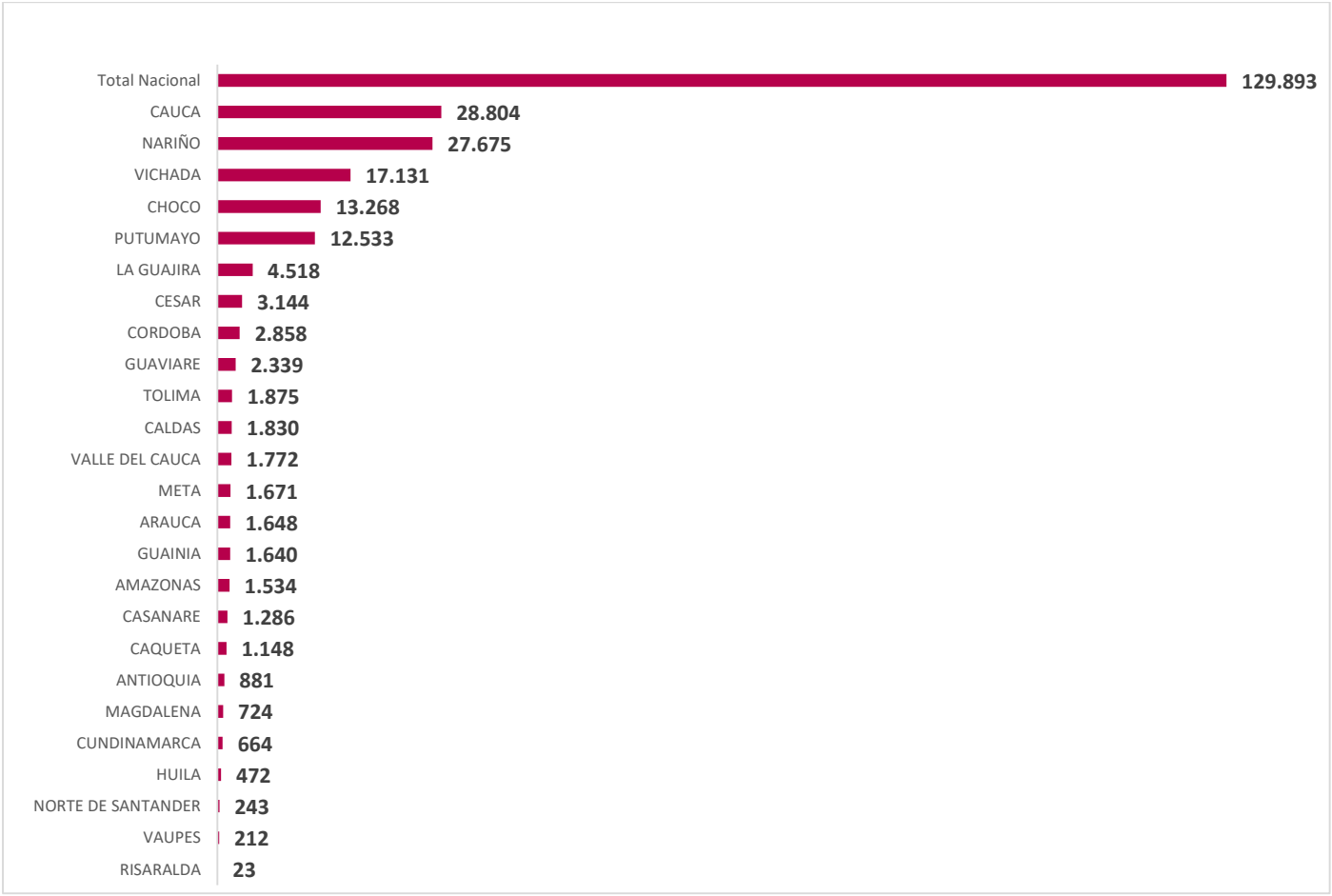
PTc es la población total censada en el resguardo o fragmento

Por otra parte, en 31 fragmentos de resguardo se identificó problemas en los resultados de la pregunta de autorreconocimiento étnico, dados por niveles de no respuesta muy altos. En estos casos, se aplicó la proporción observada en el conjunto de resguardos que conforman la región a la cual hace parte el resguardo, por considerarse más cercana a la realidad que la observada en el mismo resguardo.

Finalmente, se presentaron 14 resguardos con jurisdicción en un solo municipio donde hubo omisión censal total (la población censada fue cero), ninguno de los métodos muestra un ajuste superior a cero, y existe evidencia de presencia de población en el territorio como imágenes de satélite, cartografía social, autocensos, y/o certificaciones anteriores. En estos casos se tomó como población indígena ajustada la certificación de población indígena en resguardos de la vigencia 2018.

La gráfica 18, muestra que total de indígenas omitidos en resguardos a nivel nacional se estima en de 129.893 , siendo los departamentos de Cauca, Nariño, Vichada, Chocó y Putumayo los mayores receptos de dicha población.

Gráfica 18: Indígenas omitidos en resguardos



4.1.2 Metodología: Calculo de omisión NMAA en manzanas y secciones rurales

Adoptando los resultados de fase dos, se cuenta con una nueva participación de los grupos étnicos a niveles subnacionales. Esta información permite aplicar la metodología propuesta por (Stevens, 2015),

que consiste en aplicar un modelo de aprendizaje, usando las participaciones a dichos niveles, para luego predecir sobre niveles inferiores a los censales.

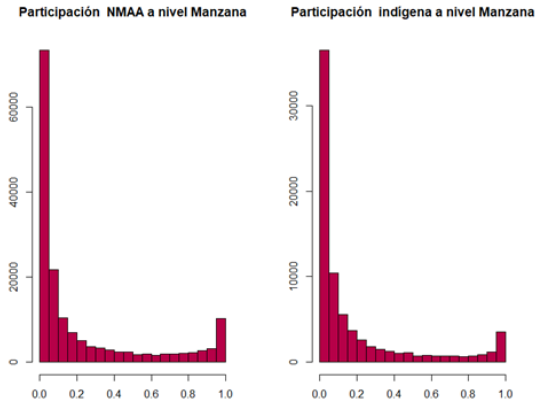
Unas de las primeras preguntas que surgen luego de la caída en la participación de negros, mulatos, afrodescendientes y afrocolombianos (NMAA) son:

- ¿Cómo lo omisión censal afecto la participación afro a nivel nacional?
- ¿Todas las personas omitidas en el censo pertenecen a mismo grupo étnico?
- ¿La omisión censal está concentrada en zonas específicas, o tiene un comportamiento aleatorio?

Gracias a que los datos del CNPV 2018 se encuentran georreferenciados, es posible determinar el nivel de participación étnico en cualquier nivel de desagregación manejado por la entidad (municipio, sector, sección o manzana). La visualización de dichas participaciones permite dar una idea de los patrones de asentamiento de los grupos étnicos, además, de caracterizar los grupos étnicos demográficamente.

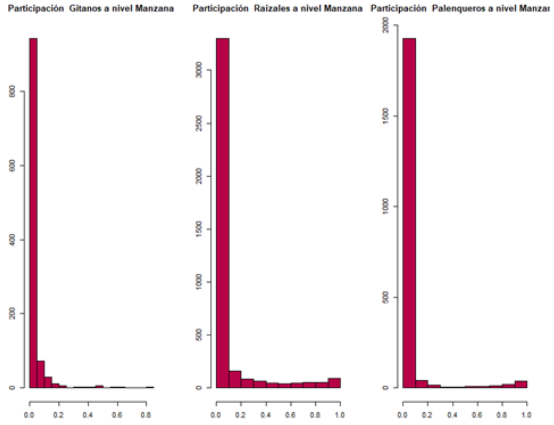
La participación obtenida a nivel de manzana de todos los grupos étnicos, muestra unos patrones distribucionales particulares, que corroboran en el caso indígena y NMAA, la existencia de territorios en los que estos grupos étnicos predominan, la gráfica 19 muestra como las manzanas o secciones rurales con baja prevalencia y las manzanas o secciones rurales con altas prevalencias son las más numerosas:

