



Departamento Nacional
de Planeación - **DNP**

Realizar levantamiento de información que permita continuar con el seguimiento periódico a la percepción ciudadana sobre temas estratégicos del Plan Nacional de Desarrollo.

Producto 3: Informe de resultados del levantamiento de información

ARD-649 DOC_09

Bogotá, D.C. Diciembre 26 de 2023



	Realizar levantamiento de información que permita continuar con el seguimiento periódico a la percepción ciudadana sobre temas estratégicos del Plan Nacional de Desarrollo.	 Departamento Nacional de Planeación - DNP
	PRODUCTO No. 3	

Departamento Nacional de Planeación (DNP)
Dirección de Seguimiento y Evaluación de Políticas Públicas (DSEPP)
Carlos Alberto Castañeda

Comité Técnico de Seguimiento
Lilian Andrea Castro Villarreal
Ana María Arias Díaz
Edwin Andrés Cruz

Sistemas Especializados de Información S.A. - SEI
Equipo Consultor
Luz Adriana Arana - Directora
Antonio José Rey González – Coordinador de trabajo de campo
Edilneyi Zúñiga Avirama - Estadística

Cítese este documento así:

DNP. (2023). Realizar levantamiento de información que permita continuar con el seguimiento periódico a la percepción ciudadana sobre temas estratégicos del Plan Nacional de Desarrollo. Producto 3. Departamento Nacional de Planeación (DNP). Bogotá: Sistemas Especializados de Información S.A. - SEI

 Soluciones Estratégicas en Información	Realizar levantamiento de información que permita continuar con el seguimiento periódico a la percepción ciudadana sobre temas estratégicos del Plan Nacional de Desarrollo.	 Departamento Nacional de Planeación - DNP
	PRODUCTO No. 3	

Tabla de Contenido

1. INFORME DE RESULTADOS DEL LEVANTAMIENTO DE INFORMACIÓN.	4
1.1. Introducción	4
1.2. Descripción del procesamiento de los datos recolectados	4
1.2.1 Síntesis de la operación estadística por muestreo probabilístico implementada	4
1.2.2. Tratamiento de la no respuesta y ajuste de factores de expansión	11
1.2.2.1. Ajuste de la no respuesta con base en un modelo de clasificación (propensión)	11
1.2.2.2. Calibración	29
1.2.2.2. Suavizamiento	30
1.2.3. Directrices para el procesamiento de los resultados	30
1.2.4. Parámetros sugeridos para determinar la calidad de las estimaciones	32
1.2.5. Programas usados para la selección de la muestra y la estimación de indicadores	33
1.3. Proceso de revisión, validación, edición e imputación de los datos	33
1.3.1. Revisión de la información en campo	33
1.3.2. Captura de información	34
1.3.3. Metodología de depuración de datos	34
1.3.4. Codificación de preguntas de la opción otro cuál	35
1.4. Cuadros de salida y visualizador de datos	36
1.5. Bases de datos	37
2. TERCERA ENTREGA DE LA DOCUMENTACIÓN DE LAS BASES DE DATOS DE LA ENCUESTA	39
3. BIBLIOGRAFÍA	40
4. ANEXOS	42
ANEXO 1. Estadístico.zip	42
ANEXO 2. Bases de datos.zip	42
ANEXO 3. Formulario VF.pdf	42
ANEXO 4. Formulario maestro.pdf	42
ANEXO 5. Manual del encuestador.pdf	42
ANEXO 6. Cronograma del proyecto.pdf	42
ANEXO 7. PND_2022-2026.nesstar_entrega 3.zip	42

 Soluciones Estratégicas en Información	Realizar levantamiento de información que permita continuar con el seguimiento periódico a la percepción ciudadana sobre temas estratégicos del Plan Nacional de Desarrollo.	 Departamento Nacional de Planeación - DNP
	PRODUCTO No. 3	

1. INFORME DE RESULTADOS DEL LEVANTAMIENTO DE INFORMACIÓN.

1.1. Introducción

El presente documento contiene el informe final con los resultados del operativo de recolección de información, y ha sido preparado por el equipo técnico de SEI para presentar los resultados del procesamiento de la información para la EPC-PND 2023, donde se incluyen los aspectos más importantes relacionados con el tratamiento de los datos.

Se han organizado dos grandes capítulos, el primero de los cuales describe en detalle el procesamiento de la información, los cuadros de salida y el visualizador de datos. En el segundo capítulo se presenta la tercera y última entrega de la documentación de las bases de datos de la encuesta. Todo lo anterior se soporta con los anexos relacionados al final del documento.

1.2. Descripción del procesamiento de los datos recolectados

1.2.1 Síntesis de la operación estadística por muestreo probabilístico implementada

Para la Encuesta de Percepción Ciudadana, EPC-PND 2023, la población objetivo fueron las personas de 18 años y más, hombres y mujeres, de todos los municipios del país, residentes en la zona urbana (cabecera) y en la zona rural (centros poblados y rural disperso conformado por veredas), que además, pertenecen a todos los niveles socioeconómicos (estratos del 1 a 6 y hogares sin estratificación de acuerdo con el servicio de energía, estrato 0, los cuales se encuentran generalmente en la zona rural dispersa o veredas); esta población se estima en 38.195.238 personas (73,15 % del total nacional), de la cual el 22,35 % (8.535.459) reside en la zona rural (centros poblados o veredas) y el restante (29.659.779, es decir, el 77,65 %) en la zona urbana y, por nivel socioeconómico (estrato según el servicio de energía), su distribución (relativa aproximada), según la Superintendencia¹ de Servicios Públicos Domiciliarios es: Estrato 1, 25,7 %; Estrato 2, 41,5 %; Estrato 3, 23,3 %; Estrato 4, 5,9 %; Estrato 5, 2,3 %; Estrato 6, 1,3 %.

En cuanto a los dominios de análisis objeto de interés, estos corresponden a²:

- Total, nacional;
- 33 entidades territoriales (agregado) conformadas por cada uno de los 32 departamentos más Bogotá D.C. De estos, se tienen 7 entidades territoriales

¹ http://reportes.sui.gov.co/fabricaReportes/frameSet.jsp?idreporte=ele_com_096, reporte con corte a agosto de 2023, fuente: "http://sui.superservicios.gov.co/Reportes-del-sector/Energia/Reportes-comerciales/Consolidado-Energia-por-Empresa-Departamento-y-Municipio".

² Página 4 de 44. "ii) Estrategia de muestreo. Ítem: Subpoblaciones (estratos, dominios o post-estratos) de interés". Proceso de selección por concurso de méritos no. CM-003-19, Anexo no. 1, Anexo Técnico del Pliego de Condiciones Definitivo.

	Realizar levantamiento de información que permita continuar con el seguimiento periódico a la percepción ciudadana sobre temas estratégicos del Plan Nacional de Desarrollo.	 Departamento Nacional de Planeación - DNP
	PRODUCTO No. 3	

principales, las cuales son: Bogotá D.C., Antioquia, Valle del Cauca, Cundinamarca, Santander, Atlántico y Magdalena;

- Cada una de las 33 entidades territoriales;
- Total, nacional por zona: 1. Urbana; 2. Rural (centros poblados y veredas o zona rural dispersa);
- Total 7 entidades territoriales principales por zona (agregado);
- 26 entidades territoriales restantes (se excluyen las 7 entidades territoriales principales) por clase (1. Cabeceras, 2. Centros poblados y 3. Rural disperso).
- Regiones³ definidas⁴ como aglomeraciones de departamentos y de sus áreas metropolitanas: 1. Caribe⁵, 2. Santanderes⁶, 3. Eje Cafetero y Antioquia⁷, 4. Orinoquía⁸, 5. Amazonía⁹, 6. Central¹⁰, 7. Pacífica¹¹, 8. Insular¹²;
- Capitales de departamento y/o principales áreas metropolitanas;
- Clúster de categorías (departamentos o municipios) según la Contraloría General de la Nación¹³ (“CT01 - Categorización de Municipios 2023” y “Categorización por Departamentos 2023”, CGN, 2023);
- Clúster de capitales por región y zona.
- Municipios PDET.

En cuanto a los dominios relacionados con desagregaciones temáticas, específicamente, con características socioeconómicas que responden a un enfoque inclusivo (diferencial e interseccional¹⁴), tal como lo recomienda el Sistema Nacional Estadístico (SEN) de Colombia¹⁵, se definieron los siguientes:

- Sexo;
- Grupo etario/ciclo de vida;
- Nivel socioeconómico según el estrato del recibo de energía;
- Ocupación de la Población Económicamente Activa (PEA);

³ Se incluye la regionalización del PND 2018-2022 con el fin de aportar resultados a este nivel para que la DSEPP-DNP pueda realizar un análisis comparativo con los resultados del actual en relación con el primer levantamiento.

⁴ “Nota: la totalidad de la región (en los casos de Caribe, Eje Cafetero y Pacífica), se define como la suma de la región sin su área metropolitana más el área metropolitana correspondiente”, Proceso de selección por concurso de méritos no. CM-003-19, página 75 de 107, Pliego de Condiciones Definitivo.

⁵ Atlántico (sin A.M de Barranquilla), Bolívar, Cesar, Córdoba, La Guajira, Magdalena y Sucre.

⁶ Norte de Santander y Santander,

⁷ Caldas, Quindío, Risaralda y Antioquia (sin A.M del Valle de Aburra).

⁸ Arauca, meta, Casanare y Vichada.

⁹ Vaupés, Guainía, Amazonas, Guaviare, Caquetá y Putumayo.

¹⁰ Cundinamarca, Boyacá, Huila y Tolima.

¹¹ Cauca, Chocó, Nariño y Valle del Cauca (sin A.M de Cali).

¹² Archipiélago de San Andrés, Providencia y Santa Catalina.

¹³ Con base en: i) Población; ii) ICLD Contraloría (Miles de Pesos); iii) Gastos Funcionamiento Contraloría (Miles de Pesos); iv) % Gastos Funcionamiento / ICLD. Seis categorías más una especial (Bogotá D.C, Medellín, Cali, Bucaramanga, Barranquilla y Cartagena).

¹⁴ [chrome-extension://efaidnbmnnnibpcajpcgclefindmkaj/https://www.dane.gov.co/files/investigaciones/genero/guia-inclusion-enfoque-difencias-intersecciones-produccion-estadistica-SEN.pdf](https://efaidnbmnnnibpcajpcgclefindmkaj/https://www.dane.gov.co/files/investigaciones/genero/guia-inclusion-enfoque-difencias-intersecciones-produccion-estadistica-SEN.pdf). DANE, 2020.

¹⁵ <https://www.sen.gov.co/conozca-el-sen/que-es-el-sen>

 Soluciones Estratégicas en Información	Realizar levantamiento de información que permita continuar con el seguimiento periódico a la percepción ciudadana sobre temas estratégicos del Plan Nacional de Desarrollo.	 Departamento Nacional de Planeación - DNP
	PRODUCTO No. 3	

- Discapacidad;
- Víctima según RUV en el hogar;
- Grupo étnico;
- Estado conyugal;
- Migración;
- Parentesco;
- Nivel educativo;
- Acceso a la salud.

Con respecto a las unidades estadística del estudio, estas fueron:

Unidades de observación: son aquellas sobre las cuales se recaba la información de interés para el estudio, que para este caso son las personas de 18 años o más.

