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 1: Informe metodológico y primera entrega de la documentación de las bases de datos de la encuesta

ARD-649 DOC_01

Bogotá, D.C. Octubre 10 de 2023

SEI Soluciones Estratégicas en Información

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 1. Departamento Nacional de Planeación (DNP). Bogotá: Sistemas Especializados de Información S.A. - SEI

Tabla de Contenido

1. INFORME METODOLÓGICO 5
1.1. Introducción 5
1.2. Diseño temático 5
1.2.1. Variables e indicadores estadísticos. 5
1.2.2. Plan de resultados. 5
1.2.3. Estándares estadísticos utilizados. 6
1.2.4. Proceso de validación del cuestionario. 6
1.2.5. Fuente de información. 9
1.2.6. Normas que se aplicarán en la recolección y procesamiento de datos. 9
1.2.7. Clasificaciones y/o nomenclaturas. 10
1.2.8. Definición de términos o conceptos centrales de la encuesta. 11
1.3. Diseño estadístico 12
1.3.1. Descripción del universo de estudio. 12
1.3.2. Descripción de la población objetivo. 12
1.3.3. Cobertura y desagregaciones geográficas. 13
1.3.4. Desagregaciones temáticas, selección y tamaño de la muestra. 15
1.3.5. Descripción de la fuente primaria de los datos. 16
1.3.6. Unidades estadísticas. 16
1.3.7. Periodos de referencia. 17
1.3.8. Periodo de recolección. 17
1.3.9. Descripción y justificación del marco muestral. 17
1.3.10. Diseño muestral. 21
1.3.11. Especificación de ponderadores. 26
1.3.12. Metodología de la estimación. 29
1.3.13. Tamaño de muestra. 39
1.3.14. Precisión de los resultados. 42
1.3.15. Selección de la muestra. 49
1.3.16. Ajuste de cobertura. 49

1.4. Diseño del operativo de recolección en campo 61
1.4.1. Método, estrategia de recolección y dispositivo móvil de captura a usar. 64
1.4.2. Plan de prueba piloto. 67
1.4.3. Cronograma de recolección. 70
1.4.4. Estructura organizacional del operativo. 70
1.4.5. Plan de entrenamiento del personal. 72
1.4.6. Estrategias de sensibilización. 76
1.4.7. Seguimiento y control. 79
1.4.8. Sistemas de captura y georreferenciación de los datos recolectados. 81
1.4.9. Transmisión de datos. 83
1.5. Diseño del procesamiento de los datos 83
1.5.1. Consolidación de archivos de datos. 83
1.5.2. Codificación. 84
1.5.3. Diccionario de datos. 84
1.5.4. Revisión y validación de los archivos de datos. 85
1.5.5. Instrumentos de edición (validación y consistencia) e imputación de datos. 85
1.5.6. Cuadros de salida. 85
1.5.7. Definición de variables y construcción de indicadores. 85
2. PRIMERA ENTREGA DOCUMENTACIÓN DE BASES DE DATOS DE ENCUESTA 86
3. BIBLIOGRAFÍA 87
4. ANEXOS 89
ANEXO 1. RELACIÓN DE INDICADORES PRINCIPALES.pdf 89
ANEXO 2. FORMULARIO.pdf 89
ANEXO 3. ANEXO ESTADÍSTICO.zip 89
ANEXO 4. CRONOGRAMA DE RUTAS.pdf 89
ANEXO 5. CRONOGRAMA DEL PROYECTO.pdf 89
ANEXO 6. MANUAL ENCUESTADOR_ARD649.pdf 89
ANEXO 7. PND_2022-2026.Nesstar 89
ANEXO 8. GLOSARIO TERMINOLOGÍA EPC.pdf 89

1. INFORME METODOLÓGICO

1.1. Introducción

La medición periódica de la percepción ciudadana acerca de las apuestas incluidas en el Plan Nacional de Desarrollo se convierte en una importante herramienta para ajustar las políticas públicas teniendo en cuenta el sentir ciudadano en términos de ejecución y expectativas. Esta medición es de carácter nacional y se hace a través de la aplicación de encuestas presenciales, cara a cara, a hombres y mujeres de zonas urbanas y rurales de todo el país. Los datos obtenidos serán representativos a nivel nacional, departamental y regional, y por zona urbana y rural. En esta ocasión se visitarán más de 200 municipios[1] y se levantarán 13.400 encuestas, durante el mes de noviembre de 2023. Lo anterior, con el propósito de que el Gobierno nacional tenga la mayor cantidad de insumos para orientar sus estrategias y políticas, con el fin de lograr un mejor entendimiento y relación entre la sociedad civil y el ejecutivo.

El presente documento contiene las generalidades metodológicas de la operación, tanto a nivel temático, como estadístico, de recolección y procesamiento de los datos. Lo aquí planteado será el punto de partida para los preparativos de la recolección de información.

1.2. Diseño temático

1.2.1. Variables e indicadores estadísticos.

En el Anexo 1 se presenta una relación con algunos de los indicadores más relevantes por cada capítulo del formulario, con su respectiva descripción, fórmula de cálculo y forma de interpretación.

1.2.2. Plan de resultados.

Los resultados serán presentados en diferentes momentos y a través de diferentes medios. Es así como los resultados de la recolección podrán visualizarse en un dashboard que será actualizado permanentemente mientras se desarrolla el operativo en campo; adicionalmente se realizarán reportes periódicos de avance, y por último se entregarán los cuadros de salida con la información procesada, considerando las plantillas y contenidos acordados con el Comité Técnico de Seguimiento (CTS). En la sección 1.3.13. se describen además las medidas de calidad que acompañan a cada estimación puntual y se expone en detalle el contenido de los cuadros de salida. Es de anotar que, adicionalmente los resultados se presentarán en una plataforma de

[1] Como se explicará más adelante, en la primera etapa de muestreo se seleccionarán 238 UPM (Unidades Primarias de Muestreo).

visualización de datos con la información tanto de las estimaciones como de las estadísticas descriptivas de las diferentes variables.

1.2.3. Estándares estadísticos utilizados.

Al tratarse de una operación estadística dirigida a la población civil de 18 años o más, los lineamientos conceptuales de tipo demográfico guardan similitud con los utilizados por el DANE[2].

Con respecto a la nomenclatura para la identificación de las unidades administrativas, así como las de muestreo, se utilizará la división político-administrativa del país (DIVIPOLA), el Marco Geoestadístico Nacional (MGN) actualizado en 2021[3] y el inventario de veredas del Geoportal del DANE (vigencia 2020)[4], generado a partir del Censo Nacional Agropecuario (CNA), 2014, DANE.

Para la clasificación de la población en función de su participación en el mercado laboral se utilizarán las definiciones que el DANE usa en la Gran Encuesta Integrada de Hogares[5] (GEIH) para este fin.

En cuanto a las variables a medir, éstas se encuentran en el instrumento (encuesta) que hace parte de los anexos de este documento (Anexo 2) y sobre las unidades estadísticas, éstas se definen más adelante dentro de la sección de Diseño estadístico (numeral 1.3.).

Finalmente, para la documentación de la operación estadística y la disposición de los microdatos, se utilizará Nesstar Publisher, el cual permite (y es ampliamente utilizado, en particular por la Dirección de Seguimiento y Evaluación de Políticas Públicas (DSEPP) del Departamento Nacional de Planeación (DNP) documentar y publicar sus evaluaciones) realizar la documentación de operaciones estadísticas y migrarlas al portal centralizado llamado Archivo Nacional de Datos[6] (ANDA / Sinergia).

1.2.4. Proceso de validación del cuestionario.

En cuanto al proceso de validación y actualización del formulario este se ha desarrollado conforme lo planteado en la propuesta de SEI, del cual se citan en este documento los principales aspectos:

---

[2] https://www.dane.gov.co/index.php/estadisticas-por-tema/demogra a-y-poblacion/censo-nacional-de-poblacion-y-vivenda-2018/informacion-tecnica. Ver Manual de conceptos, junio de 2019, censo de poblaci n y vivienda, 2018.

[3] https://geoportal.dane.gov.co/servicios/descarga-y-metadatos/descarga-mgn-marco-geoestadistico-nacional/#gsc.tab=0, Descarga MGN Marco Geoestadistico Nacional, año 2021, DANE.

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

[5] chrome-extension://efaidnbmnnnibpcajpcglcle ndmkaj/https://www.dane.gov.co/ les/investigaciones/ chas/glosario_GEIH13.pdf

[6] https://anda.dnp.gov.co/index.php/catalog

- Lectura detallada del formulario para verificar que contiene las preguntas que cubren los aspectos a evaluar en sintonía con las apuestas del Plan Nacional de Desarrollo 2022-2026.
- Elaboración de nuevas preguntas y/o ajustes a las ya existentes.
- Revisión de preguntas de carácter sensible que puedan afectar el buen desarrollo de las encuestas, principalmente relacionadas con el conflicto armado.
- Revisión e inclusión de preguntas con enfoque diferencial e interseccional.
- Revisión de pases y flujos dentro de la encuesta.