Gráfica 19. Participación de grupos étnicos a nivel de manzana



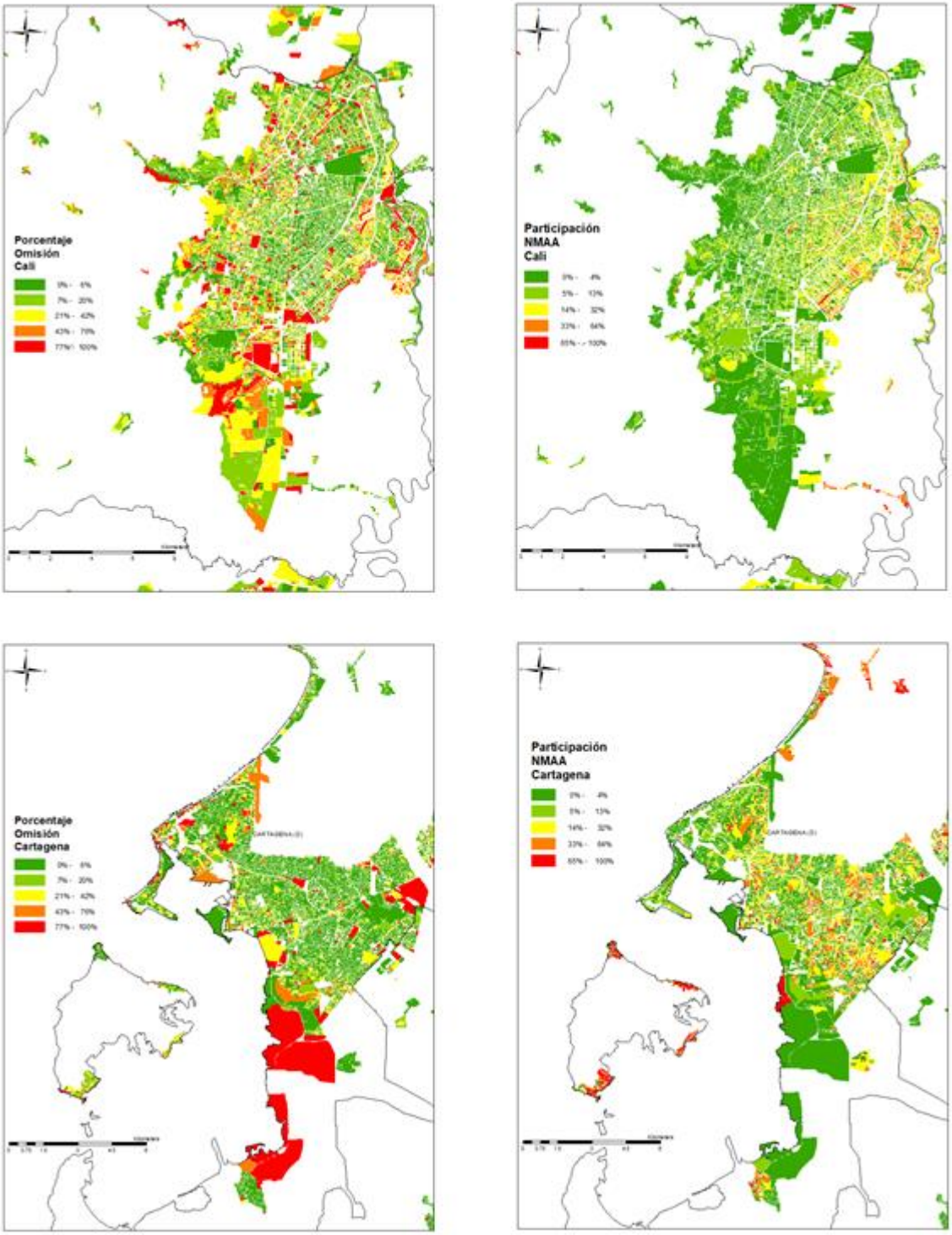
Las etnias gitanas, raizales y palanqueros no presentan estos mismos patrones de asentamiento, como se puede observar en la gráfica 20.

Gráfica 20. Distribución de la participación de Gitanos, Raizales y Palenqueros



Adicionalmente, se cuenta con el porcentaje de personas omitidas en todas las manzanas y secciones rurales del MGN, con los cuales se elaboraron mapas comparativos de participación NMAA y omisión para las ciudades de Cali y Cartagena.

Ilustración 3. Mapas de porcentaje de omisión y participación de población NMAA para Cali y Cartagena



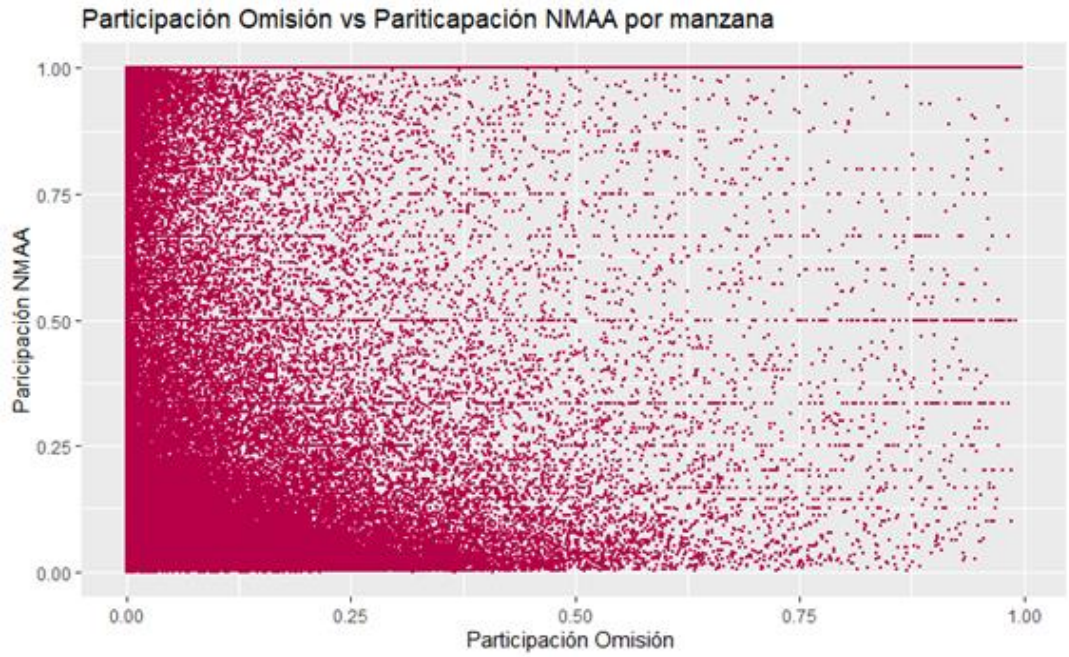
Como se evidencia en la ilustración 3, en la ciudad de Cali, la omisión censal está distribuida en toda cabecera municipal al parecer de forma aleatoria[3], lo que no ocurre con el asentamiento de las personas que se autorreconocen como NMAA, ya que predominantemente se encuentran en el distrito de Agua Blanca. De forma similar en la ciudad de Cartagena, pero con una menor intensidad en el caso de la omisión respecto a Cali, la omisión censal se encuentra distribuida de forma aleatoria en toda la cabecera.

Estas dos visualizaciones permiten entender, que el patrón de omisión censal se traslapa con el patrón de participación NMAA en algunas áreas (Agua Blanca en Cali). No obstante, también se destaca, que, aunque existen zonas de alta omisión y alta participación, por lo tanto, no podría afirmase que la caída en la partición de NMAA es enteramente atribuible a la omisión censal.

La gráfica 21 muestra el comportamiento de la omisión en las **162.556** manzanas en las que por lo menos una persona se autorreconocen como NMAA en total nacional, dicha participación es calculada sobre la población efectivamente censada dentro de la manzana:

[3] Se verifica en el estudio postcensal del profesor Urrea "Análisis de la dinámica intercensal del autorreconocimiento en la población negra, afrocolombiana, raizal y palenquera en el periodo 2005-2018. (FASE I)"

Gráfica 21. Dispersión de la participación y la omisión NMAA por manzana



Se destaca que las zonas con altas omisiones y altas participaciones de NMAA son considerablemente menores a las zonas con altas concentraciones afro y bajas omisiones. Sin embargo, los patrones de asentamiento deben ser considerados con el fin de no distribuir el comportamiento de zonas con altas participaciones en zonas que temáticamente no cuentan prevalencias altas de estos grupos. Por lo tanto, se adopta la regionalización trabajada en FASE II, formando los siguientes grupos:

- Grupo 1: Pacífico, Caribe y Suroccidente.
- Grupo 2: Noroccidente, Bogotá, Centro Oriente y Centro Sur.
- Grupo 3: Amazonas y Orinoquia.