Unidades de análisis: son los distintos niveles de desagregación (es decir, dominios de análisis) en los que, de acuerdo con el diseño probabilístico planteado y el interés de los investigadores, se deben presentar los resultados. El objetivo es inferir resultados precisos y confiables para el total de la población (total nacional) de 18 años o más, así como por cada uno de los dominios antes mencionados.

Unidades de muestreo: son aquellas a seleccionar probabilísticamente en cada etapa de muestreo. En el contexto de la operación estadística probabilística, estratificada (5 niveles de estratificación), multietápica y de elementos (personas de 18 o más años), las unidades de muestreo posteriores a la primera etapa (Unidad Primaria de Muestreo – UPM) y que corresponden a un municipio o área metropolitana, se definieron como a continuación se presenta:

- **Área urbana:** manzana (Unidad Secundaria de Muestreo - USM), vivienda (Unidad Terciaria de Muestreo - UTM), hogar (Unidad Cuaternaria de Muestreo - UCM) y persona (Unidad Final de Muestreo - UFM).
- **Área rural:** centro poblado o vereda (Unidad Secundaria de Muestreo - USM).
- **Área rural - centro poblado:** manzana (Unidad Terciaria de Muestreo – UTM), vivienda (Unidad Cuaternaria de Muestreo - UCM), hogar (Unidad Quinaria de Muestreo – UQM) y persona (Unidad Final de Muestreo - UFM).
- **Área rural - vereda:** segmento (Unidad Terciaria de Muestreo – UTM) y persona (Unidad Final de Muestreo - UFM).

	Realizar levantamiento de información que permita continuar con el seguimiento periódico a la percepción ciudadana sobre temas estratégicos del Plan Nacional de Desarrollo.	 Departamento Nacional de Planeación - DNP
	PRODUCTO No. 3	

En cuanto a los períodos de referencia del estudio, en el instrumento (encuesta) se abordaron los siguientes lapsos: semana pasada, último mes, últimos seis meses, últimos 12 meses, últimas 24 horas y momento actual.

Es relevante considerar que la información acopiada y relacionada con estos períodos para cada unidad de observación estará condicionada al momento de la encuesta según el cronograma de visita definido para el trabajo de campo. Ahora, en cuanto al período de recolección, el levantamiento de información primaria a través de la encuesta de percepción ciudadana se realizó entre el 30 octubre y el 18 de diciembre de 2023.

De otra parte, con respecto a los marcos de muestreo por tipo de unidad (o etapa), en el caso de las UPM se tuvo en cuenta la división político-administrativa del país (DIVIPOLA), la proyección de población 2018-2026¹⁶ a nivel municipal, por zona, sexo y grupos de edad del Departamento Administrativo Nacional de Estadística (DANE). Para la conformación del marco muestral de los centros poblados (CP), se tuvo como referente el Marco Geoestadístico Nacional (MGN) actualizado en 2021¹⁷, y para la conformación del marco muestral de veredas se utilizó el inventario de veredas del Geoportal del DANE (vigencia 2020)¹⁸, generado a partir del Censo Nacional Agropecuario (CNA), 2014, DANE.

En algunos casos, fue necesario gestionar la cartografía generada por los entes territoriales municipales (mapa veredal solicitado en la alcaldía, específicamente, en la secretaría de planeación) para confirmar y validar la información oficial y actualizarla. Lo anterior, solo para el nivel de los municipios incluidos en la muestra.

Con respecto a las unidades de muestreo correspondientes a las manzanas cartográficas en la cabecera y centros poblados, se utilizó el MGN. Al interior de cada manzana cartográfica seleccionada se realizó un recuento de viviendas¹⁹ (listado exhaustivo de viviendas ocupadas) para seleccionar una muestra de éstas y, en cada una, identificar mediante un recuento los hogares²⁰ y seleccionar uno. Finalmente, en cada hogar seleccionado, se identificaron y listaron los residentes habituales, para identificar,

¹⁶ <https://www.dane.gov.co/index.php/estadisticas-por-tema/demografia-y-poblacion/proyecciones-de-poblacion>, Proyecciones y retroproyecciones de población municipal para el periodo 1985-2019 y 2020-2035 con base en el CNPV 2018, Serie municipal de población por área, sexo y edad para el periodo 2020-2035, DANE.

¹⁷ <https://geoportal.dane.gov.co/servicios/descarga-y-metadatos/descarga-mgn-marco-geoestadistico-nacional/#gsc.tab=0>, Descarga MGN Marco Geoestadístico Nacional, año 2021, DANE.

¹⁸ <https://geoportal.dane.gov.co/servicios/descarga-y-metadatos/descarga-nivel-de-referencia-de-veredas/#gsc.tab=0>, año 2020, DANE.

¹⁹ Espacio independiente y separado, habitado o destinado para ser habitado por una o más personas.

²⁰ Es una persona o grupo de personas, parientes o no, que ocupan la totalidad o parte de una vivienda; atienden necesidades básicas con cargo a un presupuesto común y generalmente comparten las comidas.

	Realizar levantamiento de información que permita continuar con el seguimiento periódico a la percepción ciudadana sobre temas estratégicos del Plan Nacional de Desarrollo.	 Departamento Nacional de Planeación - DNP
	PRODUCTO No. 3	

con base en la edad, todas las personas de 18 años o más y de esta manera, seleccionar una, por hogar.

En el caso de la zona rural dispersa (veredas seleccionadas) se implementó un procedimiento en campo para identificar viviendas ocupadas y construir segmentos definidos en función de estas (3 a 8 viviendas contiguas, es decir, conglomerado), con el fin de visitar todos los hogares de las viviendas incluidas en el segmento, y en cada uno, seleccionar una persona objeto de estudio.

Teniendo como referente lo ya mencionado, se definió una operación estadística por muestreo probabilístico para garantizar que el levantamiento de información primaria relacionado con la percepción ciudadana a los ejes transversales del Plan Nacional de Desarrollo - PND 2022-2026²¹ tuviera un alcance inferencial. La metodología definida se enfocó en los aspectos técnicos relacionados con la operación estadística por muestreo probabilístico, teniendo como referente, los Lineamientos del Proceso Estadístico en el Sistema Estadístico Nacional,^{22 23} así como el anexo técnico del contrato. Esta operación estadística guarda relación con los 10 levantamientos de corte trasversal realizados entre 2015 y 2022 para evaluar los resultados de acuerdo con la percepción ciudadana de los planes de desarrollo anteriores.

Para el diseño probabilístico multietápico y estratificado antes definido se utilizó la técnica de estratificación en distintos niveles:

- Nivel I: entidad territorial, es decir, departamentos más Bogotá D.C;
- Nivel II: municipio;
- Nivel III: zona;
- Nivel IV: manzanas en cabecera;
- Nivel V: zona rural – centro poblado o vereda) y de elementos (personas de 18 o más años).

Los niveles de estratificación I, III y V se conformaron con base en criterios administrativos, usando para esto la información oficial del DANE. En cuanto a los niveles de estratificación II (UPM) y IV (manzana) se utilizó un algoritmo para estratificar, en función de la cantidad de personas de 18 años o más proyectadas a junio de 2023 (Nivel II) y la cantidad de viviendas ocupadas (Nivel IV), el cual fue propuesto por Lavallée, P. and

²¹ "PROPUESTA PRELIMINAR CUESTIONARIO 2023.pdf" Son cinco ejes: i) Ordenamiento del territorio alrededor del agua; ii) Seguridad humana y justicia social; iii) DDHH a la alimentación; iv) Transformación productiva, internacionalización y acción climática; v) Convergencia regional.

²² https://www.dane.gov.co/files/sen/normatividad/Lineamientos_Proceso_Estadistico.pdf, julio 2017, DANE.

²³ <chrome-extension://efaidnbmnnnibpcjpcqlclefindmkaj/https://www.dane.gov.co/files/sen/lineamientos/Guia-elaboracion-documento-metodologico.pdf>

	Realizar levantamiento de información que permita continuar con el seguimiento periódico a la percepción ciudadana sobre temas estratégicos del Plan Nacional de Desarrollo.	 Departamento Nacional de Planeación - DNP
	PRODUCTO No. 3	

Hidiroglou, M.A. (1988)²⁴, desarrollado por Kozak, M. (2004)²⁵ e implementado en el software libre R por Sophie Baillargeon and Louis-Paul Rivest (2014)²⁶.

De otra parte, para la selección de las unidades de muestreo de inclusión probabilística, se utilizó el método de muestreo aleatorio simple (M.A.S.) sin reemplazamiento y el algoritmo Coordinado Negativo, tal como lo describe Gutiérrez, H. A. (2016).

Con respecto a la asignación del tamaño de muestra de unidades finales de muestreo para el Nivel I de estratificación (entidad territorial), esta se realizó de manera proporcional, utilizando como información auxiliar el total poblacional proyectado, y se ajustó por zona (urbana o rural), a partir de los escenarios que se generaron a través de la simulación, para encontrar el tamaño óptimo, con el cual se garantizó el cumplimiento de los márgenes de error exigidos en el anexo técnico del contrato para: i) El total nacional; ii) Por zona; iii) Para cada una de las 7 entidades territoriales principales; iv) Para el resto de las 26 entidades territoriales.

Es pertinente resaltar que el diseño probabilístico implementado fue resultado de definir una estrategia muestral eficiente²⁷ (en términos de sesgo y mínima varianza), para un estudio poblacional de la complejidad de este. Lo anterior, aclarando que el estimador propuesto para el total es insesgado (estimador de Horvitz-Thompson) y que la estimación de razones (cociente entre dos totales estimados con base en el estimador de Horvitz-Thompson) es levemente sesgado, con estimación aproximada de la varianza del estimador a través de la técnica de linealización de Taylor.

En cuanto a la cobertura alcanzada por unidad de muestreo (o etapa), sólo no se logró trabajar una UPM en el departamento de Guainía, debido a problemas con el acceso. Esto se documentó en el producto 2 (Informe de trabajo de campo) y como anexo se adjuntó una certificación en la que se presentan los detalles del caso.

²⁴Lavallée, P. and Hidiroglou, M.A. (1988). On the stratification of skewed populations. Survey Methodology, 14, 33-43.

²⁵Kozak, M. (2004). Optimal stratification using random search method in agricultural surveys. Statistics in Transition, 6(5), 797-806.

²⁶Univariate Stratification of Survey Populations. R package version. 2.2-5. <https://CRAN.R-project.org/package=stratification>. Usando la función strata. LH

²⁷ "La estrategia para conformar el plan muestral tiene como meta conseguir la mejor estimación, o las mejores estimaciones, en términos de sesgo y de varianza, consumiendo para ello la menor cantidad posible de recursos y sujeto a un cronograma establecido. En el proceso de selección de las muestras los sesgos son producidos por defectos graves en los marcos muestrales, tales como subcobertura, sobrecobertura y repetición. En el proceso de recolección de información se puede pensar en tres fuentes de error que generan sesgos importantes en los resultados, esas fuentes son el entrevistador, el proceso de medición, y el objeto observado. Los sujetos que realizan la observación o medición son también encargados de realizar los recuentos de los diseños multietápicos. Los graves errores que se pueden cometer en estas tareas están ligados a: una deficiente capacitación en los procesos y pasos a seguir; un proceso impropio de selección; formas de pago inadecuadas; procesos insuficientes de control, supervisión y corrección de errores". Bautista L., Técnica de diseño de encuestas, Simposio de Estadística 2000, páginas 37 y 38.