De acuerdo con estos parámetros metodológicos el proceso se ha surtido como se describe enseguida.

El 26 de septiembre de 2023, y tal como se acordó en la reunión de inicio efectuada entre SEI y el CTS el día 25 de septiembre, el DNP socializó el contenido del formulario a la firma e inmediatamente después envío por correo electrónico una versión preliminar del formulario (con los ajustes discutidos en la reunión), que será aplicado en el levantamiento de la encuesta de percepción al PND 2022-2026, y el día 27 del mismo mes SEI envió la primera versión diagramada del formulario. El 28 de septiembre el CTS envió comentarios de la versión diagramada y el 29 del mismo mes se enviaron los ajustes a las observaciones y comentarios realizados.

Es importante mencionar que su contenido fue propuesto por la DSEPP con base en las preguntas aplicadas en los levantamientos de la Encuesta de Percepción Ciudadana (EPC) al PND 2018-2022, y a un ejercicio de identificación de necesidades de información adelantado al interior del DNP con las diferentes direcciones técnicas que tienen relación con los diferentes aspectos contenidos el PND.

Como resultado del ejercicio de validación se presentaron las siguientes novedades con respecto a la propuesta del formulario entregado por el DNP el 26 de septiembre:

- Ajuste al consentimiento informado.
- Inclusión de nuevas preguntas en el módulo de Pobreza.
- Ajustes de pases en el control operativo
- Inclusión de una pregunta sobre estado civil en el módulo de sociodemográficos.
- Ajuste de escalas en preguntas del módulo de Gobierno y en el de Recreación y Cultura.
- Ajuste en la redacción de algunas preguntas en los módulos de Barrios y Ambiente.
- Ajuste en instrucciones de leer y rotar opciones de respuesta en los módulos de Trabajo y Ambiente.

El día 29 de septiembre por correo electrónico, fue aprobado por el DNP con el fin de ser programado en el aplicativo de captura y posteriormente para ser piloteado (ver Anexo 2). La encuesta será aplicada de manera presencial (cara a cara) y mediante el uso de Dispositivos Móviles de Captura en un formulario que tiene la siguiente estructura:

Tabla 1 Estructura del formulario

CAPÍTULO	NOMBRE DEL CAPÍTULO	QUIEN RESPONDE
A	CONTROL OPERATIVO	ENCUESTADOR
IA.	SOCIODEMOGRÁFICOS –Matriz del hogar	PERSONA DEL HOGAR DE 18 AÑOS O MÁS QUE CONOZCA LA INFORMACION DE LOS OTROS MIEMBROS DEL HOGAR
IB.	DATOS DE LA PERSONA SELECCIONADA Y SOCIODEMOGRÁFICOS	PERSONA DE 18 AÑOS O MÁS SELECCIONADA POR EL SISTEMA
II.	CONECTIVIDAD E INTERNET	
III.	GOBIERNO Y GENERALES	
IV.	TRANSPARENCIA	
V.	POBREZA Y DESIGUALDAD	
VI.	SEGURIDAD	
VII.	NIÑOS, NIÑAS Y ADOLESCENTES	
VIII.	SITUACIÓN DE CAMPO	
IX.	JUSTICIA	
X.	EDUCACIÓN	
XI.	SALUD	
XII.	TRABAJO	
XIII.	CULTURA Y RECREACIÓN	
XIV.	INFRAESTRUCTURA Y TRANSPORTE	
XV	MINAS Y ENERGÍA	
XVI.	BARRIOS, VIVIENDA Y SERVICIOS PÚBLICOS	
XVII	AMBIENTE	
XVIII	PAZ Y MIGRACIONES	
XIX	OTROS TEMAS ESTRATÉGICOS DEL GOBIERNO NACIONAL	
XX	EXCLUSIVO PARA EL ENTREVISTADOR	ENCUESTADOR

Fuente. SEI. Formulario EPC 2023.

1.2.5. Fuente de información.

El informante idóneo definido metodológicamente desde la concepción de esta encuesta es cualquier residente habitual del hogar, seleccionado bajo las características que determina el diseño muestral, lo cual implica una selección probabilística, de manera que el aplicativo de muestra tendrá en consideración a todos los miembros del hogar para hacer esa selección. Dicho informante deberá ser igual o mayor de 18 años y deberá estar en capacidad de responder la encuesta por sus propios medios, sin asistencia de otros. Al inicio del formulario se tienen unas preguntas filtro que buscan descartar de la selección a quienes por avanzada edad, condiciones de salud o discapacidad cognitiva no se encuentran en condiciones de responder de forma autónoma- Ahora bien, para aquellos casos en los que se encuentren personas con discapacidad auditiva no se hará su exclusión siempre que se cuente con un intérprete que le asista en su hogar. Esto es debido a que logísticamente no es viable contar con un intérprete de señas en cada equipo de recolección, y resulta difícil planear la asistencia de un intérprete de la consultoría.

Para minimizar el desgaste de la fuente de información, esta consultoría comparte los criterios sugeridos por el DANE, y adoptados por el equipo del DNP, en cuanto a la necesidad de ajustar el tamaño del cuestionario a una duración que resulte práctica tanto para el informante como para el encuestador, en adoptar las soluciones tecnológicas para la recolección de datos, y en aplicar criterios estadísticos para la selección de la muestra, además de llevar a cabo rigurosamente los procesos de validación de los procedimientos y del instrumento del encuesta, con el propósito de lograr un adecuado acercamiento con las herramientas necesarias que permitan que fluya la aplicación y por ende, todo el proceso de recolección.

1.2.6. Normas que se aplicarán en la recolección y procesamiento de datos.

Previo al diseño del aplicativo de captura de la información, el equipo operativo de SEI elabora una malla de validación que permite controlar las reglas de diligenciamiento (pases, saltos, rangos permitidos), con el fin de hacer una captura inteligente del formulario de la encuesta.

Esta malla de validación permite programar que el aplicativo solo acepte valores válidos y haga cruces entre preguntas, además que emita alertas al encuestador para indicar que la información registrada es inconsistente o debe ser verificada. Basados en esta malla, se programan cada una de las restricciones en el aplicativo, y de esta forma, al ingresar un registro se controla en cada variable, los rangos y la información inconsistente, dando como resultado la revisión en tiempo real (apenas se están ingresando los datos), y por ende la consistencia de la información.

Un ejemplo de validación sucede cuando se ingresa el número celular, pues la malla cuenta con un rango que solicita el ingreso de 10 dígitos. Si el encuestador ingresa

solo 9 números el sistema no lo deja pasar y le genera un aviso de alerta para que verifique la información ingresada, e igual sucede si ingresa más de 10 números. De esta manera se garantiza que la depuración sucede en tiempo real y con esto se evitan reprocesos.

Es de anotar que el programa no permite pasar a otra pregunta sin marcar una respuesta válida, con lo cual se elimina la posibilidad de imputar datos faltantes, y para asegurar que funcione correctamente se realizan pruebas antes de la capacitación al personal.

En simultánea se realiza un análisis de frecuencias simples, como un mecanismo de control y verificación de la consistencia de la información. Estos procesos de depuración se justifican porque permiten asegurar la entrega de tablas con datos confiables.

1.2.7. Clasificaciones y/o nomenclaturas.

Para el registro de direcciones se tendrán en consideración los siguientes parámetros, basados en lo determinado por el DANE para el registro completo de direcciones, haciendo uso de nomenclatura estandarizada.

- La nomenclatura incluye la denominación de la vía (calle, carrera, avenida) sobre la que se encuentra y el número de placa asignado al acceso.
- Si la vivienda tiene varias placas, debe registrar todas como parte de la dirección.
- Después de la vía (calle, carrera, avenida) se escribe el número de la vía principal y la placa de la vivienda, sin utilizar símbolos como #, No., punto, coma o guion.
- Si la orientación (Este, Sur, etc.) forma parte de la nomenclatura, debe incluirse la palabra completa.
- Debe incluir todos los datos que identifican la vivienda, como número del edificio, bloque, interior, número de apartamento, etc.
- Si no existe nomenclatura, es necesario registrar datos o puntos de referencia fijos, que permitan ubicar la vivienda, como por ejemplo lugares de fácil reconocimiento como instituciones.

Las abreviaturas válidas para registrar las direcciones se presentan en la Tabla 2.