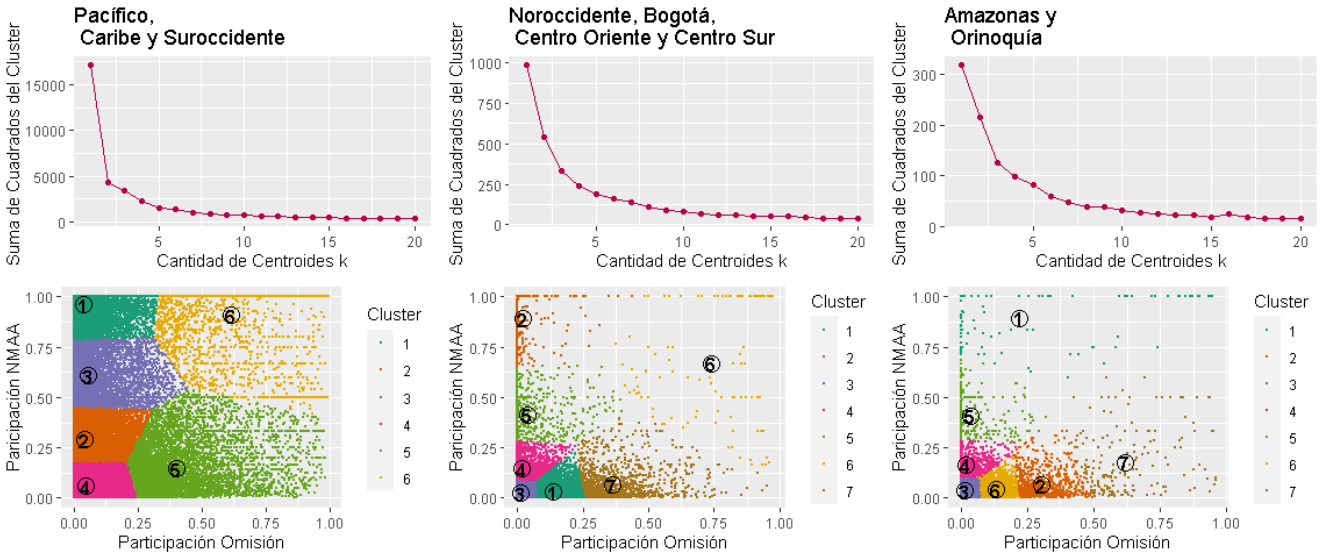
4.2.2 Resultados de las estimaciones realizadas para población NMAA

Adoptando los tres grupos de regiones propuestos, se determinan los conjuntos de aprendizaje, prueba y predicción mediante la implementación del algoritmo k-means[4], con el fin de usar las manzanas y secciones rurales con participaciones bajas de omisión como las mejores candidatas a ser incluidas en el conjunto de aprendizaje. Así mismo, las manzanas y secciones rurales con altas participaciones de omisión serán las mejores candidatas a ser incluidas en el conjunto de predicción. Empleando el método

[4] Ver (Hartigan, 1979)

del codo, se seleccionan el número óptimo de clúster para cada uno de los grupos de regiones, la gráfica 22 muestra el resultado de esta implementación.

Gráfica 22. Resultados estimaciones de población NMAA por región

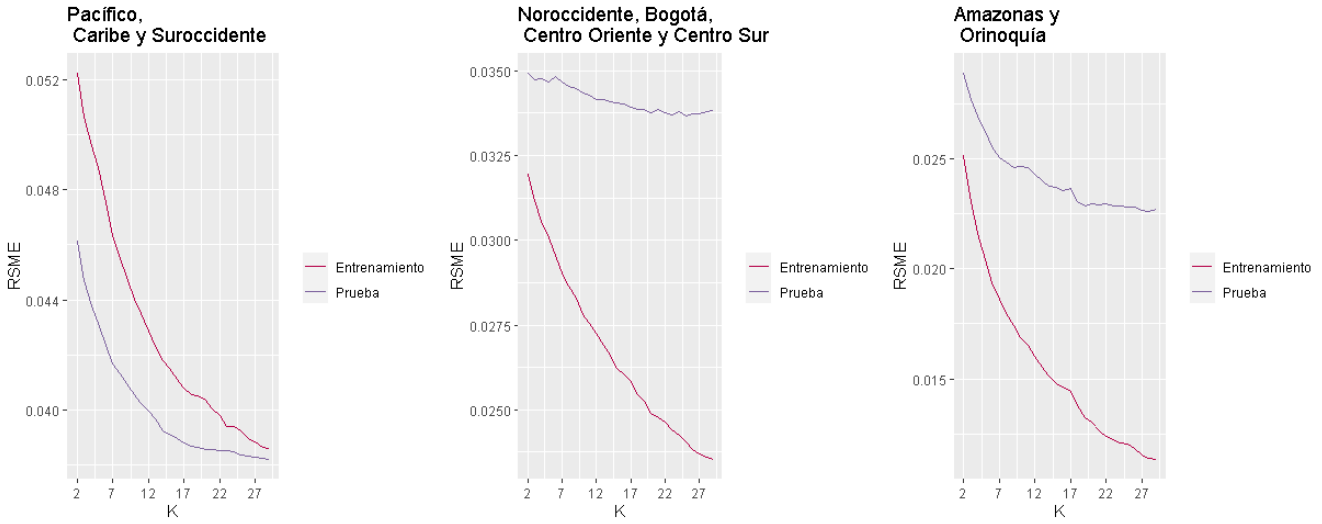


De esta forma, en el grupo de regiones 1, se obtiene un número óptimo de clúster de 6, donde los clústeres 1, 2, 3 y 4 son seleccionados como el conjunto de aprendizaje y los clústeres 5 y 6 como los conjuntos de predicción. En los grupos de regiones 2 y 3, se determina un número óptimo de clúster de 7, siendo los clústeres 1,2, 3, 4 y 5 el conjunto de aprendizaje y los clústeres 6 y 7 el conjunto de predicción para la región 1. Por último, en la región 3 los clústeres 3,4,5 y 6 como el conjunto de aprendizaje dejando como conjuntos de predicción los clústeres 1,2 y 7.

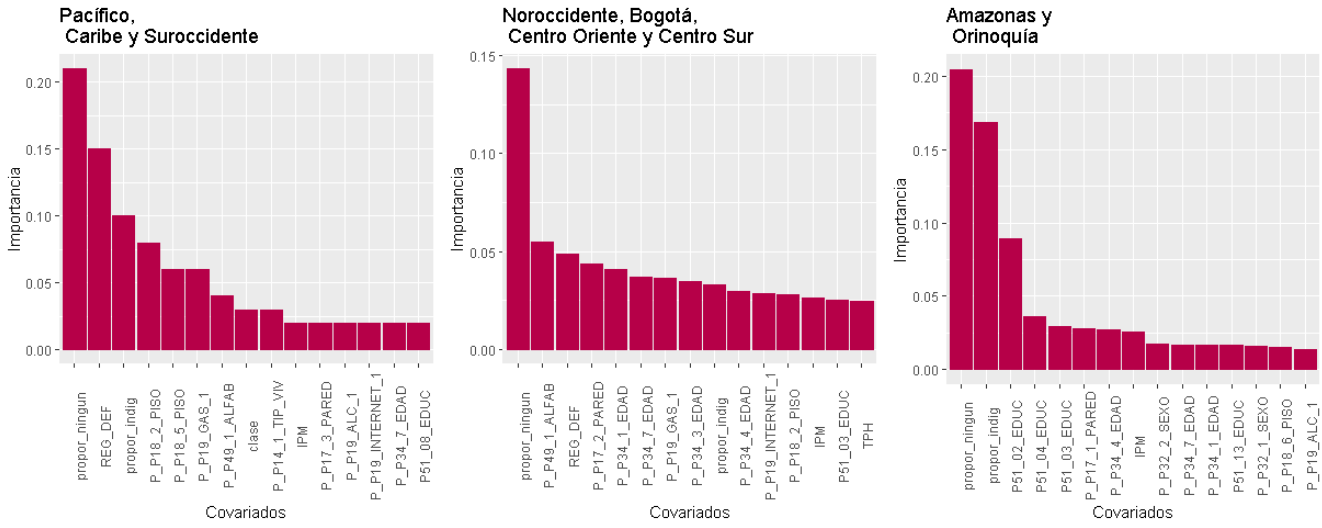
Una vez determinados los conjuntos de entrenamiento, de forma aleatoria se selecciona un conjunto de prueba, que servirá para diagnosticar la mejor combinación de hiperparámetros de un modelo de predicción GBTR, de esta manera se procede a determinar el hiperparámetro óptimo de profundidad, la gráfica 22 muestra el resultado de la validación cruzada para los grupos de regiones, de donde se determina una profundidad de 14, 30 y 20 para el grupo de regiones 1, 2 y 3 respectivamente. Seguido, se ejecuta el ajuste de los modelos para cada grupo de regiones con sus respectivos parámetros óptimos y se determina la importancia de los 58 covariados en cada modelo, seleccionado aquellos cuyo aporte supera el 1%, la gráfica 24 muestra el comportamiento de dichos covariados, destacando la participación de ningún grupo étnico (propor_ningun) con el mayor aporte en todos los modelos, así