 Soluciones Estratégicas en Información	Realizar levantamiento de información que permita continuar con el seguimiento periódico a la percepción ciudadana sobre temas estratégicos del Plan Nacional de Desarrollo.	 Departamento Nacional de Planeación - DNP
	PRODUCTO No. 3	

Con respecto a la muestra de manzanas (3.162) fue necesario realizar un incremento del 42,5 % (seleccionar 1.344 manzanas adicionales) para lograr una cobertura del 105,8 %, es decir, realizar recuento de viviendas en 3.536. Lo anterior, a raíz de la no respuesta que se observó en este nivel (etapa de muestreo) por diversas causas. En las manzanas²⁸ de las cabeceras municipales o de los centros poblados donde se logró adelantar el trabajo de campo y realizar encuestas, es decir, donde se realizó recuento de viviendas (3.536) se seleccionaron 28.816 viviendas²⁹ (utilizando el aplicativo desarrollado para la implementación de los métodos y algoritmos de selección probabilísticos en campo) y en éstas se contactaron 17.220 hogares³⁰ (en 3.429 manzanas).

De estos (los 17.220) hogares contactados en viviendas, 15.811 (lo cual representa el 118% de los hogares esperados, es decir, 13.400) atendieron al personal de campo y suministraron la información relacionada con la matriz sociodemográfica, a partir de la cual (criterios de inclusión) se selecciona al informante idóneo con base en métodos probabilísticos.

Finamente, en los 15.811 hogares con información se logró encuestar a 13.455 personas³¹, lo cual representa una cobertura, frente al tamaño esperado (exigido en el anexo técnico del contrato) del 100,4 %. En la Tabla 1 a continuación se presenta el comparativo por unidad de muestreo de los tamaños según la seleccionado, lo esperado y lo efectivo. Para conocer más detalles acerca de la cobertura se sugiere consultar el producto 2.

Tabla 1. Cobertura alcanzada por unidad de muestreo

Unidad de muestreo	Tamaños seleccionados (A)	Tamaños esperados (B)	Unidades efectivas (C)	Cobertura efectiva con respecto a lo esperado (C/B)
Área metropolitana/Municipios (cabeceras) ¹	<u>240</u>	<u>238</u>	<u>237</u>	<u>99,6%</u>
Centros poblados (rural)	<u>179</u>	<u>160</u>	<u>160</u>	<u>100,0%</u>
Veredas (rural disperso)	<u>197</u>	<u>154</u>	<u>152</u>	<u>98,7%</u>
Manzanas ²	4.506	3.162	3.536	111,8%
Viviendas	28.816	13.400	17.220	128,5%
Hogares	17.220	13.400	15.811	118,0%

²⁸ En la Tabla 10 se presenta la distribución de las manzanas seleccionadas según resultado de contacto o visita en campo.

²⁹ En la Tabla 11 se presenta la distribución de las viviendas seleccionadas según resultado de contacto.

³⁰ En la Tabla 12 se presenta la distribución de los hogares seleccionados según resultado de contacto.

³¹ En la Tabla 13 se presenta la distribución de las personas seleccionadas según resultado de contacto.

 Soluciones Estratégicas en Información	Realizar levantamiento de información que permita continuar con el seguimiento periódico a la percepción ciudadana sobre temas estratégicos del Plan Nacional de Desarrollo.		 Departamento Nacional de Planeación - DNP	
	PRODUCTO No. 3			
Unidad de muestreo	Tamaños seleccionados (A)	Tamaños esperados (B)	Unidades efectivas (C)	Cobertura efectiva con respecto a lo esperado (C/B)
Personas	15.811	13.400	13.454	100,4%

Notas:

¹ San Felipe (CD) en Guainía, en reemplazo de Pana Pana (CD), no se trabajó.

² Manzanas del MGN, DANE, 2021, en cabecera y centros poblados.

Fuente. Datos operativos de campo, reporte final de cobertura, 20 de diciembre de 2023. Información incluida en el anexo estadístico ("A649_MUESTRA_20DIC23_FINAL.xlsx")

1.2.2. Tratamiento de la no respuesta y ajuste de factores de expansión

1.2.2.1. Ajuste de la no respuesta con base en un modelo de clasificación (propensión)

Teniendo en cuenta que la no respuesta se presenta principalmente en el nivel de la unidad de muestreo manzana, a continuación, se presentan algunos modelos de clasificación planteados para estimar la probabilidad de respuesta, es decir, la probabilidad de realizar al menos una encuesta, reconociendo que no es posible asumir el supuesto clásico donde los respondientes y no respondientes tienen el mismo comportamiento y corrigiendo de esta manera, el sesgo potencial de selección que se puede inducir en la muestra por este fenómeno.

Ahora, una vez seleccionado el modelo de clasificación para el ajuste de no respuesta (NR) a los factores de expansión básicos en el nivel de manzana, este se multiplica por los factores básicos del resto de etapas, ajustados por no respuesta con base en el enfoque propuesto por Valliant et al. (2018), en el cual es posible implementar un ajuste de no respuesta clasificando las unidades de observación seleccionadas en: elegibles, no elegibles y desconocidos. De esta manera se genera el factor básico final ajustado por no respuesta para cada persona encuestada, el cual, luego, se somete a la técnica de calibración. En este contexto, a continuación se presentan los pasos que se siguieron para escoger el modelo con el mejor ajuste de no respuesta en el nivel de manzana:

Lectura de datos

Base de entrada: variables sociodemográficas consolidadas por el DANE a nivel manzana junto con la información de cobertura final a este mismo nivel, mientras que la variable respuesta corresponde a una indicadora de si hubo respuesta en la manzana (1) o no (0).

 Soluciones Estratégicas en Información	Realizar levantamiento de información que permita continuar con el seguimiento periódico a la percepción ciudadana sobre temas estratégicos del Plan Nacional de Desarrollo.	 Departamento Nacional de Planeación - DNP
	PRODUCTO No. 3	

Para mejorar el análisis, las variables DANE de conteos de viviendas y personas con ciertas características fueron consideradas como proporciones de los totales de hogares y personas respectivamente (hoja “Datos_prop” del archivo “datos_NR_LEV1.xlsx”)

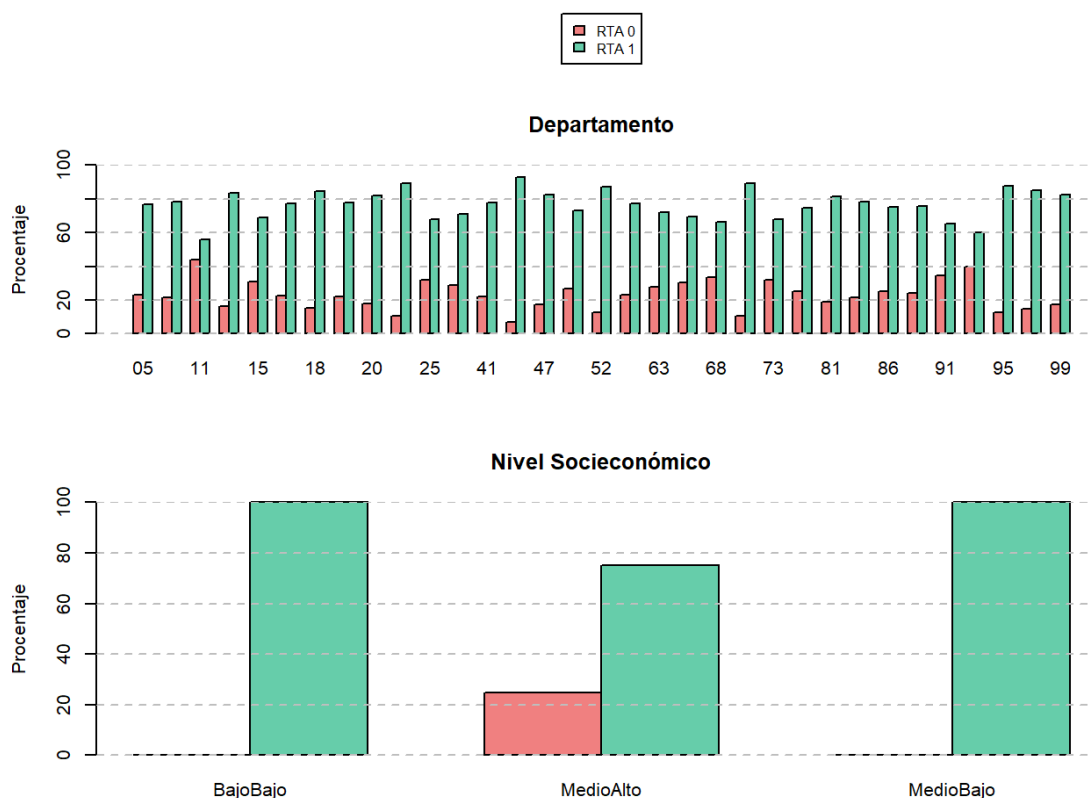
```
## # A tibble: 5 × 38
##   RTA DPTO DENSIDAD TIP_CASAS TIP_APTO HOGARES ENERGIA `NSE_BAJO-BAJO`
##   <dbl> <chr>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1  1 05    0.00143  0.267  0.667   10 0.667    0.7
## 2  1 05    0.134   0.0833  0.917   12 1        1
## 3  1 05    0.0236  0.383   0.596   38 0.809    0
## 4  1 05    0.110   0.163   0.837   39 0.907    0.821
## 5  1 05    0.00958  1      0      5 1        0.2
## #> 30 more variables: NSE_BAJO <dbl>, `NSE_MEDIO-BAJO` <dbl>,
## # `NSE_MEDIO-ALTO` <dbl>, NSE <chr>, CON_ACU <dbl>, SIN_ACU <dbl>,
## # CON_ALC <dbl>, SIN_ALC <dbl>, CON_GAS <dbl>, SIN_GAS <dbl>,
## # NO_INF_GAS <dbl>, CON_RECOL <dbl>, SIN_RECOL <dbl>, CON_INT <dbl>,
## # SIN_INT <dbl>, NO_INF_INT <dbl>, MUJERES <dbl>, `EDAD_20-29` <dbl>,
## # `EDAD_30-39` <dbl>, `EDAD_40-49` <dbl>, `EDAD_50-59` <dbl>,
## # `EDAD_60-69` <dbl>, `EDAD_70-79` <dbl>, EDAD_80 <dbl>, NE_BAJO <dbl>, ...
## [1] 3837 38
```

Para la relación entre las variables y su nombre, consultar la tabla en la hoja “Variables Datos” del archivo “datos_NR_LEV1.xlsx”.

Breve análisis exploratorio

Variables de tipo cualitativo: en este caso se cuenta con dos variables cualitativas: el Departamento y el Nivel socioeconómico, como se puede observar en la Figura 1 a continuación.