De las abreviaturas presentadas, las resaltadas con sombra corresponden a la identificación de clases de vías. Toda dirección debe comenzar por cualquiera de estas opciones.

Tabla 2. Abreviaturas válidas para registrar direcciones

ABREVIA	UNIDAD	ABREVIA	UNIDAD	ABREVIA	UNIDAD
AD	Administración	***CV***	Circunvalar	***MN***	Mezzanine
AG	Agrupación	***CC***	Cuenta Corrida	***MD***	Módulo
AL	Altillo	***DPTAL***	Departamental	***MPAL***	Municipal
AP	Apartamento	***DP***	Depósito	***OF***	Oficina
AV	Avenida	***DG***	Diagonal	***PQ***	Parqueadero
AC	Avenida Calle	***DS***	Depósito sótano	***PA***	Paseo
AK	Avenida Carrera	***ED***	Edificio	***PH***	Penthouse
AU	Autopista	***EN***	Entrada	***PI***	Piso o planta
B	Barrio	***ES***	Escalera	***PT***	Portería
BQ	Bloque	***ESC***	Escuela	***PU***	Puesto
BG	Bodega	***ET***	Etapa	***PS***	Puesto de salud
CL	Calle	***FCA***	Finca	***SA***	Salón
CN	Callejón	***GJ***	Garaje	***SECT***	Sector
CM	Camino	***GR***	Grupo	***SS***	Semisótano
KR	Carrera	***HDA***	Hacienda	***SOT***	Sótano
CT	Carretera	***HPTAL***	Hospital	***SM***	Supermanzana
CA	Casa	***INST***	Instituto	***TZ***	Terraza
CRIO	Caserío	***IN***	Interior	***TO***	Torre
CE	Célula	***KM***	Kilómetro	***TV***	Transversal
CS	Centro de salud	***LM***	Local Mezzanine	***UN***	Unidad
COL	Colegio	***LC***	Local	***UR***	Unidad Residencial
CO	Conjunto	***LT***	Lote	***URB***	Urbanización
CONS	Consultorio	***MZ***	Manzana	***VDA***	Vereda
CGTO	Corregimiento	***ESQ***	Esquina	***ZN***	Zona

Fuente: SEI

1.2.8. Definición de términos o conceptos centrales de la encuesta.

En el Anexo 8 se presentan los términos empleados en el formulario de encuesta, organizados por capítulo, y con base en lo sugerido por el Sistema de Consulta de Conceptos Estandarizados del DANE (https://conceptos.dane.gov.co/conceptos/area_tematica).

1.3. Diseño estadístico

Para inferir resultados con validez estadística a la población de referencia, es decir, de manera confiable y precisa, se implementará con rigor técnico un método de muestreo probabilístico, lo cual exige: un marco de muestreo en las distintas etapas y que, en cada una de éstas, a nivel de los elementos, la asignación de las probabilidades de inclusión sea conocida, mayor a cero y que los mecanismos de selección implementados condicionen las probabilidades asignadas a priori. Garantizando lo anterior y teniendo como base los modelos de probabilidad teóricos, establecidos según el diseño a partir de la muestra observada, los resultados tendrán la validez que requiere el estudio.

1.3.1. Descripción del universo de estudio.

El universo de estudio está conformado por 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), 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). Según la información oficial del DANE, la población total residente en el país y proyectada a junio de 2023 es de 52.215.503[7], de la cual el 76,13 % (39.751.902) se encuentra localizada en la zona urbana y el restante en la zona rural (centros poblados o veredas); el 51,19 % (26.729.489) son mujeres y el restante (25.486.014, es decir, el 48,81 %) son hombres.

1.3.2. Descripción de la población objetivo.

Con respecto a la población objetivo, son las personas de 18 años o más y se estima en 38.195.238 personas (73,15 %), 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. Ahora, por nivel socioeconómico (estrato según el servicio de Energía), la distribución relativa aproximada de la población, según la Superintendencia[8] 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 %.

[7] https://www.dane.gov.co/index.php/estadisticas-por-tema/demografia-y-poblacion/proyecciones-de-poblacion, con base en el censo de población y vivienda 2.018, DANE. Tabla "ACTUALIZACIÓN POST COVID-19. PROYECCIONES DE POBLACIÓN MUNICIPAL POR ÁREA, SEXO Y EDAD." "PROYECCIONES DE POBLACIÓN A NIVEL MUNICIPAL. PERIODO 2020 - 2035.".

[8] 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".

En la Figura 1 se observa la distribución espacial de los municipios del país, según las 33 entidades territoriales (departamentos) de interés que se encuentran en el anexo técnico del contrato.

1.3.3. Cobertura y desagregaciones geográficas.

La cobertura geográfica del estudio, la cual según DANE[9] "Es la extensión territorial sobre la cual se ejecuta una operación estadística. Como ejemplo se puede mencionar: cobertura nacional, regional o departamental.", corresponde al área geográfica que es incluida dentro del marco muestral en sus distintas etapas para garantizar que contiene la población objetivo, y está conformada por 1.102 municipios y 20 áreas no municipalizadas que se encuentran dentro de 33 entidades territoriales (32 departamentos más Bogotá D.C) objeto de estudio.

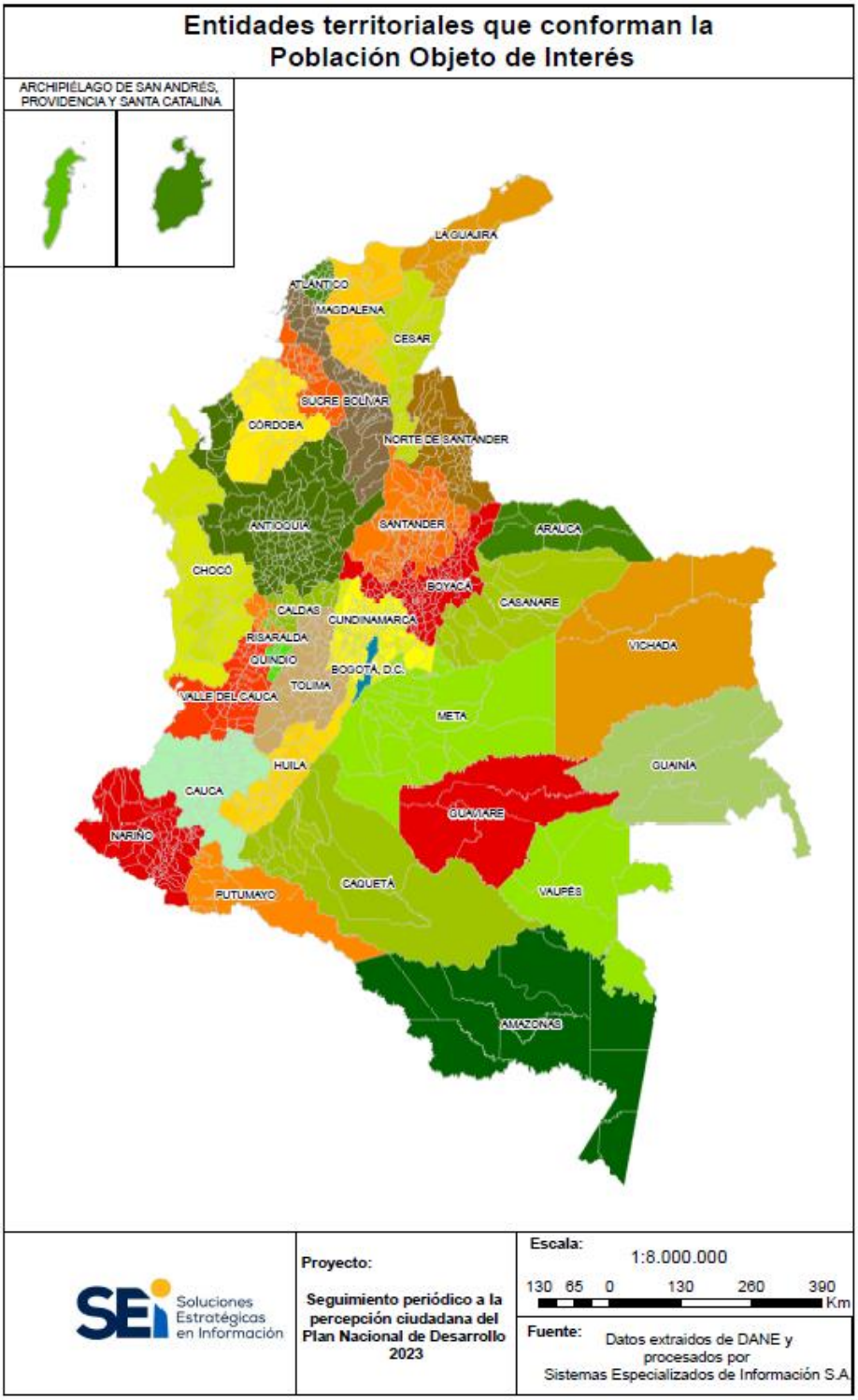
De otra forma, la operación estadística definida cubre todo el territorio nacional (continental e insular) y en él, se investigan todos y cada uno de los elementos que conforman la población objetivo (personas de 18 años o más), de tal forma que se obtiene información válida en términos estadísticos (precisa y confiable) para inferir resultados con desagregación geográfica en el nivel de entidades territoriales, principalmente, así como de todos los dominios de análisis propuestos.