como el índice de pobreza multidimensional censal (IPM[5]) de forma común en todos los grupos de regiones.

Gráfica 23. Comportamiento de los conjuntos de entrenamiento y prueba



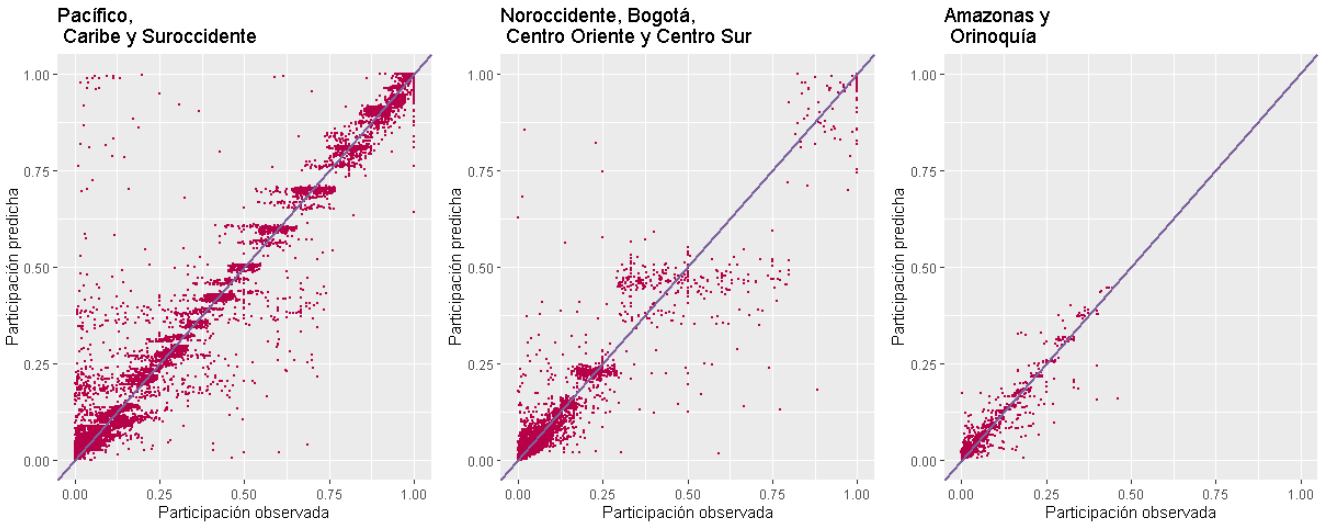
Gráfica 24. Covariados que superan el 1% en aporte por regiones agrupadas



[5] Ver [Medida de pobreza multidimensional municipal de fuente censal 2018]

Por último, la gráfica 25 muestra el comportamiento de los valores observados versus las predicciones sobre lo conjuntos de prueba, evidenciando el típico comportamiento de ojiva para los grupos de regiones 1 y 3. Para la región 3, este comportamiento es evidente para participaciones inferiores al 25%.

Gráfica 25. Dispersión de los valores observados respecto a las predicciones



Como métricas de desempeño de los modelos, se muestran la raíz del error cuadrático medio (siglas en ingles RSME) y el coeficiente de determinación ($\boldsymbol{R^2}$). La tabla 13 muestra los resultados para cada uno de los grupos de regiones sobre os conjuntos de prueba. Se resalta el modelo ajustado para el grupo de regiones 1, ya que explica el 98.8% de variabilidad de las participaciones NMAA. Así mismo, el modelo para el grupo de regiones 3, que presenta la menor medida de variabilidad RSME. En general, las métricas de desempeño son óptimas, ya que el coeficiente de determinación de los 3 modelos supera el 89%, y una variabilidad que no supera el 5%.

Tabla 13. RMSE y coeficiente de determinación para cada grupo de regiones

Grupo de Regiones	Nombre	RMSE	R^2
1	Pacífica, Atlántica y Suroccidente	0,0430	0,9879
2	Noroccidente, Bogotá y área de influencia, centro oriente, centro sur	0,0365	0,8942
3	Amazonía y Orinoquía	0,0237	0,9217

Dado que el diagnóstico de los modelos aplicados es aceptable, se aplican los ajustes a los conjuntos de predicción bajo las siguientes consideraciones:

1. Los resultados del modelo son condicionados a las participaciones observadas, es decir si el modelo predice una menor proporción de NMAA dentro de la manzana se toma la proporción observada.

2. Dado que el conjunto de aprendizaje tiene niveles bajos de omisión (**147.888** manzanas o sectores rurales), se considera que el comportamiento étnico dentro de las personas omitidas debe conservar la estructura observada de las personas efectivamente censadas, por lo tanto, se aplica la participación étnica observada a la población omitida.

3. Los modelos predicen la participación dentro de las manzana (**202.598** con por lo menos una persona omitida), por lo tanto el ajuste al volumen se realiza como:

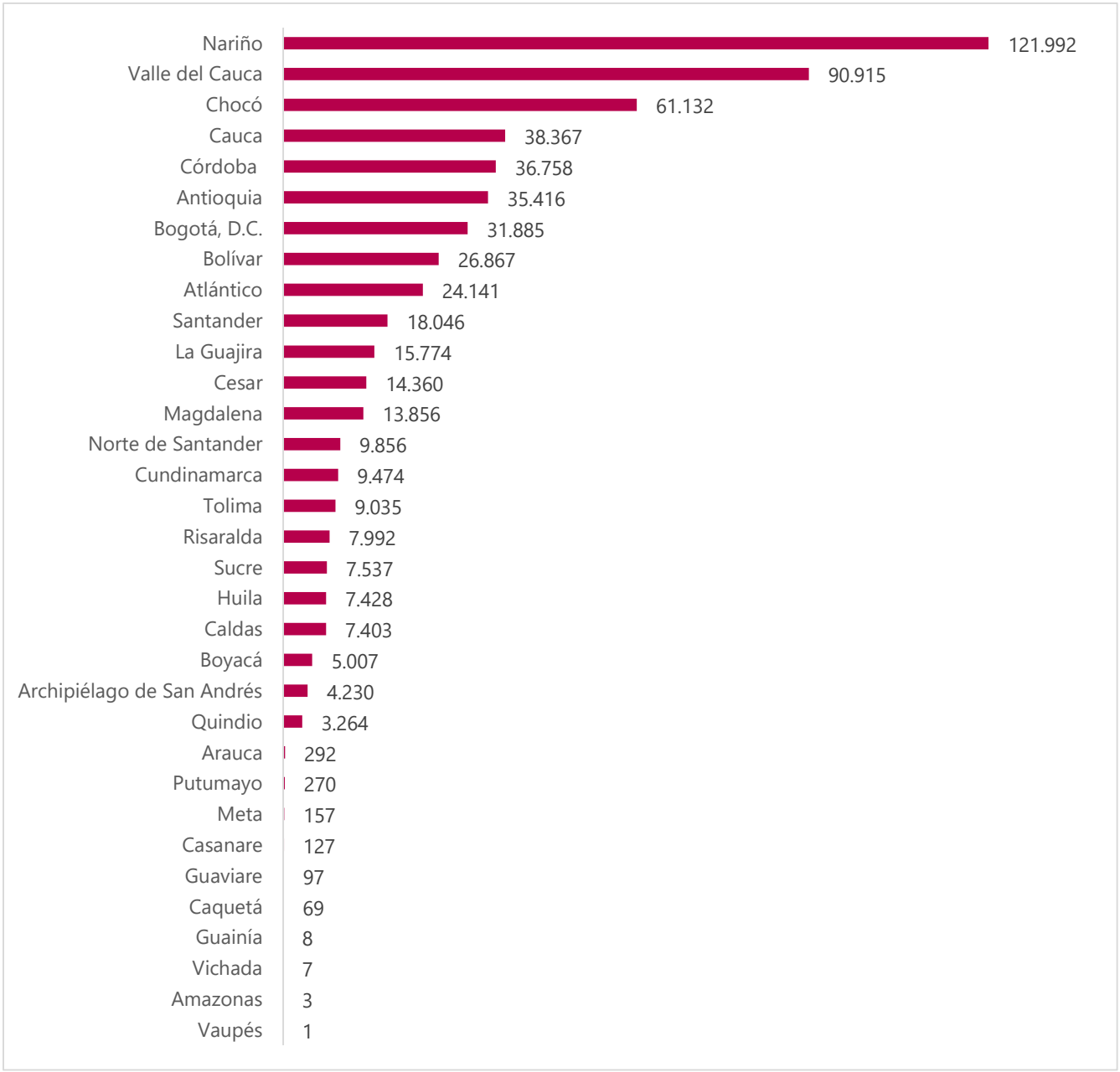
   $$P_{ajustada} = P_{observada} + P_{omitida} \times (\bar{p}_{NARP} - p_{NARP})$$