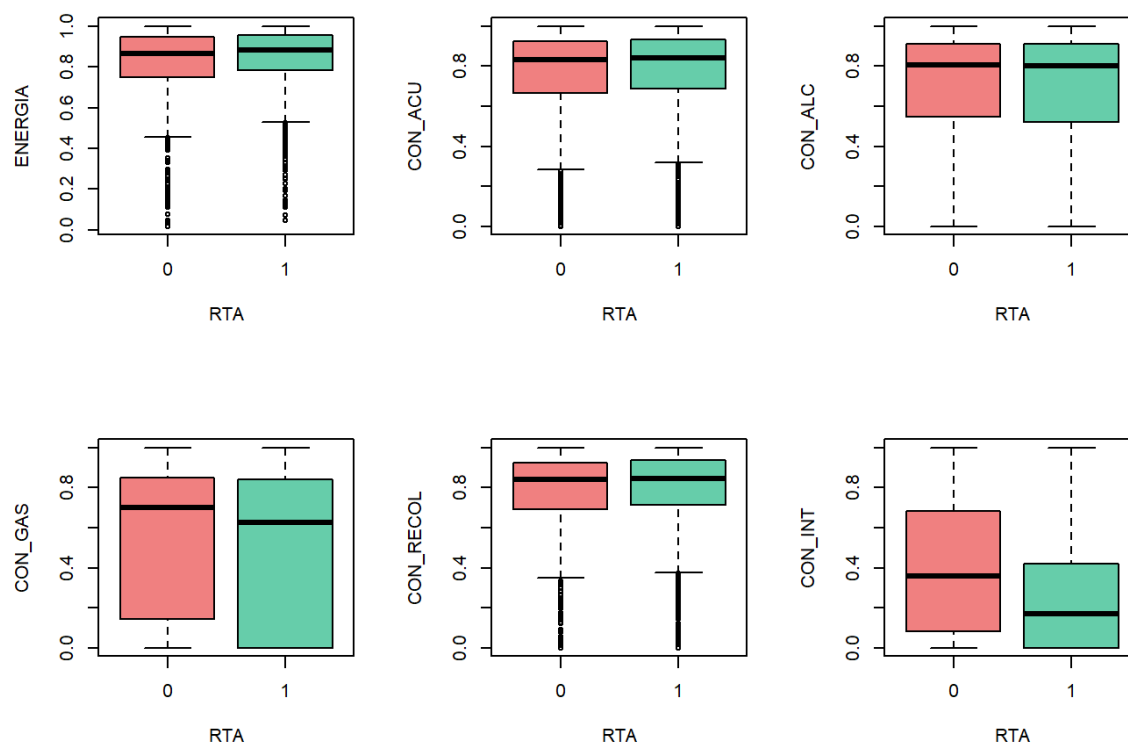
Figura 1. Distribución de las manzanas seleccionadas por entidad territorial a partir del código DANE según respuesta (1: Con encuestas / 0: Sin encuestas)



Fuente: Datos operativo de campo, reporte final de cobertura, 20 de diciembre de 2023. Información incluida en el anexo estadístico ("A649_MUESTRA_20DIC23_FINAL.xlsx").

Variables de tipo cuantitativo: inicialmente se grafica en la Figura 2, la distribución del conteo de viviendas por manzana **con** acceso a diferentes servicios públicos discriminado por tipo de respuesta (respondió o no). Esto para elegir uno de dichos servicios para ser incluido en el modelo.

Figura 2. Distribución de las manzanas seleccionadas por acceso a diferentes servicios públicos según respuesta (1: Con encuestas / 0: Sin encuestas)

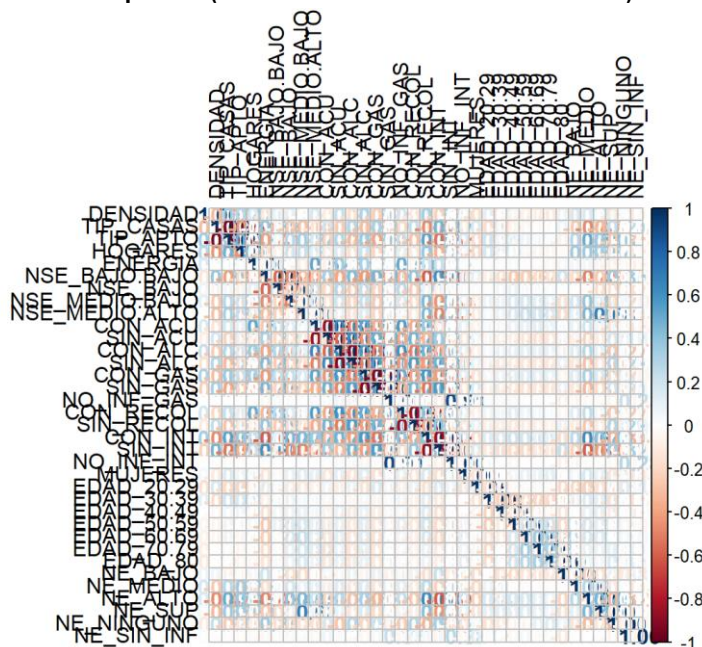


Fuente: Datos operativo de campo, reporte final de cobertura, 20 de diciembre de 2023. Información incluida en el anexo estadístico ("A649_MUESTRA_20DIC23_FINAL.xlsx").

Ahora, un ejercicio similar se puede realizar para los conteos de viviendas **sin** acceso a esos mismos servicios.

Entonces, si bien las variables referentes a acceso a internet son las más relevantes para el proceso, se decide dejar todas las variables de servicios. Adicionalmente, se calcula la matriz de correlación de las variables involucradas en el análisis. Se concluye que hay una alta correlación entre algunas de las variables, pero para el modelo no se quita ninguna, como se observa en la Figura 3.

Figura 3. Distribución de las manzanas seleccionadas por acceso a diferentes servicios públicos según respuesta (1: Con encuestas / 0: Sin encuestas)



Fuente: Datos operativo de campo, reporte final de cobertura, 20 de diciembre de 2023. Información incluida en el anexo estadístico ("A649_MUESTRA_20DIC23_FINAL.xlsx").

Preparación de las variables a utilizar en el modelo

De las 38 variables explicativas disponibles, se dejan de usar para la NR las correspondientes a los conteos por nivel socioeconómico ya que se incluye una variable categórica con el nivel socioeconómico modal de la manzana (NSE).

Conjunto de variables a trabajar

Para el modelo se utilizan 34 variables explicativas siendo la mayoría numéricas, excepto dos que son tipo factor (Departamento y NSE).

```
## [1] 3837 34
```

División en conjunto Train y Test

Se divide el conjunto de datos en un 80% para entrenamiento y un 20% para prueba.

```
set.seed(12345)

train <- createDataPartition(y = logros$RTA, p = 0.8, list = FALSE, times = 1)

datos_train <- logros[train,]
datos_test <- logros[-train,]
```

Estandarización de variables numéricas

Se estandarizan todas las variables numéricas, como sigue:

```
# Test
for(i in 4:34){
  datos_test[,i] <- (datos_test[,i] - mean(datos_train[,i])) / sd(datos_train[,i])
}

# Train
datos_train[,c(4:34)] <- map_df(datos_train[,c(4:34)], function(columna) {
  columna <- (columna - mean(columna)) / sd(columna) })
```

Para los modelos de regresión penalizada (modelo 1 y 3) se ajustan las variables predictoras como un conjunto de variables dummy (0 y 1) y la variable respuesta de tipo numérica.

Ajuste de modelos

A continuación, se ajustan cuatro modelos logísticos y se muestra la matriz de confusión construida en el conjunto Test para evaluar la calidad de cada uno de ellos.

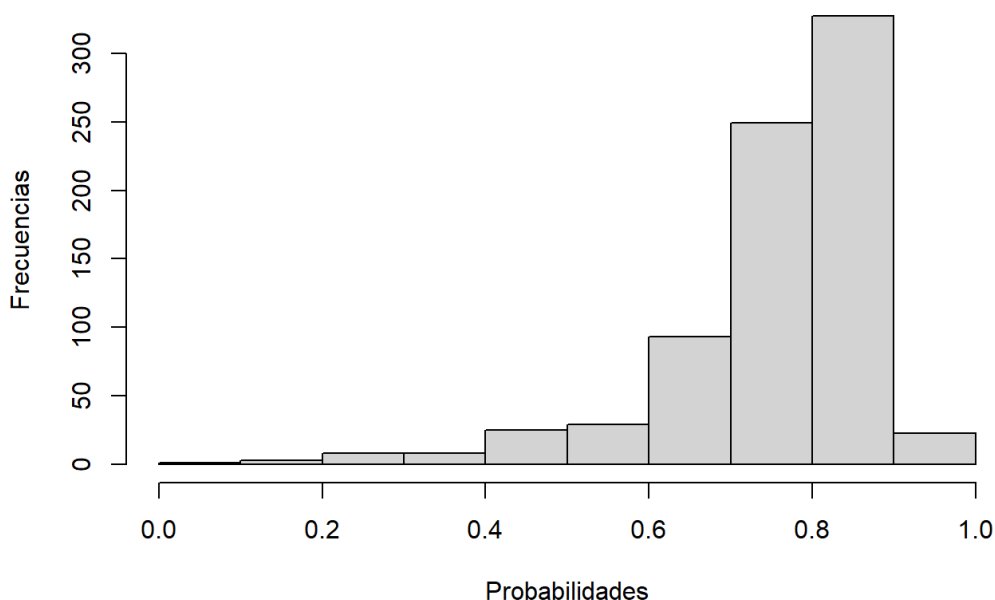
Modelo 1. Saturado

Se ajusta modelo con todas las variables disponibles, el cual será llamado **M.full**. Debido a la posibilidad de multicolinealidad entre las variables explicativas, el modelo completo o saturado se ajusta usando penalización Ridge.

```
set.seed(123)
cv.full <- cv.glmnet(x, y, alpha = 0, family = "binomial")
# Ajuste final en datos Train
M.full <- glmnet(x, y, alpha = 0, family = "binomial",
  lambda = cv.full$lambda.min)
```

```
##
##
## Cell Contents
## |-----|
## |          N |
## |      N / Row Total |
## |-----|
##
##
## Total Observations in Table: 766
##
##
##      | Observados
## Predictos |    0 |    1 | Row Total |
## -----|-----|-----|-----|
##      0 |    33 |    12 |    45 |
##      |    0.73 |    0.27 |    0.06 |
## -----|-----|-----|-----|
##      1 |   155 |   566 |   721 |
##      |    0.21 |    0.79 |    0.94 |
## -----|-----|-----|-----|
## Column Total |   188 |   578 |   766 |
## -----|-----|-----|-----|
```

Histograma de las probabilidades del modelo M.full



Modelo 2. Usando selección de variables

El modelo **M.step** resulta del método de selección de variables “Stepwise” (en su versión “both”, en ambos sentidos) usando la función `stepAIC()` del paquete MASS de R.

```
M.full1 <- glm(RTA ~ ., family = binomial(link = "logit"), data = datos_train)
```

```
# Selección de variables "Both"
```

```
M.step = M.full1 %>% stepAIC(trace = FALSE, direction = "both")
```

```
summary(M.step)
```

```
##
```

```
## Call:
```

```
## glm(formula = RTA ~ NSE + TIP_CASAS + HOGARES + SIN_ACU + CON_GAS +
```

```
## SIN_GAS + CON_RECOL + CON_INT + SIN_INT + EDAD_20.29 + EDAD_30.39 +
```

```
## EDAD_50.59 + NE_MEDIO + NE_SUP + NE_SIN_INF, family = binomial(link = "logit"),
```

```
## data = datos_train)
```

```
## Coefficients:
```

```
## Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept) 14.45006 271.63278 0.053 0.957575
```