En cuanto a los dominios de análisis objeto de interés estos son[10]:

- 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 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, incluida San Andrés (se excluyen las 7 entidades territoriales principales) por clase (1. Cabeceras, 2. Centros poblados y 3. Rural disperso).

[9] "chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.dane.gov.co/files/sen/normatividad/Lineamientos_Proceso_Estadistico.pdf."
[10] 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.

Figura 1. Distribución de los municipios que conforman las 33 entidades territoriales de interés.



Fuente: Según División Política y Administrativa, DANE, agosto 2023.

- Regiones[11] definidas[12] como aglomeraciones de departamentos y de sus áreas metropolitanas: 1. Caribe[13], 2. Santanderes[14], 3. Eje Cafetero y Antioquia[15], 4. Orinoquía[16], 5. Amazonía[17], 6. Central[18], 7. Pacífica[19], 8. Insular[20];
- Capitales de departamento y/o principales áreas metropolitanas;
- Clúster de categorías (departamentos o municipios) según la Contraloría General de la Nación[21] ("CT01 - Categorización de Municipios 2023" y "Categorización por Departamentos 2023", CGN, 2023);
- Clúster de capitales por región y zona.
- Municipios PDET.

Para los dominios relacionados con la entidad territorial, la zona y el total nacional, como se especifica más adelante, los márgenes de error para un parámetro tipo proporción del 50 % y con un nivel de confianza del 95 %, corresponden a los definidos por la entidad en el anexo técnico del contrato.

1.3.4. Desagregaciones temáticas, selección y tamaño de la muestra.

Se mantienen (también para fines comparativos) los dominios de análisis considerados en los levantamientos anteriores (EPC PND 2014-2018 y EPC PND 2018-2022), relacionados con características socioeconómicas. Específicamente[22]:

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

Además, se incluyen algunas preguntas relacionadas con el enfoque diferencial e interseccional[23] que complementen aquellas inicialmente consideradas sobre el tema

[11] 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.
[12] *"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.
[13] Atlántico (sin A.M de Barranquilla), Bolívar, Cesar, Córdoba, La Guajira, Magdalena y Sucre.
[14] Norte de Santander y Santander,
[15] Caldas, Quindío, Risaralda y Antioquia (sin A.M del Valle de Aburra).
[16] Arauca, meta, Casanare y Vichada.
[17] Vaupés, Guainía, Amazonas, Guaviare, Caquetá y Putumayo.
[18] Cundinamarca, Boyacá, Huila y Tolima.
[19] Cauca, Chocó, Nariño y Valle del Cauca (sin A.M de Cali).
[20] Archipiélago de San Andrés, Providencia y Santa Catalina.
[21] 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).
[22] https://anda.dnp.gov.co/index.php/catalog/151/study-description.
[23] chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.dane.gov.co/files/investigaciones/genero/guia-inclusion-enfoque-difencias-intersecciones-produccion-estadistica-SEN.pdf. DANE, 2020.

(zona, sexo y ciclo de vida) en los levantamientos anteriores, lo cual permite analizar los resultados considerando las "brechas" preexistentes a través de un enfoque inclusivo, tal como lo recomienda el Sistema Nacional Estadístico (SEN) de Colombia[24]. Estas preguntas se relacionan con:

- Discapacidad;
- Víctima según RUV;
- Grupo étnico;
- Orientación sexual e identidad de género;
- Estado conyugal;
- Migración;
- Parentesco;
- Nivel educativo;
- Acceso a la salud.

Lo anterior, está condicionado a que los márgenes de precisión observados (error estándar – E.E o coeficiente de variación estimado – c.v.e.[25]) sean tolerables, en concordancia con los parámetros acordados con el CTS del proyecto de la Dirección de DSEPP del DNP y con el supervisor del contrato.

1.3.5. Descripción de la fuente primaria de los datos.

Para la operación estadística por muestreo probabilístico se esperan aplicar 13.400 encuestas en 238 municipios o áreas metropolitanas (más adelante esto se relacionará con las Unidades Primarias de Muestreo[26] - UPM.) La información primaria es recopilada a través de una encuesta de percepción ciudadana cuya temática aborda los principales ejes del PND 2022-2026. Este instrumento hace parte de los anexos de este documento metodológico.

1.3.6. Unidades estadísticas.

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.

[24] https://www.sen.gov.co/conozca-el-sen/que-es-el-sen
[25] En el levantamiento anterior, los niveles de calidad se determinaron considerando el c.v.e, así: < 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. c.v.e.> 20,0 % se asocia a estimaciones imprecisas.
[26] Corresponde a municipios o áreas metropolitanas.

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 definen como:

- 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).

1.3.7. Periodos de referencia.

En el instrumento (encuesta) se abordan los siguientes períodos de referencia: 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 en el lapso del trabajo de campo.

1.3.8. Periodo de recolección.

El levantamiento de información primaria a través de la encuesta de percepción ciudadana se realizará entre octubre y diciembre de 2023.

1.3.9. Descripción y justificación del marco muestral.

En el contexto de la literatura especializada de muestreo, la implementación de cada etapa de un diseño probabilístico demanda su propio marco de muestreo, que contenga la información actualizada, así como necesaria y suficiente, relacionada con la identificación y localización (así como de contacto) de las unidades de muestreo objeto de selección.

De esta manera, para la selección de los municipios se tuvo en cuenta la división político-administrativa del país (DIVIPOLA), la proyección de población 2018-2026[27] 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 2021[28], y para la conformación del marco muestral de veredas se utilizó el inventario de veredas del Geoportal del DANE (vigencia 2020)[29], generado a partir del Censo Nacional Agropecuario (CNA), 2014, DANE. Así mismo, de ser necesario, se gestionará la cartografía generada por los entes territoriales municipales (mapa veredal, solicitado en cada alcaldía 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 correspondiente a las manzanas cartográficas en la cabecera y centros poblados, se utilizó el MGN.

Ahora, al interior de cada manzana cartográfica seleccionada se realizará un recuento de viviendas[30] (listado exhaustivo de viviendas ocupadas) para seleccionar una muestra de éstas y, en cada una, identificar mediante un recuento los hogares[31] y seleccionar uno. Finalmente, en cada hogar seleccionado, se identificarán y listarán los residentes habituales para identificar, 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 implementará 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.

La descripción anterior, destaca la relevancia técnica que tiene dentro del esquema probabilístico la construcción de los marcos de muestreo en cada etapa definida, para lo cual se hizo necesario articular diferentes fuentes de información, con el fin de garantizar su integridad. En este contexto, es importante enfatizar el valor de la cartografía como herramienta de diseño, planificación y ejecución de las tareas relacionadas con los marcos de muestreo.

---

[27] 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.**

[28] https://geoportal.dane.gov.co/servicios/descarga-y-metadatos/descarga-mgn-marco-geoestadistico-nacional/#gsc.tab=0, Descarga MGN Marco Geoestadistico Nacional, año 2021, DANE.