   Donde:
   p_{NMAA}:es la proporción de NMAA observada en la manzana
   $\bar{p}_{NMAA}$: es la predicción de la proporción de NMAA en manzana.
   $P_{(-)}$: Número de personas

4. Por último, en caso en que $\bar{p}_{NARP} < p_{NARP}$, el número de personas NMAA omitidas en la manzana es igual al número de personas omitidas.

La gráfica 26 muestra los valores estimados a nivel departamental, en donde el departamento que en donde se encuentran mayor población omitida NMAA es Nariño con 121.992 personas, seguido de Valle del Cauca con 90.915 y Chocó con 61.132. Resaltando que los departamentos pertenecientes a la Amazonía y Orinoquía tienen la menor estimación, sumando 1.031 personas omitidas NMAA.

Gráfica 26. Valores estimados de población NMAA omitida por departamento



5. Fase IV – Distribución proporcional de la población omitida no asignada en Fase II y III por pertenencia étnica

5.1. Marco teórico

En el contexto demográfico, es de interés obtener tasa cuyos numeradores y denominadores provienen de conteos básicos, desagregados en características de las personas como edad y sexo, sin embargo, las fuentes de información oficiales podrían contener distorsiones en dichos conteos producto inconvenientes en la recolección de información u omisión de personas en las operaciones estadísticas que se implementan alrededor del mundo. No obstante, cualquier discrepancia en dichos conteos, afectaría drásticamente el resultado de las tasas, por lo tanto, es indispensable implementar estrategias que permitan tratar los conteos y conseguir tasas que tengan sentido demográfico.

En (Arriaga, 1994) se describen las técnicas de Rele y Brass para ajustar las tasas específicas de fecundidad, mediante el uso de tablas cuadradas, ajustando el conteo de mujeres en edad fértil y niños menores a cinco años. Estas técnicas pueden ser extrapoladas al contexto étnico, ya que al contar con una el total de población ajustada por cobertura del CNPV 2018, la participación étnica a total nacional de la ECV 2018 y la desagregación geográfica departamental o municipal, es posible ajustar los conteos para cada una de las etnias dentro de la desagregación geográfica y por ende obtener participaciones étnicas razonables.

5.2. Metodología

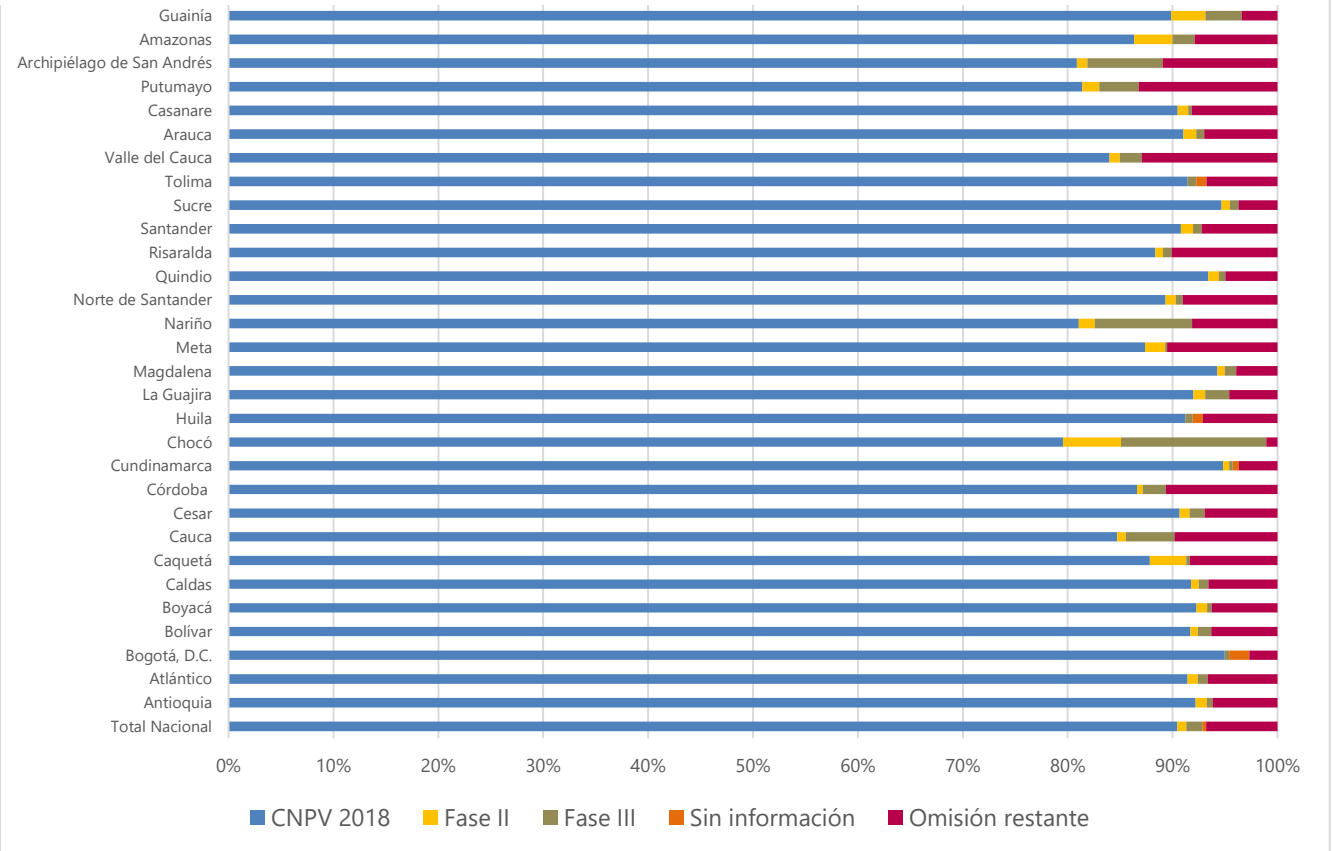
La propuesta metodológica de las fases II y III se concentran en mitigar la caída del autorrecnocimiento étnico, asignando de forma probabilística pertenencia étnica a las personas sin información y omitidas. Sin embargo, dichos procedimientos no logran asignar un grupo étnico a la totalidad objeto del estudio, la tabla 14 muestra los resultados a total nacional de las personas efectivamente censadas por grupo étnico y los resultados adoptando los ajustes propuestos. El grupo NMAA recibe 649.135 personas siendo el mayor receptor de población, llegando a 3.599.207; seguido de ningún grupo étnico que aumenta 518.255 personas, e indígenas, a los cuales se suman 130.987 personas. Las etnias Rrom, Raizal y Palenquero, no aumentan su población ya que, por sus bajas participaciones, no fueron incluidos en los procedimientos propuestos. Por último, se tienen 182.484 personas sin información étnica y 3.391.286 personas omitidas que no fueron asignadas en fase II y III.