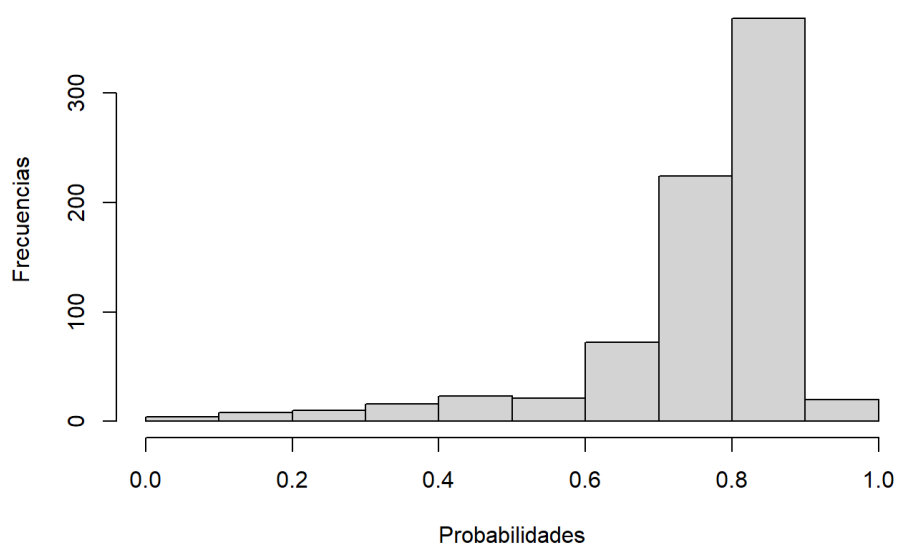
PRODUCTO No. 3

```
## NSEMedioAlto -13.22930 271.63279 -0.049 0.961156
## NSEMedioBajo 0.87062 923.59100 0.001 0.999248
## TIP_CASAS 0.19852 0.05214 3.807 0.000141 ***
## HOGARES -0.16067 0.05107 -3.146 0.001654 **
## SIN_ACU -0.11104 0.05464 -2.032 0.042125 *
## CON_GAS -2.19055 1.30941 -1.673 0.094341 .
## SIN_GAS -2.16440 1.27424 -1.699 0.089397 .
## CON_RECOL 0.11619 0.06226 1.866 0.062022 .
## CON_INT 1.69751 0.96172 1.765 0.077551 .
## SIN_INT 2.11851 1.02178 2.073 0.038139 *
## EDAD_20.29 -0.17332 0.04893 -3.542 0.000397 ***
## EDAD_30.39 -0.06961 0.04818 -1.445 0.148500
## EDAD_50.59 -0.09213 0.04803 -1.918 0.055103 .
## NE_MEDIO -0.08328 0.04761 -1.749 0.080247 .
## NE_SUP -0.30765 0.04866 -6.322 2.57e-10 ***
## NE_SIN_INF -0.12407 0.04119 -3.013 0.002591 **
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 3425.6 on 3070 degrees of freedom
## Residual deviance: 3072.2 on 3054 degrees of freedom
## AIC: 3106.2
##
## Number of Fisher Scoring iterations: 13
##
## Cell Contents
## |-----|
## | N |
## | N / Row Total |
## |-----|
## Total Observations in Table: 766
## | Observados
```

PRODUCTO No. 3

##	Predichos	0	1	Row Total
##	----- ----- ----- -----			
##	0	43	18	61
##		0.70	0.30	0.08
##	----- ----- ----- -----			
##	1	145	560	705
##		0.21	0.79	0.92
##	----- ----- ----- -----			
##	Column Total	188	578	766
##	----- ----- ----- -----			

Histograma de las probabilidades del modelo M.step



Modelo 3. Usando Penalización Lasso

A continuación, se ajusta un modelo usando penalización Lasso dentro de la función `glmnet` del paquete del mismo nombre. El parámetro `Lambda` es encontrado usando validación cruzada 10-fold. Este modelo se identificará como ***M.lasso***.

```
set.seed(123)
cv.lasso <- cv.glmnet(x, y, alpha = 1, family = "binomial")
# Fit the final model on the training data
```

```
M.lasso <- glmnet(x, y, alpha = 1, family = "binomial",
```

```
lambda = cv.lasso$lambda.min)
```

```
coef(M.lasso)
```

```
## 66 x 1 sparse Matrix of class "dgCMatrix"
```

```
##          s0
```

```
## (Intercept) 1.92846892
```

```
## DPTO08      .
```

```
## DPTO11      .
```

```
## DPTO13      0.28656164
```

```
## DPTO15     -0.21207071
```

```
## DPTO17      .
```

```
## DPTO18      .
```

```
## DPTO19      .
```

```
## DPTO20      0.07277423
```

```
## DPTO23      0.00310350
```

```
## DPTO25     -0.15962510
```

```
## DPTO27     -0.18613373
```

```
## DPTO41      .
```

```
## DPTO44      0.45320296
```

```
## DPTO47      .
```

```
## DPTO50      .
```

```
## DPTO52      0.50306650
```

```
## DPTO54      .
```

```
## DPTO63      .
```

```
## DPTO66     -0.23283670
```

```
## DPTO68     -0.06680652
```

```
## DPTO70      0.19025243
```

```
## DPTO73     -0.26317754
```

```
## DPTO76      .
```

```
## DPTO81      .
```

```
## DPTO85      .
```

```
## DPTO86      .
```

PRODUCTO No. 3

DPTO88 .

DPTO91 -0.12289931

DPTO94 -0.59600751

DPTO95 0.30015830

DPTO97 .

DPTO99 .

NSEMedioAlto -0.71450045

NSEMedioBajo .

DENSIDAD .

TIP_CASAS 0.17330287

TIP_APTO .

HOGARES -0.14519987

ENERGIA 0.03143537

CON_ACU 0.07857044

SIN_ACU .

CON_ALC 0.01267578

SIN_ALC .

CON_GAS .

SIN_GAS .

NO_INF_GAS .

CON_RECOL 0.04850969

SIN_RECOL .

CON_INT .

SIN_INT 0.23934219

NO_INF_INT -0.00257327

MUJERES .

EDAD_20.29 -0.10169742

EDAD_30.39 -0.01491350

EDAD_40.49 .

EDAD_50.59 -0.02241742

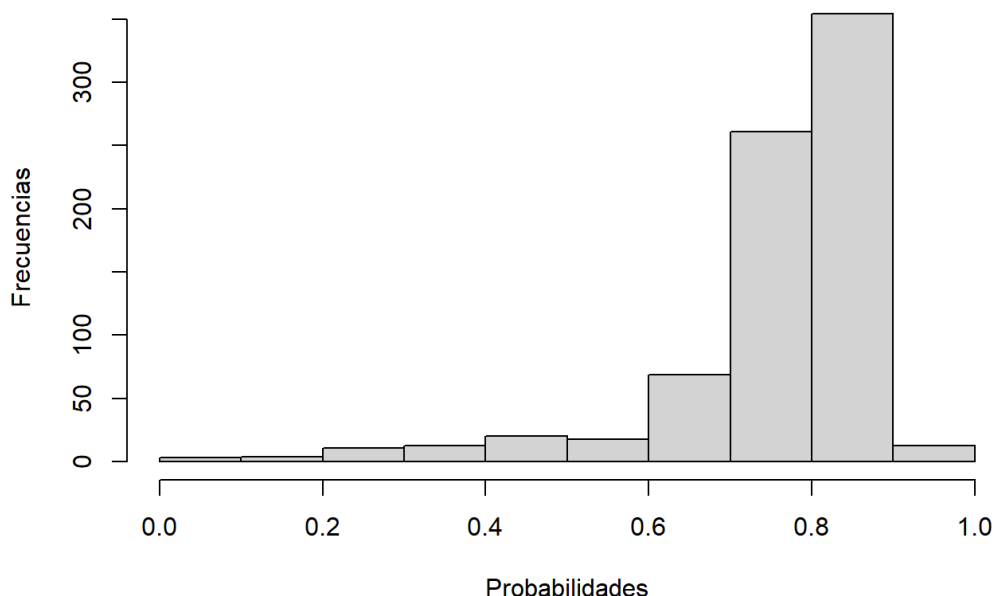
EDAD_60.69 .

EDAD_70.79 .

PRODUCTO No. 3

```
## EDAD_80      .
## NE_BAJO      0.02679489
## NE_MEDIO     -0.04183270
## NE_ALTO      .
## NE_SUP       -0.29759797
## NE_NINGUNO   .
## NE_SIN_INF   -0.09507431
##
##
## Cell Contents
## |-----|
## |          N |
## |      N / Row Total |
## |-----|
##
##
## Total Observations in Table: 766
##
##
##      | Observados
## Predichos |    0 |    1 | Row Total |
## -----|-----|-----|-----|
##      0 |   35 |   16 |   51 |
##      |  0.69 |  0.31 |  0.07 |
## -----|-----|-----|-----|
##      1 |  153 |  562 |  715 |
##      |  0.21 |  0.79 |  0.93 |
## -----|-----|-----|-----|
## Column Total |   188 |   578 |   766 |
## -----|-----|-----|-----|
```

Histograma de las probabilidades del modelo M.lasso



Modelo 4. Random Forest

A continuación, se ajusta un modelo de clasificación Random Forest (**M.Rf**) dentro de la metodología de Statistical Learning usando validación cruzada para elegir los parámetros que optimicen el AUC de la curva ROC (roc_auc) del modelo. Se usa el comando `step_upsample(RTA, over_ratio = 1)` para balancear las categorías de la variable respuesta RTA.

```
## # A tibble: 1 × 9
##   mtry trees max.depth .metric .estimator mean   n std_err .config
##   <int> <dbl>   <dbl> <chr>  <chr>    <dbl> <int>  <dbl> <chr>
## 1     8   150     5 roc_auc binary  0.705   5 0.00953 Preprocessor1_Mo...
## Warning: `pull_workflow_fit()` was deprecated in workflows 0.2.3.
## ⓘ Please use `extract_fit_parsnip()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
## parsnip model object
```

##

Ranger result

##

Call:

ranger::ranger(x = maybe_data_frame(x), y = y, mtry = min_cols(~8L, x), num.trees = ~150, max.dept
h = ~5, importance = ~"none", seed = ~123, num.threads = 1, verbose = FALSE, probability = TRUE)