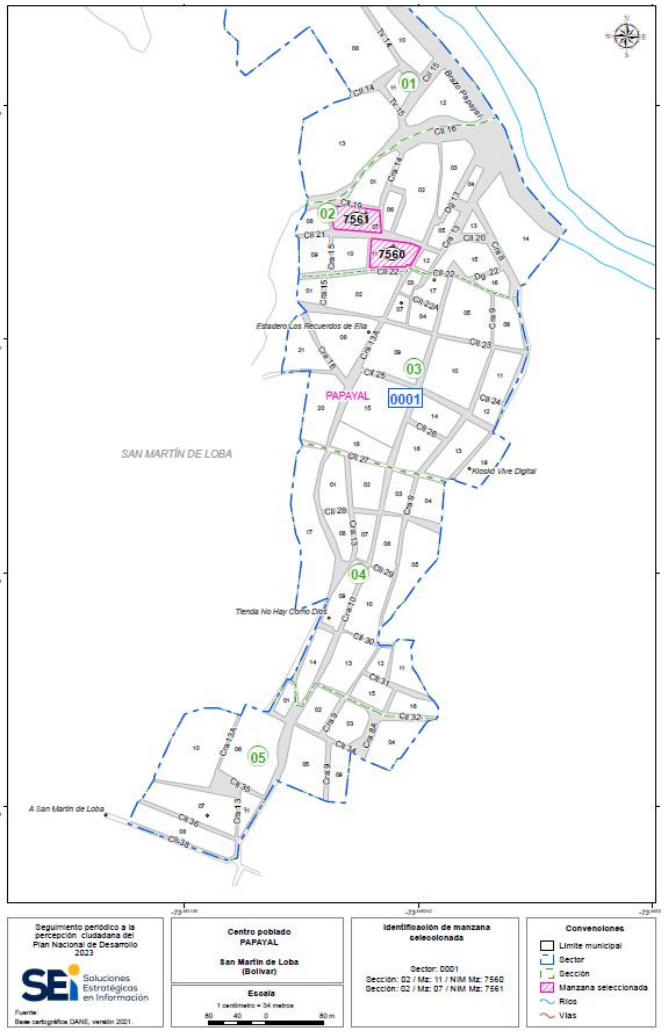
[29] https://geoportal.dane.gov.co/servicios/descarga-y-metadatos/descarga-nivel-de-referencia-de-veredas/#gsc.tab=0, año 2020, DANE.

[30] Espacio independiente y separado, habitado o destinado para ser habitado por una o m s personas.

[31] Es una persona o grupo de personas, parientes o no, que ocupan la totalidad o parte de una vivienda; atienden necesidades b sicas con cargo a un presupuesto com n y generalmente comparten las comidas.

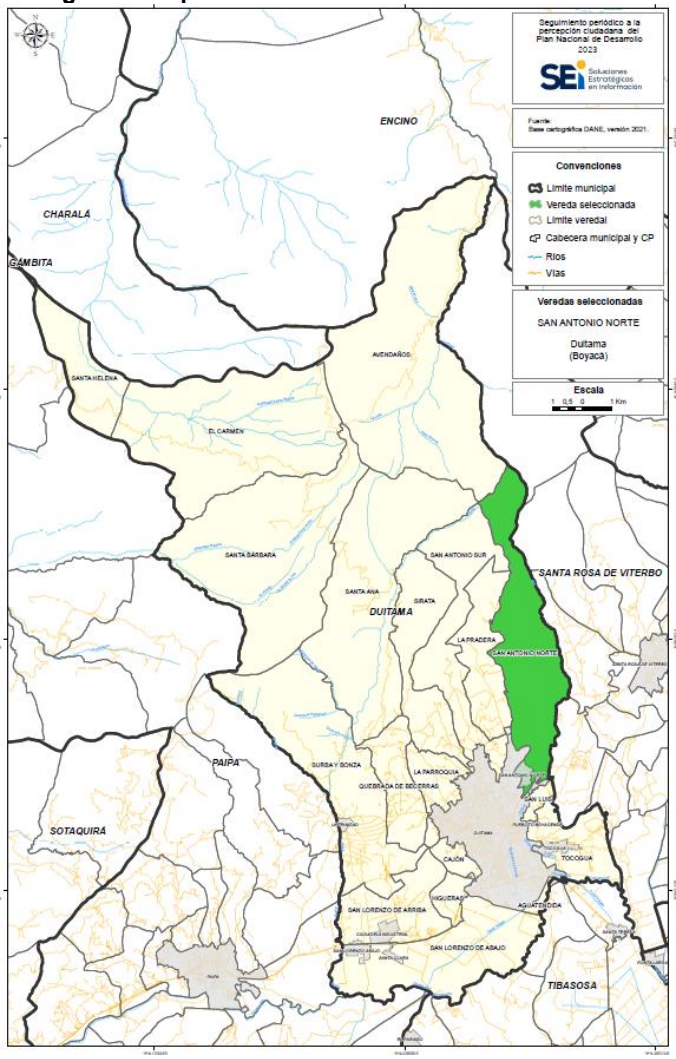
La georreferenciación de manzanas en las cabeceras municipales, así como en los centros poblados y, la cartografía de las unidades rurales (centros poblados y veredas), además de la identificación de las viviendas dentro de segmentos en una vereda, a partir de un croquis a mano alzada construido con la orientación de un líder veredal, permite que el personal de campo localice de manera exacta las unidades de muestreo. Contar con este insumo, además facilita la planeación logística, lo cual impacta en un operativo de recolección eficiente en términos de la optimización de recursos y facilita la supervisión al personal encargado del operativo. Al respecto, en las siguientes figuras (2 y 3) se presentan ejemplos de los insumos que se generarían para la planeación y ejecución del trabajo de campo.

Figura 2. Mapa de manzanas cartográficas seleccionadas en un centro poblado.



Fuente: Elaboración propia, con base en el MGN 2021, DANE.

Figura 3. Mapa de una vereda seleccionada.



Fuente: Elaboración propia, con base en el MGN 2021, DANE.

Importante resaltar que el recuento surge de la necesidad de construir un marco de muestreo de viviendas y hogares actualizado, que refleje las condiciones de la población durante el período de recolección de la información primaria (Trabajo de campo), tanto en el nivel de las manzanas (cabecera y centros poblados), como en el de las veredas y segmentos seleccionados. Dicho procedimiento (recuento o inventario) es el resultado de un proceso de listar las edificaciones y unidades de vivienda que se adelanta para trabajar con información actualizada que no está disponible dentro de las fuentes oficiales, lo cual permite garantizar que la selección probabilística de las unidades de muestreo en las últimas etapas se realice con base en las condiciones requeridas por el método probabilístico.

La parte logística y operativa del recuento consiste en un recorrido exhaustivo por todos los lados de la manzana seleccionada, con el objetivo de identificar, por observación, las unidades de vivienda con uso exclusivo residencial o uso mixto (con hogares). Este procedimiento se realizó teniendo en cuenta los lineamientos técnicos que, por ejemplo, utiliza el DANE para hacerlo en sus operaciones estadísticas.

Para este fin, dentro del aplicativo instalado en los Dispositivos Móviles de Captura (DMC), con el que se realiza la captura y validación en tiempo real de la encuesta, se registran las técnicas de supervisión y se controla la cobertura e implementación del diseño muestral probabilístico en las últimas etapas de muestreo. Dentro del módulo de administración de la muestra para los recuentos en dicho aplicativo, se programa el algoritmo de selección planteado en el diseño muestral (Coordinado negativo); dicho algoritmo cumple estrictamente los parámetros de identificación de las unidades de selección en las últimas tres etapas de muestreo, se almacenan todas las variables involucradas en el algoritmo para su ejecución, y esto permite hacer seguimiento y replicar.

En resumen, los encuestadores ingresan la información y el aplicativo se encarga de almacenarla para estructurar el marco de muestreo, y luego, de ejecutar la selección, genera un reporte con las viviendas/hogares/personas a encuestar. Dicho algoritmo está diseñado y programado para ejecutarse una única vez, por lo tanto, el resultado de la selección no puede ser alterado, ajustado o manipulado por el encuestador ni por nadie. Al respecto, la evidencia de los aleatorios con distribución Uniforme (0,1) y la “marca” de la persona seleccionada se puede replicar a partir de la semilla que se guarda en una tabla cifrada dentro del DMC, para validar el procedimiento.

1.3.10. Diseño muestral.

a) Características del diseño muestral probabilístico, la estratificación y el esquema definido para cada etapa[32].

Para este estudio se definió una operación estadística a partir de un 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 2022-2026[33] tenga un alcance inferencial.

[32] El contenido de esta sección se desarrolla según lo mencionado en la página 20 de: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.dane.gov.co/files/sen/lineamientos/Guia-elaboracion-documento-metodologico.pdf

[33] “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 alimentaci n; iv) Transformaci n productiva, internacionalizaci n y acci n clim tica; v) Convergencia regional.

Dicha metodología profundiza los aspectos técnicos relacionados con la operación estadística por muestreo probabilístico, teniendo como referente los Lineamientos del Proceso Estadístico en el Sistema Estadístico Nacional,[34][35], así como el anexo técnico del contrato. Adicionalmente, para esta sección se incluye un anexo estadístico (Anexo 3), con los detalles computacionales y salidas de los resultados comentados, con el fin de responder al requerimiento de reproducibilidad solicitado.

Esta operación estadística guarda relación con 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.

En el contexto de lo anterior, el diseño probabilístico para la operación estadística descrita se define como: multietápico y estratificado (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 conforman 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 utilizará un algoritmo que se describe más adelante.

En este contexto, se reitera cuáles son las etapas y unidades de muestreo, posteriores a la primera etapa (Unidad Primaria de Muestreo – UPM) y que corresponden a un municipio o área metropolitana):

- **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).