Tabla 14: Resultados étnicos a nivel nacional con ajuste y sin ajuste

Ajuste	Indígena	Rrom	Raizal	Palenquero(a)	NMAA	Ningún	Sin información	Omisión
Sin	1.905.617	2.649	25.515	6.637	2.950.072	38.678.341	595.586	4.094.077
Con	2.036.604	2.649	25.515	6.637	3.599.207	39.196.596	182.484	3.391.286
Diferencia	**130.987**	**0**	**0**	**0**	**649.135**	**518.255**	**-413.102**	**-702.791**

La gráfica 27 muestra el porcentaje de población a nivel departamental que fue efectivamente censado, estimado en fase II y III, sin información de pertenencia étnica y omitido, destacando que la omisión restante a nivel departamental no supera el 15% en ningún departamento, siendo Putumayo, con 13% de omisión sin asignación étnica el de mayor valor y Chocó con el 1.1% el de menor valor. Por lo tanto, es plausible suponer, que el restante de omisión a nivel departamental tendría la misma estructura étnica obtenida de la suma de la población efectivamente censada y el ajuste propuesto.

Gráfica 27:Participación por tipo de ajuste a nivel departamental



El insumo a nivel departamental para el procedimiento propuesto se muestra en la tabla 15, en donde el objetivo principal, es distribuir el restante de omisión a nivel departamental, usando como condición columna las participaciones étnicas de la ECV 2018 a nivel nacional y la población ajustada por cobertura del CNPV 2018 como condición columna.

Tabla 15: Insumos tabla cuadrada

Departamento	Indígena	Rrom	Raizal	Palenquero(a)	NMAA	Ningún	Sin información	Restante Omisión CNPV 2018	P. BASE CNPV 2018
Participaciones étnicas ECV 2018	**4,34%**	**0,01%**	**0,04%**	**0,02%**	**9,28%**	**86,31%**			
Total Nacional	**2.036.604**	**2.649**	**25.515**	**6.637**	**3.599.207**	**39.196.596**	**182.484**	**3.391.286**	**48.258.494**
Antioquia	38.811	140	640	183	348.095	5.622.424	0	396.809	6.407.102
Atlántico	39.061	101	478	852	163.282	2.162.330	0	169.413	2.535.517
Bogotá, D.C.	19.131	603	1.060	218	97.541	7.094.869	142.564	199.144	7.412.566
Bolívar	5.305	31	573	3.988	342.289	1.584.174	0	133.750	2.070.110
Boyacá	7.409	18	62	16	9.176	1.124.181	0	76.514	1.217.376
Caldas	57.730	37	107	30	21.983	852.659	0	65.709	998.255
Caquetá	11.372	21	30	14	5.112	345.570	0	39.730	401.849
Cauca	338.442	39	93	86	287.295	684.503	0	154.030	1.464.488
Cesar	55.057	20	128	75	156.877	903.421	0	84.996	1.200.574
Córdoba	205.483	142	167	77	139.543	1.249.124	0	190.247	1.784.783
Cundinamarca	19.621	98	148	60	22.358	2.769.734	16.820	107.041	2.919.060
Chocó	81.683	36	130	124	418.970	21.861	0	12.022	534.826
Huila	17.547	35	43	29	12.455	992.220	10.334	78.057	1.100.386
La Guajira	399.841	29	108	111	76.050	365.276	0	39.145	880.560
Magdalena	21.869	39	95	80	120.011	1.135.841	0	63.811	1.341.746
Meta	24.613	32	96	30	8.867	889.526	0	116.558	1.039.722
Nariño	234.231	141	114	101	369.192	879.096	0	147.717	1.630.592
Norte de Santander	4.841	238	38	22	15.266	1.336.452	0	134.832	1.491.689
Quindío	3.120	6	22	2	9.300	500.638	0	26.816	539.904
Risaralda	29.935	18	96	29	24.613	792.687	0	96.023	943.401
Santander	1.412	347	156	49	40.600	1.984.470	0	157.803	2.184.837
Sucre	104.890	134	135	46	110.541	655.677	0	33.440	904.863
Tolima	47.312	161	60	37	14.145	1.178.126	12.766	90.346	1.330.187
Valle del Cauca	32.638	136	474	290	742.879	3.105.998	0	593.471	4.475.886
Arauca	8.308	4	51	13	10.286	222.846	0	20.666	262.174
Casanare	8.590	12	28	13	6.216	366.770	0	38.875	420.504
Putumayo	63.236	17	12	30	10.582	221.721	0	52.584	348.182
Archipiélago de San Andrés	1.871	0	20.332	10	10.766	21.392	0	6.909	61.280
Amazonas	40.904	5	13	0	476	25.584	0	9.607	76.589
Guainía	35.068	5	5	4	459	9.446	0	3.127	48.114
Guaviare	10.228	3	6	5	3.136	63.024	0	6.365	82.767
Vaupés	33.108	0	10	8	271	5.582	0	1.818	40.797
Vichada	61.709	1	5	5	577	29.374	0	16.137	107.808

Por lo tanto, se plantea armonizar las estimaciones de fase II y III mediante la distribución proporcional de la población omitida que no fue asignada en las anteriores fases, mediante la distribución

proporcional de la población omitida a la que aún no se le ha asignado pertenecía étnica, bajos el siguiente esquema:

1. Tomar como referente las participaciones étnicas dadas por la ECV 2018 a nivel nacional – Condición Columna.
2. Tomar como referente el total de poblacional ajustada por cobertura – Condición Fila.
3. Distribuir la población omitida sin asignar, mediante una tabla cuadrada a nivel departamental.
4. Condicionar el procedimiento para no estimar un total poblacional inferior a lo efectivamente censado.
5. Una vez se tenga convergencia, aplicar el mismo procedimiento a nivel municipal dentro del departamento.
6. Una vez se tenga convergencia, agregar a nivel departamental para obtener los resultados finales

Para ilustrar el mecanismo de ajuste de la tabla consistencia, es necesario dividir el ajuste por ciclos, en donde cada ciclo consta de dos (2) pasos. El primer paso es un ajuste horizontal (en este caso se ajusta hacia el total departamental o municipal corregido por cobertura), para esto se calcula una relación del total población y la suma de la distribución de la población según pertenencia étnica correspondiente a esa área geográfica, por último, se multiplica la relación calculada por la distribución del área (departamento o municipio). El segundo paso, es un ajuste vertical, el cual busca respetar los totales departamentales de la distribución según pertenencia étnica corregida por omisión censal, este ajuste se realiza de la misma manera, es decir calculando la relación entre el total a respetar y la suma de la distribución resultante del paso uno (1). Fue necesaria la construcción seis (6) ciclos para lograr la estimación final. Adicionalmente, se agregó una restricción en la construcción de la tabla de consistencia, la cual, consiste en asegurar que las estimaciones finales no fueran inferiores a la población efectivamente censada.

Finalmente, las estimaciones no son enteras, para lo cual se acude al uso de un redondeado acumulativo, que busca respetar los totales de población corregidos por cobertura departamentales y municipales.

6.3 Resultados

En los resultados de la tabla cuadrada, cada fila puede ser modelada mediante una distribución multinomial con 6 categorías[6], de la misma forma las columnas con 33 categorías. Con esta aproximación es posible generar una distribución empírica mediante un Bootstrap[7] paramétrico con 1000 réplicas para las filas y las columnas, de la cual se obtienen intervalos del 95% de confianza para la participación de cada grupo étnico dentro de cada departamento (fila) y la participación del departamento dentro de cada grupo étnico (columna), tomando el percentil 2,5 y 97,5 de la distribución empírica, así como sus respectivos coeficientes de variación.

La tabla 16 muestra los coeficientes de variación (cv´s) para las proporciones fila, resaltando que los cv´s para los grupos Indígena, NMAA y Ningún no sobrepasan el 5%, no obstante, los grupos Rrom, Raizal y Palenquero presentan cv´s hasta del 51%, que se explica por la baja participación de estos en algunos territorios, sin m embargo en el caso de Raizales en el Archipiélago de San Andrés y Palanquero en Bolívar presentan cv´s inferiores al 1,5% dada su presencia mayoritaria en dichos departamentos.