##

Type: Probability estimation

Number of trees: 150

Sample size: 4632

Number of independent variables: 33

Mtry: 8

Target node size: 10

Variable importance mode: none

Splitrule: gini

OOB prediction error (Brier s.): 0.1992008

##

Cell Contents

##

|-----|

| N |

| N / Row Total |

|-----|

Total Observations in Table: 766

| Observados

Predichos | 0 | 1 | Row Total |

-----|-----|-----|-----|

0 | 93 | 92 | 185 |

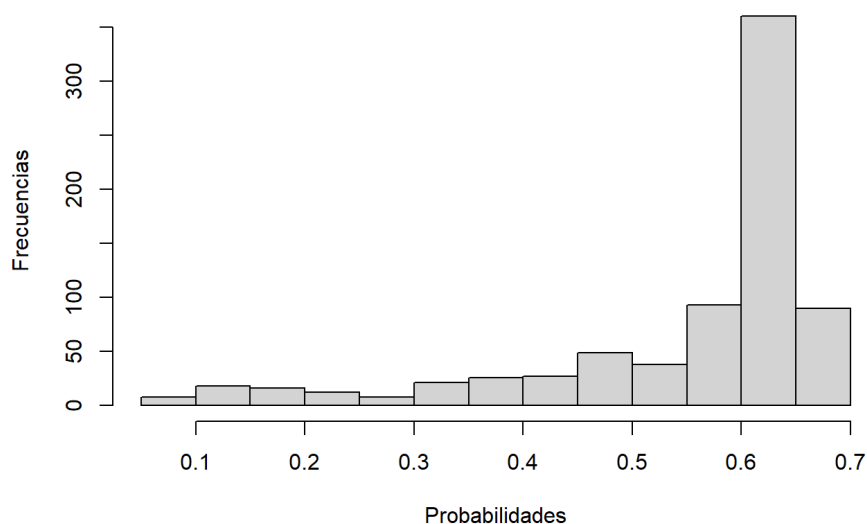
| 0.50 | 0.50 | 0.24 |

-----|-----|-----|-----|

1 | 95 | 486 | 581 |

##		0.16		0.84		0.76	
##	----- ----- ----- -----						
##	Column Total		188		578		766
##	----- ----- ----- -----						

Histograma de las probabilidades del modelo RF

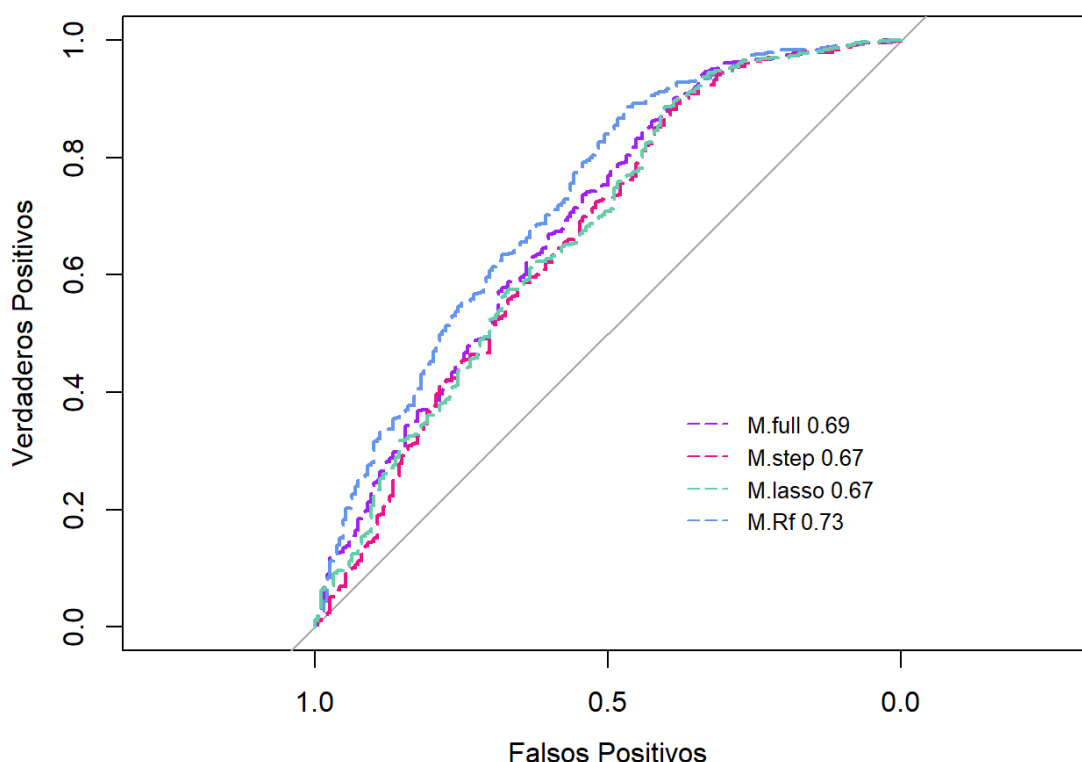


Comparación de modelos

Se compararon los modelos usando como parámetros de calidad el accuracy así como las probabilidades de falsos positivos y de falsos negativos para las matrices de confusión mostradas anteriormente, utilizando como punto de corte 0.5. Los resultados se presentan a continuación.

##	M.full	M.step	M.lasso	M.Rf
## accuracy.test	0.78198433	0.78720627	0.77937337	0.7558747
## FalsoNeg.test	0.02076125	0.03114187	0.02768166	0.1591696
## FalsoPos.test	0.82446809	0.77127660	0.81382979	0.5053191

Adicionalmente, se graficó la curva ROC de los cuatro modelos usando el conjunto de Test. Al respecto, se concluyó que el mejor modelo es **M.Rf** ya que maximiza el área bajo la curva.



Conclusión

Teniendo como referente el modelo de no respuesta ajustado en el levantamiento 4 - EPC PND 2018-2022 y con base en los datos obtenidos del levantamiento EPC PND 2022-2026, los modelos propuestos tienen un desempeño aceptable para clasificar pues el área bajo la curva ROC es próxima a 0.7. En particular para el modelo de **M.Rf** este criterio es de 0.73, siendo el valor más alto encontrado.

Adicionalmente, con base en el resumen de la tabla anterior, la evaluación del modelo a través de validación cruzada (y con lo realizado en el levantamiento 4 - EPC PND 2018-2022, como referente), la propuesta del equipo estadístico es utilizar para el ajuste de la no respuesta en el nivel de manzana, el modelo **M.Rf** ya que este es el que minimiza la probabilidad de falsos positivos y falsos negativos conservando un accuracy de aproximadamente 0.75.

A partir del *propensity score*, resultado de la implementación del modelo **M.Rf**, el factor básico de las manzanas fue ajustado por $1/\hat{\phi}_i$, donde $\hat{\phi}_i$, es la propensión de respuesta estimada para la unidad (manzana) i . De esta manera, el factor básico de expansión ajustado por no respuesta se expresa como:

 Soluciones Estratégicas en Información	Realizar levantamiento de información que permita continuar con el seguimiento periódico a la percepción ciudadana sobre temas estratégicos del Plan Nacional de Desarrollo.	 Departamento Nacional de Planeación - DNP
	PRODUCTO No. 3	

$$Flex_{manzana,ajusteNR} = \frac{Flex_{básico_manzana}}{\hat{\phi}_i}$$

Ahora, para el resto de las unidades de muestreo, se determinó utilizar la metodología de ajuste propuesta por Valliant, R., et al (2013). Este ejercicio permitió ajustar los tamaños poblacionales, y de esta manera, reducir los sesgos que son inducidos por la presencia de la omisión. Específicamente, esta metodología, consiste en construir tres clústeres a partir de los resultados de contacto observados y generar un factor de ajuste excluyendo los casos “no elegibles”, los cuales son considerados “fuera del universo de estudio”. El resultado de este procedimiento es factor que ajusta el factor básico de cada unidad de muestreo.

Al respecto, los pasos implementados fueron:

- 1) Se asignó probabilidad 0 a las unidades no elegibles.
- 2) Se calculó el factor básico de expansión para elegibles y no elegibles por convocatoria.
- 3) Se realizó la suma de los factores básicos de expansión por convocatoria sin tener en cuenta a los desconocidos.
- 4) Se calculó el tamaño del universo de estudio (población objetivo) a partir de la fuente administrativa suministrada por la entidad para la construcción del marco (directorio).
- 5) Se calculó un cociente, por convocatoria, entre lo obtenido en el paso 3 (numerador) y lo obtenido en el paso 4 (denominador), lo cual representa el factor de ajuste por omisión muestral según técnica de Valliant.

Así, las expresiones matemáticas relacionadas con la metodología anterior son:

$$f_{ajuste,i} = a_{1b} * d_{oi}$$

Donde:

d_{oi} , factor de expansión básico;

$$a_{1b} = \frac{\sum_{i \in S_b} d_{oi}}{\sum_{i \in S_{b,KN}} d_{oi}};$$

S_b , muestra seleccionada en el estrato b;

$S_{b,KN}$, subconjunto de la muestra seleccionada para la cual es posible determinar si existen unidades finales de muestreo que pertenecen o no a la población objetivo (factibles y no

	Realizar levantamiento de información que permita continuar con el seguimiento periódico a la percepción ciudadana sobre temas estratégicos del Plan Nacional de Desarrollo.	 Departamento Nacional de Planeación - DNP
	PRODUCTO No. 3	

factibles de encuestar); es decir, es equivalente a excluir de la muestra solo las unidades que generan duda sobre su pertenencia a la población objetivo (elegibilidad desconocida). De esta manera, el $f_{ajuste,i} = 0$, para los elementos de S_b en los que se desconoce si hacen parte o no de la población objetivo.

Por lo tanto, el factor de expansión final para la unidad k queda definido por la siguiente expresión:

$$F_{final,k} = Fex_{básico,k} * f_{ajuste,k}$$

Finalmente, al realizar el producto de los factores básicos ajustados por no respuesta para cada unidad (persona) se genera el factor final ajustado por no respuesta, el cual se somete al procedimiento de calibración.

1.2.2.2. Calibración

Fundamentalmente, el método de calibración consiste en el ajuste a los factores de expansión que ya fueron ajustados por la no respuesta observada, buscando afectar en menor medida los pesos muestrales, pero que al final, la restitución se realice sobre el total del universo, con base en información auxiliar, en este caso, de las proyecciones de población municipales por zona, DANE.

Para esto, existen varios métodos entre lineales y logísticos, que buscan construir un conjunto de nuevos pesos, similares en tamaño a los que resultaron del inverso de la probabilidad de selección y que cumplan con la ecuación de calibración:

$$\sum_{k \in S} w_k y_k = t_y.$$

El objetivo que se persigue con este proceso es expandir los resultados de los r respondientes, los cuales son un subconjunto de la muestra inicialmente planteada s , al total de la población U . El escenario perfecto es aquel en el que se cuente con información auxiliar para todos los elementos seleccionados en la última etapa y para los diferentes niveles de conglomeración seleccionadas en el muestreo multietápico.

Estevao & Särndal (2002) proponen 10 escenarios que contemplan todos los posibles casos en que puede implementarse un estimador de calibración, dependiendo de la información auxiliar disponible para un muestreo en dos fases.

Los pesos finales w_k se construyen a partir de los pesos iniciales f_k ajustados a partir de un vector de información. Estos estarán dados por:

 Soluciones Estratégicas en Información	Realizar levantamiento de información que permita continuar con el seguimiento periódico a la percepción ciudadana sobre temas estratégicos del Plan Nacional de Desarrollo.	 Departamento Nacional de Planeación - DNP
	PRODUCTO No. 3	

$$w_k = f_k v_k, \quad v_k = 1 + \lambda'_r y_k$$

Donde $\lambda'_r = (Y - \sum_r f_k y_k)' (\sum_r f_k y_k y'_k)^{-1}$.

Los detalles de la calibración, así como los resultados derivados de esta se encuentran en el Anexo 1 (Estadístico) que hace parte de este producto.

1.2.2.2. Suavizamiento

La metodología de suavizamiento permite controlar aquellos valores atípicos que se producen debido a las condiciones de la matriz de calibración y a la cobertura alcanzada en cada dominio definido. Para este propósito se implementó la metodología propuesta por Plotter (1990, 1993) para “suavizar” los valores atípicos de los factores identificados luego del ajuste por no respuesta y calibración (w_k). De esta manera, se genera el factor final ajustado para realizar el procesamiento de resultados.

Así como en el caso de la calibración, Los detalles computacionales y los resultados derivados de esta se encuentran en el Anexo 1 que hace parte de este producto.