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

[35] chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.dane.gov.co/files/sen/lineamientos/Guia-elaboracion-documento-metodologico.pdf

b) Proceso para la conformación de los estratos según etapa de muestreo (variables y algoritmos).

Con respecto a los niveles de estratificación, el procedimiento implementado fue:

- Inicialmente, se construyeron 33 estratos, uno por cada entidad territorial[36] (departamento más Bogotá D.C), para la selección de las Unidades Primarias de Muestreo (U.P.M), es decir, municipio o área metropolitana;
- Los municipios de Bogotá D.C, San Andrés y Providencia-Santa Catalina, son U.P.M. de inclusión forzosa para las entidades territoriales Bogotá D.C y Archipiélago de San Andrés, Providencia y Santa Catalina;
- En el resto de las entidades territoriales (en total, 31) las U.P.M. se clasifican en estratos (Nivel II) con base en el algoritmo mencionado más adelante para este fin y la variable auxiliar "Total población de 18 años o más proyectada a junio de 2023";
- Todas las capitales (excepto San José de Guaviare) son de inclusión forzosa. Ahora, en el caso de Antioquia, Valle del Cauca, Atlántico y Santander, además de las capitales, los municipios de las áreas metropolitanas también fueron de inclusión forzosa[37]. Por último, en el caso de Magdalena se incluyó Santa Marta y Fundación, y en el de Cundinamarca, Madrid, Mosquera, Chía, Funza, Facatativá y Soacha, como UPM de inclusión forzosa.
- Al interior de cada U.P.M seleccionada se estratificó por zona urbana o rural (Nivel III);

En la zona urbana de las U.P.M:

- Las manzanas cartográficas se estratificaron (Nivel IV) de acuerdo con la cantidad de viviendas ocupadas reportadas según censo de población y vivienda 2018, DANE;
- En cada manzana se realizó un recuento para identificar viviendas ocupadas;
- Dentro de la muestra de viviendas, se construyó un marco de hogares y se seleccionó uno;
- En cada hogar se identificaron sus residentes habituales y, con base en esto, se seleccionó una persona de 18 años o más:

En la zona rural se estratificará por clase (2: centro poblado y 3: rural disperso o vereda):

- Se seleccionó una muestra de centros poblados, de acuerdo con la información que tiene el DANE sobre esta unidad administrativa; en cada centro poblado se

[36] Cada entidad territorial, además de ser un estrato, es un dominio de análisis.
[37] Más adelante, en la Tabla 4 "", se indica qué municipios fueron incluidos forzosamente por departamento. Este aspecto se relaciona con la intención de garantizar (es un resultado) los márgenes de error para el indicador trazador y nivel de confianza solicitado en el anexo técnico del contrato por entidad territorial.

seleccionó una muestra de manzanas, con base en el MGN 2021; en cada manzana se seleccionó una muestra de viviendas; en cada vivienda, un hogar; en cada hogar, una persona de 18 años o más.

- Se seleccionó una muestra de veredas, de acuerdo con la información que sobre esta unidad administrativa tiene el DANE; en cada vereda, con la ayuda de un líder social, se construyó un croquis a mano alzada para identificar viviendas ocupadas y sus jefes de hogar para, con base en esto, construir segmentos entre 3 y 12 viviendas (en promedio, 10 viviendas) y seleccionar una muestra de estos; todas las viviendas del segmento (conglomerado) fueron visitadas para identificar TODOS los hogares existentes y en cada uno de estos, seleccionar una persona de 18 años o más.

El algoritmo que se utilizó 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) fue propuesto por Lavallée, P. and Hidiroglou, M.A. (1988)[38], desarrollado por Kozak, M. (2004)[39] e implementado en el software libre R por Sophie Baillargeon and Louis-Paul Rivest (2014)[40]. Luego, a partir de estos resultados, y con el fin de controlar la cantidad de estratos por departamento, así como mantener la eficiencia del diseño, en algunos casos se redujo (colapsando) su cantidad (de estratos). En el anexo estadístico (Anexo 3) se encuentra el código construido para este fin.

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). Lo anterior, se detallará más adelante.

La asignación del tamaño de muestra de unidades finales para el Nivel I de estratificación (entidad territorial) se realizó proporcionalmente al total poblacional proyectado, y se fue ajustando por zona, a partir de los escenarios que se generaron a través de la simulación, para encontrar el óptimo con el cual garantizar los márgenes de error requeridos según el anexo técnico para: 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.

Por otra parte, es pertinente resaltar que el diseño probabilístico sugerido parte de una estrategia eficiente[41] (en términos de sesgo y mínima varianza), para un estudio

[38]Lavallée, P. and Hidiroglou, M.A. (1988). On the stratification of skewed populations. Survey Methodology, 14, 33-43.
[39]Kozak, M. (2004). Optimal stratification using random search method in agricultural surveys. Statistics in Transition, 6(5), 797-806.
[40]Univariate Stratification of Survey Populations. R package version. 2.2-5. https://CRAN.R-project.org/package=stratification. Usando la función strata. LH

[41] "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…

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.

c) Métodos de muestreo y algoritmos de selección que se aplicarán en cada una de las etapas o fases.

Según el diseño muestral definido, se implementó el método de muestreo aleatorio simple (M.A.S.) sin reemplazamiento, con base en el algoritmo de selección Coordinado Negativo. Este algoritmo de selección probabilística se propone también como estrategia para incluir unidades de muestreo de reemplazo (las relacionadas con municipios, centros poblados, veredas y manzanas cartográficas), para mitigar la no respuesta en caso de ser necesario (debido a las imperfecciones del marco, problemas de orden público, o para mitigar el efecto de la no respuesta en relación con la ausencia o rechazo de las unidades finales de muestreo – personas seleccionadas), sin generar sesgos y sin necesidad de implementar una selección condicional o por fases, ya que se sigue el orden de inclusión aleatorio determinado desde la selección inicial realizada con una semilla y entregada al CTS antes de iniciar el trabajo de campo.

Tanto el método de muestreo como el algoritmo citado, documentados en Gutiérrez (2016), se implementaron en la función strata del software libre estadístico R versión 4.1.1, del paquete sampling (Tillé, Y. & Matei, A., 2022).

Al respecto, el método y algoritmo consisten en:

M.A.S. (Bautista, 1998):

- Se selecciona aleatoriamente y sin repetición, un conjunto de n elementos dentro de los N del universo;
- Todos los elementos tienen igual probabilidad de ser seleccionados: para el primero seleccionado 1/N;

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.

- Para el segundo elemento, el elemento extraído tiene probabilidad 0, y los restantes tienen probabilidad 1/(N−1), y así sucesivamente hasta obtener los n elementos deseados.
- Finalmente, todas las muestras tienen probabilidad igual a: $1/\binom{N}{n}$.

Coordinado Negativo:

- Para cada elemento de la población se genera un número aleatorio r de la distribución uniforme (0,1).
- Se ordenan de menor a mayor los elementos respecto a los aleatorios previamente generados.
- Se seleccionan los primeros n elementos.

Reiteramos que para su implementación se definió un código en R Studio incorporando una semilla en función de la fecha del acta de inicio del contrato (“210923”), con el fin de que los expertos puedan replicar los ejercicios de selección, los cuales son parte del anexo estadístico (Anexo 3).

1.3.11. Especificación de ponderadores.

Como se implementó un diseño probabilístico estratificado y multietápico, en el cual se implementaron en todas las etapas de muestreo el método de M.A.S., las probabilidades de inclusión generales toman la siguiente forma:

M.A.S (Unidades de muestreo en etapas siguientes)

$$\pi_k = \frac{n}{N}$$

- n: es la cantidad de unidades seleccionadas;
- N: es la cantidad de unidades en la población.

Específicamente las probabilidades de inclusión toman la siguiente forma en cada etapa:

Probabilidad de inclusión por etapa según método (UPM):

En la primera etapa, se hace selección de UPM con un diseño M.A.S. de manera tal que la probabilidad de inclusión queda expresada como:

$$\pi_{I,khr} = \frac{n_{I,h,r}}{N_{I,h,r}}$$

Fuente: Con base en Bautista, 1998

Donde $N_{I,h,r}$ corresponde al total de total de municipios existentes del estrato h de la entidad territorial r y $n_{I,h,r}$corresponde al total de municipios seleccionados en el estrato h de la entidad territorial r.

Ahora, manteniendo la misma estructura de probabilidad de inclusión que se menciona a nivel general para un muestreo M.A.S, a continuación, se presentan las probabilidades de inclusión para la segunda a la sexta etapa.