[6] Para los departamentos de Bogotá, Cundinamarca, Huila y Tolima las personas sin información étnica son sumadas a Ningún grupo con la finalidad de alterar lo menos posible las proporciones.
[7] Ver (González Gómez, 2015) y (Efron, 1979)

Tabla 16: Estimaciones Fila

Departamento	Indígena	cv	Rrom	cv	Raizal	cv	Palenquero	cv	NMAA	cv	Ningún	cv
Total Nacional	**4,4%**	**0,1%**	**0,0%**	**1,9%**	**0,1%**	**0,6%**	**0,0%**	**1,0%**	**9,1%**	**0,0%**	**86,4%**	**0,0%**
Antioquia	0,6%	0,5%	0,0%	8,3%	0,0%	3,8%	0,0%	5,4%	6,7%	0,1%	92,7%	0,0%
Atlántico	1,6%	0,5%	0,0%	10,0%	0,0%	4,5%	0,1%	2,5%	8,1%	0,2%	90,3%	0,0%
Bogotá, D.C.	0,3%	0,7%	0,0%	4,0%	0,0%	3,2%	0,0%	5,0%	1,6%	0,3%	98,1%	0,0%
Bolívar	0,3%	1,4%	0,0%	17,2%	0,0%	4,1%	0,3%	1,2%	20,2%	0,1%	79,2%	0,0%
Boyacá	0,6%	1,1%	0,0%	22,9%	0,0%	12,4%	0,0%	18,3%	1,0%	0,9%	98,4%	0,0%
Caldas	5,8%	0,4%	0,0%	15,8%	0,0%	9,4%	0,0%	13,2%	2,8%	0,6%	91,4%	0,0%
Caquetá	2,9%	0,9%	0,0%	21,9%	0,0%	17,5%	0,0%	18,7%	1,7%	1,2%	95,4%	0,0%
Cauca	23,8%	0,1%	0,0%	14,6%	0,0%	10,5%	0,0%	8,4%	24,2%	0,2%	51,9%	0,1%
Cesar	4,6%	0,4%	0,0%	20,9%	0,0%	8,8%	0,0%	8,2%	16,2%	0,2%	79,2%	0,0%
Córdoba	12,1%	0,2%	0,0%	8,3%	0,0%	7,8%	0,0%	8,2%	10,3%	0,2%	77,6%	0,0%
Cundinamarca	0,6%	0,7%	0,0%	10,0%	0,0%	7,9%	0,0%	9,9%	0,9%	0,6%	98,4%	0,0%
Chocó	16,9%	0,3%	0,0%	16,9%	0,0%	8,3%	0,0%	7,6%	78,9%	0,1%	4,1%	0,7%
Huila	1,6%	0,7%	0,0%	16,0%	0,0%	14,7%	0,0%	13,5%	1,4%	0,8%	97,0%	0,0%
La Guajira	46,2%	0,1%	0,0%	18,2%	0,0%	9,1%	0,0%	7,0%	10,6%	0,3%	43,1%	0,1%
Magdalena	1,7%	0,7%	0,0%	15,3%	0,0%	10,2%	0,0%	8,6%	10,9%	0,2%	87,5%	0,0%
Meta	2,4%	0,6%	0,0%	17,0%	0,0%	9,6%	0,0%	12,7%	1,1%	0,9%	96,4%	0,0%
Nariño	14,7%	0,2%	0,0%	7,8%	0,0%	9,3%	0,0%	7,8%	26,6%	0,1%	58,6%	0,1%
Norte de Santander	0,3%	1,4%	0,0%	6,6%	0,0%	16,2%	0,0%	15,6%	1,3%	0,7%	98,3%	0,0%
Quindío	0,6%	1,8%	0,0%	36,1%	0,0%	20,3%	0,0%	44,3%	2,1%	0,9%	97,3%	0,0%
Risaralda	3,2%	0,6%	0,0%	21,2%	0,0%	10,0%	0,0%	13,0%	3,4%	0,5%	93,3%	0,0%
Santander	0,1%	2,7%	0,0%	5,4%	0,0%	7,9%	0,0%	10,5%	2,4%	0,4%	97,5%	0,0%
Sucre	11,7%	0,3%	0,0%	8,4%	0,0%	8,5%	0,0%	10,9%	14,0%	0,3%	74,3%	0,1%
Tolima	3,6%	0,4%	0,0%	8,0%	0,0%	13,1%	0,0%	12,1%	1,4%	0,8%	95,1%	0,0%
Valle del Cauca	0,7%	0,5%	0,0%	8,5%	0,0%	4,6%	0,0%	4,4%	21,9%	0,1%	77,3%	0,0%
Arauca	3,2%	1,0%	0,0%	40,7%	0,0%	14,0%	0,0%	20,5%	5,0%	0,9%	91,8%	0,1%
Casanare	2,1%	1,1%	0,0%	27,2%	0,0%	18,4%	0,0%	18,7%	1,9%	1,1%	96,0%	0,0%
Putumayo	19,9%	0,3%	0,0%	21,7%	0,0%	25,9%	0,0%	12,8%	4,3%	0,8%	75,8%	0,1%
Archipiélago de San Andrés	3,2%	2,3%	0,0%	0,0%	33,2%	0,6%	0,0%	22,3%	23,7%	0,7%	39,9%	0,5%
Amazonas	59,0%	0,3%	0,0%	36,9%	0,0%	25,2%	0,0%	0,0%	0,9%	3,9%	40,1%	0,4%
Guainía	76,7%	0,3%	0,0%	35,2%	0,0%	37,0%	0,0%	33,0%	1,3%	4,2%	22,0%	0,9%
Guaviare	12,4%	0,9%	0,0%	44,0%	0,0%	34,7%	0,0%	29,8%	4,9%	1,5%	82,7%	0,2%
Vaupés	83,7%	0,2%	0,0%	0,0%	0,0%	28,8%	0,0%	23,7%	0,9%	5,0%	15,4%	1,2%
Vichada	65,8%	0,2%	0,0%	51,8%	0,0%	33,9%	0,0%	29,8%	0,8%	3,5%	33,4%	0,4%

La tabla 17 muestra los coeficientes de variación (cv´s) para las proporciones columna, donde el comportamiento de los cv´s es similar a las estimaciones por fila, mostrando que los grupos Rrom, Raizales y Palenqueros, presentan la mayor variabilidad.