1.2.3. Directrices para el procesamiento de los resultados

En el Producto No. 2, Informe Final de Trabajo de Campo, se describe la metodología de captura y depuración de la información obtenida en campo. Un elemento fundamental en este proceso es la incorporación, en el aplicativo de captura, de la malla de validación, la cual consiste en una matriz que muestra las relaciones lógicas entre los diferentes capítulos y variables del formulario; allí se establece la secuencia de las preguntas y los códigos válidos, lo cual sirve como referente para identificar inconsistencias y errores en la información, partiendo de normas y rangos de valores aceptables. Esto ocurre en tiempo real, es decir, durante el registro de la información que suministra el encuestado.

Ahora, con respecto al procesamiento de la información, este se realizó utilizando el paquete Stata (Versión 14) y calculando las medidas de calidad (Error estándar (EE), margen de error (EE*1.96), efecto de diseño (DEFF), límites del intervalo de confianza (IC) al 95 %, Deff y Error estándar relativo (cve)) asociadas.

	Realizar levantamiento de información que permita continuar con el seguimiento periódico a la percepción ciudadana sobre temas estratégicos del Plan Nacional de Desarrollo.	 Departamento Nacional de Planeación - DNP
	PRODUCTO No. 3	

De otra parte, como ya se explicó, se implementó la técnica de calibración, de acuerdo con el alcance establecido en el Informe Metodológico. Para esto, se utilizó el procedimiento Calibrate³² para generar los pesos calibrados. Además, se utilizó el método lineal predeterminado, cuyo estimador es el GREG, propuesto por Deville y Särndal (1992).

Con respecto a los factores de expansión, estos se calcularon con base en las probabilidades de inclusión por unidad de muestreo, incluidas en el producto 1 (Informe metodológico). Además, se realizó un ajuste por no respuesta, calibración y suavizamiento, lo cual ya se expuso previamente.

Así mismo, se generó el factor final de expansión ajustado por no respuesta, calibrado y suavizado, con base en el cual, se realizó la estimación puntual de los resultados. La distribución absoluta de la población objeto de inferencia según algunos de los principales dominios geográficos de análisis se incluye en el Anexo 1.

Finalmente, el DoFile (código) utilizado para la calibración (así como la matriz) se encuentra en el Anexo 1, así como el DoFile para la generación de los cuadros de salida. Al respecto, la estructura del código (.do) genérico (Svyset³³) para la estimación de resultados en el nivel nacional, por sexo y región, es:

```
/* Definición Diseño Muestral */
svyset UPM [pweight= FEX_CALIB ], fpc(N_UPM) strata (ID_subestrato)
singleunit(certain) || Llave_Clase, fpc(N_Clase) strata (Zona) || NIM_mz, fpc(N_mz) strata
(subestrato_MZ) || llave_viv, fpc(N_viviendas) || llave_viv, fpc(N_hogares) || _n
,fpc(N_personas)
```

Donde:

UPM: identifica en la base la unidad primaria de muestreo (municipio o área metropolitana);

FEX_CALIB]: es el factor final en el nivel de la persona de 18 años o más encuestada, teniendo en cuenta las etapas de muestreo, el ajuste por no respuesta y el factor de calibración por zona, sexo y nivel socioeconómico según subregión;

Fpc: son los tamaños (o factores de corrección) de unidades de muestreo por nivel de estratificación: para UPM fpc(N_UPM) , para el nivel Clase fpc(N_Clase) , para manzana

³² <http://fmwww.bc.edu/repec/bocode/c/calibrate.html>

³³ <https://www.stata.com/manuals13/svysyestimation.pdf>

 Soluciones Estratégicas en Información	Realizar levantamiento de información que permita continuar con el seguimiento periódico a la percepción ciudadana sobre temas estratégicos del Plan Nacional de Desarrollo.	 Departamento Nacional de Planeación - DNP
	PRODUCTO No. 3	

fpc(N_mz), para la vivienda fpc(N_viviendas), para los hogares , fpc(N_hogares) y para las personas fpc(N_personas);

strata : estrato en el nivel de las UPM's (ID_subestrato), zona (urbano/rural) (Zona), de manzanas en la cabecera (subestrato_MZ);

singleunit(certain): permite identificar que se incluyen UPM como estratos forzosos;

Llave_Clase: identifica la cabecera municipal, el centro poblado o la vereda, en donde se realizó la encuesta;

NIM_mz: identificador de la manzana (en cabecera o centro poblado) o del segmento (vereda);

llave_viv: identificador de la vivienda;

_n: identificador de la persona (contador=1).

Una vez se ejecuta el código anterior (.do), se declara el dominio y el parámetro a estimar, así como la variable para la cual se requiere la estimación puntual. Por ejemplo, para los dominios: nacional, sexo y región, la proporción, para la variable M1_103, se declara de la siguiente forma:

**nacional*

svy, vce(linearized): proportion M1_103

**sexo*

svy, vce(linearized) over(Sexo) : prop M1_103

**region*

svy, vce(linearized) over(region) : prop M1_103

Finalmente, las salidas de información (en extensión .xlsx) se encuentran en el Anexo 1.

1.2.4. Parámetros sugeridos para determinar la calidad de las estimaciones

Para garantizar el uso adecuado de los resultados, el CTS y SEI, acordaron incluir en cada cuadro de salida las siguientes notas técnicas para orientar al lector en la interpretación correcta de las estimaciones:

 Soluciones Estratégicas en Información	Realizar levantamiento de información que permita continuar con el seguimiento periódico a la percepción ciudadana sobre temas estratégicos del Plan Nacional de Desarrollo.	 Departamento Nacional de Planeación - DNP
	PRODUCTO No. 3	

Nota 1: Datos expandidos y ajustados por estructura poblacional con base en el censo de población y vivienda 2.018 y los datos de proyecciones de población a 2023 realizados por el DANE.

Nota 2: Los referentes para determinar estimaciones con validez inferencial son el coeficiente de variación estimado (CVE), el error estándar (EE) y el margen de error ($EE \cdot 1,96$).

Nota 3: Estimaciones resaltadas son imprecisas (con base en el Coeficiente de Variación Estimado – CVE), por lo cual es mejor abstenerse de su interpretación dado que carecen de validez inferencial, ya sea por sus márgenes de calidad o por ser calculadas con pocas unidades de observación.

Nota 4: Los márgenes de calidad para CVE son: < 2,5 % Excelente calidad; 2,6 % - 5,0 % Muy buena; 5,1 % - 10,0 % Buena calidad; 10,1 % - 15,0 % Aceptable calidad; 15,1 % - 20,0 % Calidad regular. CVE > 20,0 % se asocia a estimaciones imprecisas.

1.2.5. Programas usados para la selección de la muestra y la estimación de indicadores

En el Anexo 1 se encuentran los DoFile en Stata utilizados para la estimación de los indicadores tipo proporción para cada dominio y categoría de análisis por dominio de interés.

1.3. Proceso de revisión, validación, edición e imputación de los datos

1.3.1. Revisión de la información en campo

Dicha revisión fue hecha directamente en campo por el supervisor de grupo, mediante el uso de una tableta en la cual, por medio de un aplicativo diseñado especialmente, se podían abrir las encuestas ya realizadas para hacer verificación en una muestra, utilizando las siguientes técnicas de supervisión: acompañamiento, verificación personal posterior y verificación telefónica. Las inconsistencias encontradas fueron corregidas y enviadas a la oficina central.

Adicionalmente desde la oficina principal se verificaron, para una muestra de 826 encuestas realizadas (6%), el 100% de los audios de las entrevistas, cuyo resultado fue retroalimentado al personal de campo, principalmente en técnicas de entrevista, algunos conceptos y casos puntuales de diligenciamiento de algunas variables. Cada supervisor hizo reuniones de retroalimentación con sus encuestadores, con el fin de señalar los errores encontrados en la revisión de formularios, así como en las inconsistencias reportadas desde la oficina central.

	Realizar levantamiento de información que permita continuar con el seguimiento periódico a la percepción ciudadana sobre temas estratégicos del Plan Nacional de Desarrollo.	 Departamento Nacional de Planeación - DNP
	PRODUCTO No. 3	

1.3.2. Captura de información

El formulario de encuesta fue diseñado para ser digitado en Dispositivos Móviles de Captura (DMC), en este caso equipos celulares, y para ello se elaboró un robusto aplicativo que además permitió administrar la muestra. En el diseño de este aplicativo se incluyó una malla de validación, la cual contiene controles (pases, saltos, filtros, rangos) y validaciones de la información a capturar.

1.3.3. Metodología de depuración de datos

El proceso de depuración de las tablas de datos se realizó en cada una de las fases del operativo y se describe a continuación:

- Durante la aplicación de la encuesta: sucede en tiempo real con el informante, pues el programa de captura contiene la malla de validación programada, lo cual permite validar la información recolectada. Dichas validaciones impiden saltar preguntas que por el flujo del formulario deben ser contestadas, también impide colocar cantidades fuera de rango, y permite la marcación de una sola respuesta o de múltiples, dependiendo de cómo se requiera la respuesta.
- Durante la recepción de los datos al servidor: la información es enviada diariamente desde campo por cada encuestador, directamente al servidor de SEI y al momento de subir los datos se detectan anomalías con la sincronización. Diariamente se revisa la completitud de la información y se generan reportes de seguimiento.
- Durante la revisión de los datos: se generan tablas de datos y se realiza revisión por parte del coordinador de trabajo de campo. Esta revisión busca determinar si el programa de validación está arrojando como inconsistencias datos correctos o si está omitiendo inconsistencias. Además se hace una revisión de frecuencias simples para identificar inconsistencias u omisiones, y éste es un proceso que se repite hasta considerar que el nivel de depuración es suficiente, lo cual sucede cuando los hallazgos de anomalías son iguales a cero.

En caso de detectar alguna anomalía en el programa se solicitan los ajustes al responsable de sistemas y se realizan pruebas antes de liberar una nueva actualización del aplicativo para el equipo de campo.

Adicionalmente, también se reciben y procesan los resultados de las técnicas de supervisión (acompañamiento, verificación posterior y verificación telefónica) con el fin de hacer retroalimentación a los grupos de campo y generar reportes de errores por grupo y encuestador.

 Soluciones Estratégicas en Información	Realizar levantamiento de información que permita continuar con el seguimiento periódico a la percepción ciudadana sobre temas estratégicos del Plan Nacional de Desarrollo.	 Departamento Nacional de Planeación - DNP
	PRODUCTO No. 3	

1.3.4. Codificación de preguntas de la opción otro cuál

Con respecto a la codificación de las preguntas con opción de respuesta “otro, ¿cuál?”, el coordinador de trabajo de campo revisó y homogenizó respuestas, pero además creó algunas nuevas categorías, las cuales se presentan en la Tabla 2.

Tabla 2. Nuevas categorías producto de la revisión de respuestas de Otro: ¿cuál?