Probabilidad de inclusión veredas / centros poblados / área urbana (Segunda etapa)	**Probabilidad de inclusión de manzanas / segmentos (Tercera etapa)**	**Probabilidad de inclusión de viviendas (Cuarta etapa)**	**Probabilidad de inclusión de hogares (Quinta etapa)**
$\pi_{II,j} = \frac{n_{z,k,h,r}}{N_{z,k,h,r}}$. Con: - $\pi_{II,j} = 1$, en la zona urbana. - $n_{z,k,h,r}$: cantidad de veredas/ centros poblados (CP) seleccionados en la zona z, del municipio k que pertenece al estrato h de la entidad territorial r. - $N_{z,k,h,r}$: cantidad de veredas/ centros poblados (CP) identificados según marco, en la zona z, del municipio k que pertenece al estrato h de la entidad territorial r.	$\pi_{III,l,1} = \frac{n_{j,z,k,h,r}}{N_{j,z,k,h,r}}$ Con: - $n_{j,z,k,h,r}$: cantidad de manzanas/segmentos seleccionados en la vereda/CP/ área urbana j, de la zona z, del municipio k que pertenece al estrato h de la entidad territorial r. -$N_{j,z,k,h,r}$: cantidad de manzanas/segmentos identificados según marco en la vereda/CP/ área urbana j, de la zona z, del municipio k que pertenece al estrato h de la entidad territorial r.	$\pi_{IV,m} = \frac{n_{l,j,z,k,h,r}}{N_{l,j,z,k,h,r}}$ Con: - $n_{l,j,z,k,h,r}$: número de viviendas seleccionadas en la manzana l, del CP/área urbana j, de la zona z, del municipio k que pertenece al estrato h de la entidad territorial r. - $N_{l,j,z,k,h,r}$: total de viviendas identificadas según marco en la manzana l, del CP/ área urbana j, de la zona z, del municipio k que pertenece al estrato h de la entidad territorial r. - $\pi_{IV,m} = 1$, para la zona de veredas.	$\pi_{V,f} = \frac{n_{m,l,j,z,k,h,r}}{N_{m,l,j,z,k,h,r}}$ Con: -$n_{m,l,j,z,k,h,r}$: número de hogares seleccionados en la vivienda m, de la manzana l, del CP/ área urbana j, de la zona z, del municipio k que pertenece al estrato h de la entidad territorial r. -$N_{m,l,j,z,k,h,r}$: total de hogares según marco en la vivienda m, de la manzana l, del CP/ área urbana j, de la zona z, del municipio k que pertenece al estrato h de la entidad territorial r. - $\pi_{V,f} = 1$, para la zona de veredas.

La probabilidad de inclusión al nivel de manzana teniendo en cuenta el submuestreo, queda definida como:

$$\pi_{III,l} = \pi_{III,l,1} * \frac{1}{N_{sect}}$$

Siendo N_{sect} el número de sectores creados en la manzana.

Finalmente, con base en lo planteado anteriormente, la probabilidad de inclusión final para la p-ésima unidad seleccionada bajo el diseño probabilístico definido, queda definida como el producto de las probabilidades de inclusión de cada etapa:

$$\pi_{p_multietápico} = \pi_{I,khr} * \pi_{II,j} * \pi_{III,l} * \pi_{IV,m} * \pi_{V,f} * \pi_{VI,p}$$

Con :

- $\pi_{VI,p} = \frac{1}{N_{f,m,l,j,z,k,h,r}}$, probabilidad de inclusión de la sexta etapa (persona mayor de 18 años), en el hogar f, de la vivienda m, de la manzana/segmento l, de la vereda /CP/área urbana j, de la zona z, del municipio k que pertenece al estrato h de la entidad territorial r;
- $N_{f,m,l,j,z,k,h,r}$ total, de personas mayores de 18 años, en el hogar f, de la vivienda m, de la manzana/segmento l, de la vereda /CP/ área urbana j, de la zona z, del municipio k que pertenece al estrato h de la entidad territorial r.

De acuerdo con esto, los factores de expansión básicos para cada etapa de muestreo quedan expresados de forma general como el inverso de la probabilidad de inclusión. Por lo tanto, el factor de expansión de cada unidad de muestreo toma la siguiente expresión de acuerdo con las expresiones antes definidas:

Factor de expansión de UPM (Primera etapa)	**Factor de expansión de veredas / centros poblados / área urbana (Segunda etapa)**	**Factor de expansión de manzanas / segmentos (Tercera etapa)**
$Fex_{I,khr} = \frac{1}{\pi_{I,khr}}$.	$Fex_{II,j} = \frac{1}{\pi_{II,j}}$.	$Fex_{III,l} = \frac{1}{\pi_{III,l}}$.
Factor de expansión de viviendas (Cuarta etapa)	**Factor de expansión de hogares (Quinta etapa)**	**Factor de expansión de personas (Sexta etapa)**
$Fex_{IV,m} = \frac{1}{\pi_{IV,m}}$.	$Fex_{V,f} = \frac{1}{\pi_{V,f}}$.	$Fex_{IV,p} = \frac{1}{\pi_{VI,p}}$.

Es así como el factor de expansión final la p-ésima unidad seleccionada bajo el diseño probabilístico indicado, queda definida como el inverso del producto de las probabilidades de inclusión de cada etapa:

$$Fex_{final_p} = \frac{1}{\pi_{p_multietápico}}.$$

1.3.12. Metodología de la estimación.

i) Parámetros e indicadores fundamentales para estimar que dan cumplimiento a los objetivos de la consultoría.

De acuerdo con la información del anexo técnico del estudio, para dar alcance al contenido temático, se construyó un cuestionario con los siguientes capítulos:

- Características sociodemográficas;
- Conectividad a internet;
- Gobierno y generales;
- Transparencia;
- Pobreza y desigualdad;
- Seguridad;
- Niños, niñas y adolescentes;
- Agricultura;
- Justicia;
- Educación;
- Salud;
- Trabajo;
- Cultura y recreación;
- Infraestructura y transporte;
- Minas y energía;
- Barrios, vivienda y servicios públicos;
- Ambiente;
- Paz;
- Otros temas estratégicos del gobierno nacional.

Respecto a las medidas de calidad asociadas a las estimaciones y de manera coherente con la estructura de contenido para la publicación de resultados de los levantamientos anteriores, se acompañará cada cuadro de salida con las siguientes notas técnicas:

- 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*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.
- Nota 5: Intervalos de confianza calculados con base en la transformación Logit con un 95 % de confianza.<.

Ahora, lo antes mencionado guarda relación con las recomendaciones que ha hecho la CEPAL[42] sobre este particular. Específicamente, es necesario determinar unos criterios de calidad para determinar la validez estadística de las estimaciones puntuales. Esto es: establecer si la estimación cuenta con márgenes de calidad aceptables o con un nivel inadecuado de precisión y, por lo tanto, no debería utilizarse para realizar inferencia estadística. En consecuencia, los cuadros de salida tendrán las siguientes medidas de calidad asociadas a la estimación puntual (Figura 4):

Figura 4. Estructura y contenido de los cuadros de salida relacionados con la estimación puntual de interés según el nivel de desagregación o análisis.

1402. En los últimos 12 meses de acuerdo con lo que usted conoce o percibe del servicio de salud en su municipio o ciudad: La calidad del servicio de salud ha:

	Estimación	Error estándar	CVE	Deff	n	Límite inferior del intervalo de confianza (al 95%)	Límite superior del intervalo de confianza (al 95%)	Margen de error
1. TOTAL								
Desmejorado	39,45%	1,05%	2,65	3,84	3297	37,4%	41,6%	2,0%
Permanecido igual	38,97%	0,96%	2,47	3,28	3307	37,1%	40,9%	1,9%
Mejorado	19,46%	0,74%	3,81	2,94	1616	18,0%	21,0%	1,5%
No sabe/No responde	2,13%	0,30%	14,10	3,62	168	1,6%	2,8%	0,6%
2. ZONA								
2.1. Urbano								
Desmejorado	40,06%	1,12%	2,80	3,43	2222	37,8%	42,3%	2,2%
Permanecido igual	38,93%	1,08%	2,77	3,21	2111	36,8%	41,1%	2,1%
Mejorado	18,93%	0,85%	4,51	3,11	1013	17,3%	20,7%	1,7%
No sabe/No responde	2,08%	0,35%	16,82	3,94	93	1,5%	2,9%	0,7%
2.2. Rural								
Desmejorado	37,24%	1,97%	5,30	3,05	1075	33,4%	41,3%	3,9%
Permanecido igual	39,12%	1,83%	4,67	2,57	1196	35,5%	42,8%	3,6%
Mejorado	21,35%	1,29%	6,02	1,80	603	18,9%	24,0%	2,5%
No sabe/No responde	2,28%	0,51%	22,42	2,15	75	1,5%	3,6%	1,0%
3. SEXO								
3.1. Hombre								
Desmejorado	35,44%	1,45%	4,08	3,67	1190	32,6%	38,4%	2,8%
Permanecido igual	43,29%	1,36%	3,14	3,02	1447	40,6%	46,0%	2,7%
Mejorado	19,00%	1,07%	5,64	3,00	639	16,9%	21,2%	2,1%
No sabe/No responde	2,27%	0,37%	16,39	2,51	81	1,6%	3,1%	0,7%
3.2. Mujer								
Desmejorado	43,13%	1,37%	3,17	3,33	2107	40,4%	45,9%	2,7%
Permanecido igual	35,00%	1,38%	3,94	3,65	1860	32,3%	37,8%	2,7%
Mejorado	19,88%	0,94%	4,72	2,42	977	18,1%	21,8%	1,8%
No sabe/No responde	1,99%	0,45%	22,67	4,57	87	1,3%	3,1%	0,9%

Fuente: Elaboración propia.