Tabla 17: Estimaciones Columna

Departamento	Indígena	cv	Rrom	cv	Raizal	cv	Palenquero	cv	NMAA	cv	Ningún	cv
Antioquia	1,9%	0,5%	5,2%	8,1%	2,5%	3,7%	2,9%	5,4%	9,7%	0,1%	14,2%	0,0%
Atlántico	1,9%	0,5%	3,8%	9,9%	1,9%	4,7%	13,4%	2,4%	4,6%	0,2%	5,5%	0,1%
Bogotá, D.C.	0,9%	0,7%	22,2%	3,6%	4,1%	3,0%	3,3%	5,1%	2,7%	0,3%	17,4%	0,0%
Bolívar	0,3%	1,4%	1,2%	16,7%	2,3%	4,2%	58,3%	0,8%	9,5%	0,1%	3,9%	0,1%
Boyacá	0,3%	1,1%	0,7%	21,8%	0,2%	12,7%	0,3%	18,7%	0,3%	0,9%	2,9%	0,1%
Caldas	2,7%	0,4%	1,4%	16,0%	0,4%	9,5%	0,5%	13,6%	0,6%	0,6%	2,2%	0,1%
Caquetá	0,5%	1,0%	0,8%	20,3%	0,1%	17,3%	0,2%	19,1%	0,2%	1,3%	0,9%	0,2%
Cauca	16,4%	0,2%	1,5%	15,6%	0,4%	10,0%	1,3%	7,9%	8,1%	0,2%	1,8%	0,1%
Cesar	2,6%	0,4%	0,8%	21,1%	0,5%	9,0%	1,2%	8,7%	4,4%	0,2%	2,3%	0,1%
Córdoba	10,2%	0,2%	5,5%	8,1%	0,7%	7,7%	1,3%	8,6%	4,2%	0,2%	3,3%	0,1%
Cundinamarca	0,8%	0,7%	3,6%	9,5%	0,6%	8,2%	0,9%	9,3%	0,6%	0,6%	6,9%	0,1%
Chocó	4,2%	0,3%	1,4%	15,6%	0,5%	8,7%	1,5%	7,5%	9,6%	0,1%	0,1%	0,7%
Huila	0,8%	0,8%	1,4%	16,3%	0,2%	15,4%	0,5%	13,1%	0,4%	0,8%	2,6%	0,1%
La Guajira	19,1%	0,1%	1,1%	18,2%	0,4%	9,8%	1,7%	7,2%	2,1%	0,3%	0,9%	0,2%
Magdalena	1,0%	0,7%	1,5%	15,3%	0,4%	10,2%	1,2%	8,6%	3,3%	0,3%	2,8%	0,1%
Meta	1,2%	0,6%	1,3%	16,8%	0,4%	10,3%	0,5%	13,2%	0,3%	0,9%	2,4%	0,1%
Nariño	11,3%	0,2%	5,4%	7,8%	0,5%	8,9%	1,5%	7,7%	9,9%	0,1%	2,3%	0,1%
Norte de Santander	0,2%	1,5%	9,0%	6,1%	0,2%	16,0%	0,4%	15,6%	0,5%	0,7%	3,5%	0,1%
Quindío	0,1%	1,8%	0,3%	34,1%	0,1%	19,4%	0,0%	43,8%	0,3%	0,9%	1,3%	0,1%
Risaralda	1,4%	0,6%	0,8%	21,0%	0,4%	10,3%	0,5%	12,8%	0,7%	0,6%	2,1%	0,1%
Santander	0,1%	2,6%	12,9%	5,0%	0,6%	8,0%	0,8%	10,5%	1,2%	0,4%	5,1%	0,1%
Sucre	5,0%	0,3%	5,0%	8,5%	0,5%	8,3%	0,7%	11,2%	2,9%	0,3%	1,6%	0,1%
Tolima	2,2%	0,5%	5,9%	7,6%	0,2%	13,1%	0,6%	12,4%	0,4%	0,8%	3,0%	0,1%
Valle del Cauca	1,6%	0,6%	5,3%	8,0%	1,9%	4,2%	4,8%	4,1%	22,3%	0,1%	8,3%	0,1%
Arauca	0,4%	1,1%	0,2%	40,1%	0,2%	13,4%	0,2%	21,9%	0,3%	0,9%	0,6%	0,2%
Casanare	0,4%	1,1%	0,5%	26,2%	0,1%	18,7%	0,2%	19,9%	0,2%	1,1%	1,0%	0,2%
Putumayo	3,2%	0,4%	0,7%	21,6%	0,1%	26,4%	0,6%	12,5%	0,3%	0,8%	0,6%	0,2%
Archipiélago de San Andrés	0,1%	2,3%	0,0%	0,0%	79,5%	0,3%	0,2%	22,1%	0,3%	0,8%	0,1%	0,6%
Amazonas	2,1%	0,5%	0,2%	36,1%	0,1%	25,2%	0,0%	0,0%	0,0%	3,7%	0,1%	0,6%
Guainía	1,7%	0,5%	0,3%	35,0%	0,0%	37,5%	0,1%	34,2%	0,0%	4,1%	0,0%	0,9%
Guaviare	0,5%	1,0%	0,1%	42,7%	0,0%	33,9%	0,1%	30,1%	0,1%	1,6%	0,2%	0,4%
Vaupés	1,6%	0,6%	0,0%	0,0%	0,0%	27,7%	0,1%	24,3%	0,0%	5,2%	0,0%	1,2%
Vichada	3,3%	0,4%	0,1%	51,0%	0,0%	33,7%	0,1%	28,7%	0,0%	3,5%	0,1%	0,5%

La gráfica 28 muestra la participación étnica dentro de cada departamento y su respectivo intervalo de confianza donde se destaca el departamento del Chocó, con una participación Indígena del 17%, NMAA

del 79% y de ningún grupo étnico del 4%, siendo el departamento con mayor participación de NMAA. Así mismos, el departamento de Vaupés con una participación indígena del 84%.

Gráfica 28: Participaciones e intervalos de confianza fila

La gráfica 29 muestra la participación étnica dentro de cada grupo étnico y su respectivo intervalo de confianza donde se destaca el departamento del Valle del cauca como el departamento con un porcentaje de del 22% para el grupo NMAA, siendo el departamento con mayor presencia de dicha población. El Archipiélago de San Andrés con el 80% de la población Raizal residente en el país y, por último, Bolívar con el 58% de la población Palenquera.

Gráfica 29: Participaciones e intervalos de confianza columna

La implementación de la tabla cuadrada consigue armonizar las estimaciones realizadas en fase II y III de forma parsimoniosa. La estimación de población NMAA en total nacional cae dentro del intervalo de confianza de la ECV 2018. Aunque la población NMAA en Valle del Cauca no es cercana a lo estimado en la ECV 2018, la propuesta metodológica da cuenta de una participación del 22% dentro del departamento, así como del 22% del total nacional, siendo el departamento con mayor población NMAA en el territorio nacional.

Los comprimentos de los intervalos de confianza son pequeños, lo que indica un alto grado de precisión en las estimaciones. Por último, queda abierta la posibilidad de encontrar estrategias que permitan determinar las personas que están en la categoría de ningún grupo étnico (No realización de la pregunta, marcación arbitraria por parte del encuestador o asignación errónea por parte del informante idóneo)

6. Conclusiones y limitaciones

Este ejercicio se enmarca dentro de la categoría de estadísticas experimentales y trazó una estrategia de cuatro fases para lograr estimar la población étnica a nivel nacional, departamental y municipal. La primera fase determinó la proporción de personas pertenecientes a grupos étnicos a nivel nacional, utilizando la ECV 2018. La segunda fase asignó la pertenencia étnica de la población efectivamente censada sin información de autorreconocimiento en el CNPV 2018. La tercera fase utilizó la georreferenciación de la población efectivamente censada para determinar la participación de cada grupo étnico en la población omitida. Por último, en la cuarta fase se utilizó la metodología de tabla cuadrada para consolidar los resultados a nivel departamental y municipal.

La importancia de realizar estos ejercicios de estadísticas experimentales radica en que permite conocer y determinar las brechas que enfrentan los diferentes grupos étnicos en Colombia, en ámbitos como: salud, educación, acceso a servicios públicos y privados, mercado laboral, condiciones de pobreza, participación política, entre otros.

El ejercicio buscó inicialmente mitigar la caída de la participación de la categoría Negro, Mulato, Afrodescendiente y Afrocolombiano en el Censo Nacional de Población y Vivienda 2018 respecto al Censo General de 2005, mediante asignación probabilística de pertenecía étnica a la población sin información y omitida.

En lo que respecta a la población Negra, Afrocolombiana, Raizal y Palenquera - NARP, el ejercicio permitió estimar de manera oficial su volumen nacional en 4.671.160 personas que constituyen el 9.34% de la población nacional. Así, como una estimación a niveles departamentales y municipales de todos los grupos étnicos.

Como resultado final, la estimación logro adicionar 229.242 indígenas, 43 Gitanos, 22 Raizales, 4.469 Palenqueros, 1.446.813 Negros, Mulatos, Afrodescendientes y Afrocolombianos y 2.826.410 sin autorrecnocimiento étnico. Estos resultados, pueden ser objeto de nuevos análisis y, por lo tanto, ser actualizados metodológicamente. Su propósito es abrir el debate sobre el uso fuentes y métodos no tradicionales, así como establecer una población base que permita la realización proyecciones poblacionales con enfoque diferencial étnico

7. Bibliografía

DANE. (2021). Lecciones aprendidas del Censo Nacional de Población y Vivienda 20218, respecto a la NTC PE 1000:2017.

DANE (2020). Evaluación de cobertura nacional y subnacional CNPV 2018. Métodos y estimaciones. (documento sin publicar)

DANE. (2019). Metodología General Encuesta Nacional De Calidad De Vida – ECV.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Voicu, I. (2018). Using first name information to improve race and ethnicity classification. Statistics and Public Policy, 5(1), 1-13.

Stevens, F. R., Gaughan, A. E., Linard, C., & Tatem, A. J. (2015). Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. PloS one, 10(2), e0107042..

Friedman, J. H. (2002). Stochastic gradient boosting. Computational statistics & data analysis, 38(4), 367-378.