NUMERO DE PREGUNTA	PREGUNTA	NUEVAS CATEGORÍAS
205	Usted NO utilizó internet para realizar compras o pagos de facturas o servicios públicos porque: Otro, ¿cuál?	INTERNET MALO
		LE GUSTA PRESENCIAL
		NO ES COSTUMBRE
		NO LE GUSTA
		NO LO NECESITA
		NO REALIZA PAGOS
		NO TIENE CELULAR
		NO TIENE DINERO
507	De acuerdo con lo que usted sabe, ¿cuál cree que es la principal razón para que no todas las personas accedan a los servicios de los bancos o entidades financieras, como crédito, ahorro, etc.? Otro, ¿cuál?	FALTA DE REQUISITOS
		MIEDO A ENDEUDARSE
		NO HAY INGRESOS
1007	¿Cuál cree usted que es el principal problema que enfrenta la educación pública en primaria y bachillerato en Colombia? Otro, ¿cuál?	CORRUPCIÓN
		FALTA DE APOYO DEL GOBIERNO
		FALTA DE COMPROMISO DE LOS PADRES
		FALTA DE DISCIPLINA
		FALTA DE EDUCACIÓN EN EL HOGAR
		FALTA DE INTERÉS DE LOS ESTUDIANTES
		FALTA DE RECURSOS
		FALTA DE RESPETO
		FALTA DE VALORES
		MALA CALIDAD DE LA EDUCACIÓN
		NINGÚN PROBLEMA
1401	¿Cuál es el medio de transporte que usted utiliza principalmente para desplazarse dentro de su municipio o vereda? Otro, ¿cuál?	BICICLETA
1404	¿Cuál cree que es el principal problema del transporte público en su municipio? Otro, ¿cuál?	COSTO DE LA GASOLINA
		FALTA DE BUSES O EN MAL ESTADO
		FALTA DE INFRAESTRUCTURA
		FALTA DE VÍAS O EN MAL ESTADO
		LAS MOTOS
		NO HAY TRANSPORTE PUBLICO
		TRANCONES
		TRANSPORTE INFORMAL

 Soluciones Estratégicas en Información		Realizar levantamiento de información que permita continuar con el seguimiento periódico a la percepción ciudadana sobre temas estratégicos del Plan Nacional de Desarrollo.	 Departamento Nacional de Planeación - DNP
		PRODUCTO No. 3	
NUMERO DE PREGUNTA	PREGUNTA	NUEVAS CATEGORÍAS	
1603	¿Cuál cree usted que es el principal aspecto positivo de su barrio? Otro, ¿cuál?	BUENAS VÍAS	
		CLIMA	
		COMERCIO	
		ECONÓMICO	
		ILUMINACIÓN	
		INFRAESTRUCTURA	
		LIMPIO	
		SERVICIOS PÚBLICOS	
1604	¿Cuál cree usted que es el principal problema o aspecto negativo de su barrio? Otro, ¿cuál?	ANIMALES Y MASCOTAS	
		FALTA DE EMPLEO	
		HABITANTES DE LA CALLE	
		INUNDACIONES	
		NO ES COMERCIAL	
		ZONA DE RIESGO	
1701	¿Cuál cree usted que es el principal problema ambiental que afecta el país?	EXPLOTACIÓN DE HIDROCARBUROS	
		FALTA DE CULTURA	

Fuente. SEI. EPC 2023.

1.4. Cuadros de salida y visualizador de datos

Así como en el caso de los códigos computacionales, la tabla de datos de salida con el cruce de los 28 dominios de análisis (de tipo geográfico y de enfoque inclusivo) se encuentran en extensión .xlsx en el Anexo 1.

El tablero de visualización con los resultados (estimaciones tipo proporción) fueron desarrollados en la herramienta Power BI de Microsoft. El visor y su instructivo se incluyen en el Anexo 1.

 Soluciones Estratégicas en Información	Realizar levantamiento de información que permita continuar con el seguimiento periódico a la percepción ciudadana sobre temas estratégicos del Plan Nacional de Desarrollo.	 Departamento Nacional de Planeación - DNP
	PRODUCTO No. 3	

1.5. Bases de datos

A continuación, se describe la composición del formulario aplicado y sus datos fueron organizados en una base de datos, la cual se adjunta en el Anexo 2. La descripción también se encuentra en el archivo PDF llamado “Descripción”, en la carpeta “Documentación de las tablas de datos” de dicho Anexo.

La base de datos contiene un diccionario organizado de acuerdo con las preguntas aplicadas. La base de datos tiene los factores de expansión utilizados para el procesamiento y los resultados analizados están expandidos. En el Anexo 2 también se encuentra un archivo en PDF llamado “Léame”, que explica la estructura de la base de datos. También se encuentran dos (2) carpetas: “Documentación de las tablas de datos” y “Tablas de datos anonimizadas”. A continuación, se describe su contenido:

1. Documentación de las tablas de datos

- Diccionario de datos
 - Módulo Hogares
 - EXCEL
 - SPSS
 - Módulo Recuento viviendas
 - EXCEL
 - SPSS
- Descripción.pdf

2. Tablas de datos

- Módulo Hogares
 - CSV
 - EXCEL
 - SPSS
 - STATA
- Módulo Recuento viviendas
 - CSV
 - EXCEL
 - SPSS
 - STATA

Se encuentran disponibles archivos de las tablas de datos anonimizadas en formato .csv, .xlsx, .sav y .dta. Cada fila corresponde a un formulario aplicado, y cada columna corresponde a una variable o pregunta del formulario.

 Soluciones Estratégicas en Información	Realizar levantamiento de información que permita continuar con el seguimiento periódico a la percepción ciudadana sobre temas estratégicos del Plan Nacional de Desarrollo.	 Departamento Nacional de Planeación - DNP
	PRODUCTO No. 3	

Las tablas de datos se estructuran con el objetivo de respetar la secuencia de las preguntas dentro del formulario. Los archivos correspondientes a las tablas se encuentran en la carpeta: “Tablas de datos anonimizadas”.

El diccionario de datos contiene la descripción de las variables de las tablas del producto a entregar, en términos de su organización, estructura y formato. Los formatos de los diccionarios se encuentran en dos tipos de formato en .xlsx y en .sav.

Columna	Descripción	Columna	Descripción
Nombre	Nombre de variable	Valores	Opciones de respuesta
Tipo	Clasificación tipo variable	Perdidos	Valores perdidos
Anchura	Cantidad caracteres	Columnas	Cantidad columnas
Decimales	Cantidad decimales	Alineación	Izquierda / Derecha
Etiqueta	Pregunta	Medida	Nominal / Escala

Imputación de los datos: La malla de validación integrada en el instrumento de captura controla la información que se recoge. Controla el flujo de la encuesta, los saltos, los filtros y las respuestas de los informantes. Esto garantiza que las encuestas tengan la información completa y de calidad, evitando la necesidad de imputar datos.

	Realizar levantamiento de información que permita continuar con el seguimiento periódico a la percepción ciudadana sobre temas estratégicos del Plan Nacional de Desarrollo.	 Departamento Nacional de Planeación - DNP
	PRODUCTO No. 3	

1.5.1. Materiales de respaldo de la operación estadística

En el Anexo 3 se incluye la versión final de formulario que fue programada y en el Anexo 4 se presenta el formulario maestro que contiene los identificadores para cada variable. Adicionalmente en el Anexo 5 se adjunta el manual del encuestador en su versión final, y en Anexo 6 está el cronograma con las fechas ajustadas conforme al desarrollo de la recolección y entrega de productos.

1.5.2. Reporte técnico sobre los cálculos

Se elaboró, por solicitud del CTS, un protocolo para dar contexto al lector y orientar en la lectura y correcta interpretación de los resultados. Este protocolo se incluye en el Anexo 1.

2. TERCERA ENTREGA DE LA DOCUMENTACIÓN DE LAS BASES DE DATOS DE LA ENCUESTA

En el Anexo 7 se presenta la tercera entrega del formato Nesstar con las respectivas secciones diligenciadas. Esta versión corresponde a la entrega final del formato consolidado .nesstar utilizando la plantilla DDI para metadatos y microdatos.

	Realizar levantamiento de información que permita continuar con el seguimiento periódico a la percepción ciudadana sobre temas estratégicos del Plan Nacional de Desarrollo.	 Departamento Nacional de Planeación - DNP
	PRODUCTO No. 3	

3. BIBLIOGRAFÍA

- Bautista, L. Diseños de muestreo estadístico. Universidad Nacional de Colombia. (1998).
- Baillargeon, S. and Rivest, L. (2014). Univariate Stratification of Survey Populations. R package version. 2.2-7. Obtenido de <https://CRAN.R-project.org/package=stratification>
- Estevao, V., Hidiroglou, M., & Sarndal, C. (1995). Methodological Principles for a Generalized Estimation System at Statistics Canada. Journal of Official Statistics, 11(2), 181—204
- Estevao, V.M., and Särndal, C.-E. (2002). The ten cases of auxiliary information for calibration in two phase sampling. Journal of Official Statistics, 18, 233-255.
- Gutiérrez, H.A. (2016). Estrategias de muestreo, diseño de encuestas y estimación de parámetros. Universidad Santo Tomás. Bogotá (2016).
- Gutierrez, H. (2018). samplesize4surveys: Sample Size Calculations for Complex Surveys. Obtenido de: <https://cran.r-project.org/web/packages/samplesize4surveys/samplesize4surveys.pdf>
- Kozak, M. (2004). Optimal stratification using random search method in agricultural surveys. Statistics in Transition, 6(5), 797-806.
- Lavalley, P. and Hidiroglou, M.A. (1988). On the stratification of skewed populations. Survey Methodology, 14, 33-43.
- Lensvelt-Mulders, G. J. L. M., P. Lugtig, and M. Hubregtse. 2009. “Separating Selection Bias and Non-coverage in Internet Panels using Propensity Matching.” Survey Practice.
- Särndal, C-E. & Lundström, S. Estimation in Surveys with Nonresponse. Wiley. (2005).
- Särndal, C-E., Swensson, B., Wretman, J. Model Assisted Survey Sampling. Springer Verlag, New York. Trewin, D (2002).
- Tillé, Y. and Matei, A. (2022). Survey Sampling. R package version 2.9. Obtenido de: <https://cran.r-project.org/web/packages/sampling/sampling.pdf>
- Lumley, T. (2017). Survey: analysis of complex survey samples R package v. <http://CRAN.R-project.org/package=survey>
- Park, D., Gelman, A. & Bafumi, J., (2004). Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls. Political Analysis 12:375–385.

	Realizar levantamiento de información que permita continuar con el seguimiento periódico a la percepción ciudadana sobre temas estratégicos del Plan Nacional de Desarrollo.	 Departamento Nacional de Planeación - DNP
	PRODUCTO No. 3	

Potter F. J. (1990). A study of procedures to identify and trim extreme sampling weights. In: Proceedings of the Survey Research Methods Section, American Statistical Association, pp 225–230.

Potter F. J. (1993). The effect of weight trimming on nonlinear survey estimates. In: Proceedings of the Survey Research Methods Section, American Statistical Association, pp 758–763.

Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: Wiley.

Valliant, R., Dever, J., & Kreuter F. (2013). Practical Tools for Designing and Weighting Survey Samples-Springer.

Park, Gelman y Bafumi (2004). Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls.

	Realizar levantamiento de información que permita continuar con el seguimiento periódico a la percepción ciudadana sobre temas estratégicos del Plan Nacional de Desarrollo.	 Departamento Nacional de Planeación - DNP
	PRODUCTO No. 3	

4. ANEXOS

ANEXO 1. Estadístico.zip

ANEXO 2. Bases de datos.zip

ANEXO 3. Formulario VF.pdf

ANEXO 4. Formulario maestro.pdf

ANEXO 5. Manual del encuestador.pdf

ANEXO 6. Cronograma del proyecto.pdf

ANEXO 7. PND_2022-2026.nesstar_entrega 3.zip