[42] https://www.cepal.org/es/publicaciones/45681-criterios-calidad-la-estimacion-indicadores-partir-encuestas-hogares-aplicacion

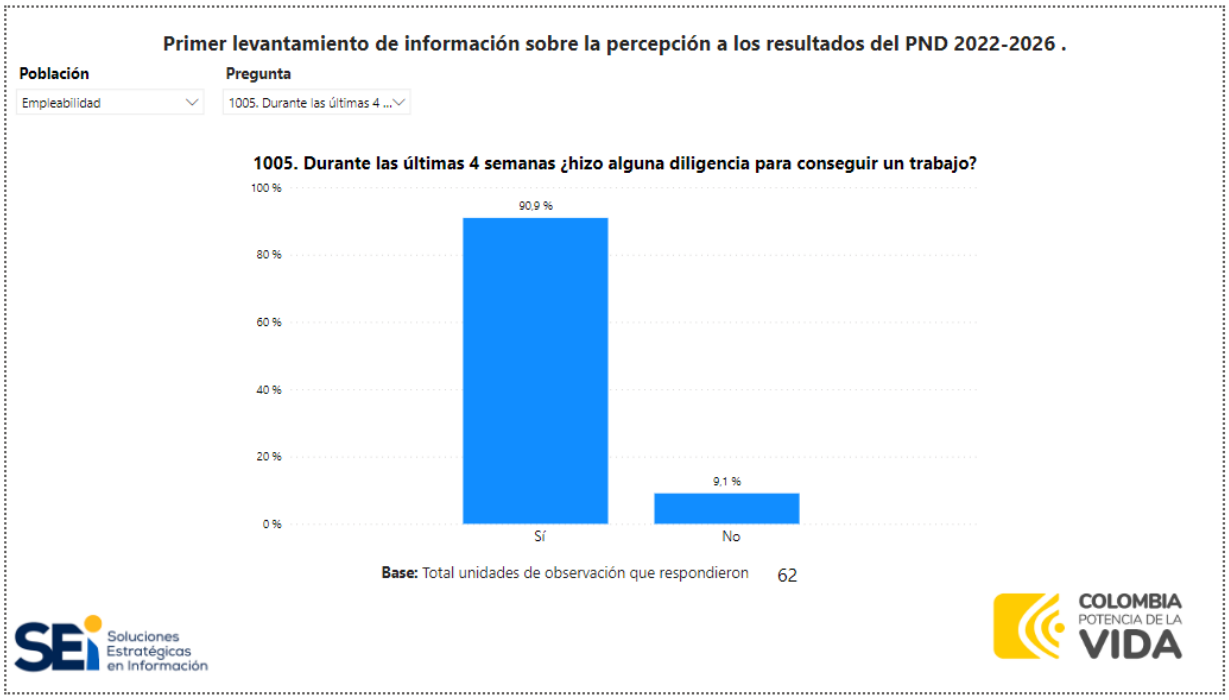
Por otra parte, como diferencial se construirá una estructura que facilite el análisis, así como la sensibilización de los resultados para la toma de decisiones, su visualización gráfica, las estimaciones puntuales junto con sus medidas de calidad (cuadro de salida) y la interpretación adecuada de las estimaciones con carácter inferencial (notas técnicas para la correcta interpretación de los resultados).

Para esto, se utilizará la herramienta Power BI de Microsoft, la cual tiene, entre otras, las siguientes ventajas:

i) acceso a los datos desde prácticamente cualquier dispositivo;
ii) os datos son accesibles, pueden ser compartidos y actualizados en tiempo real;
iii) se enfoca en el diseño y la presentación de la información;
iv) tiene la capacidad de importar datos en prácticamente todos los formatos y almacenados en cualquier fuente.

Además, facilita la transformación y el cruce de datos de distintos orígenes en una única vista. En este contexto, y a manera de ejemplo, a continuación, se presentan algunas salidas de información que se generarían (Figura 5).

Figura 5. Salidas de información a incluir en el visualizador en Power BI con los resultados inferenciales.

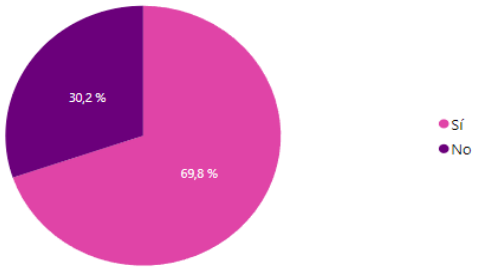


Primer levantamiento de información sobre la percepción a los resultados del PND 2022-2026.

Población Empleabilidad

Pregunta 311. ¿Es usted el/la jefe de ...

311. ¿Es usted el/la jefe de hogar?



Base: Total unidades de observación que respondieron 564

Soluciones Estratégicas en Información

COLOMBIA POTENCIA DE LA VIDA

Primer levantamiento de información sobre la percepción a los resultados del PND 2022-2026.

Población Empleabilidad

Pregunta 1005. Durante las últimas 4 ...

1005. Durante las últimas 4 semanas ¿hizo alguna diligencia para conseguir un trabajo?

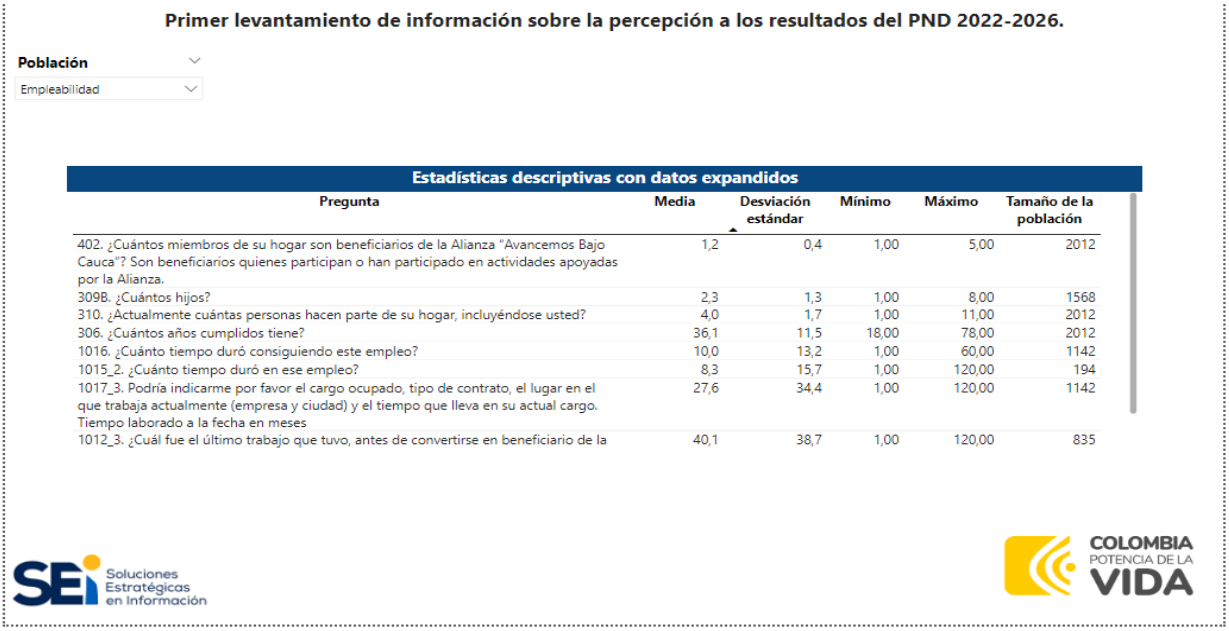
Categoría	Estimación	Error estándar	Coeficiente de variación %	Límite inferior del intervalo de confianza (95%)	Límite superior del intervalo de confianza (95%)	Margen de error	Unidades de observación que respondieron
No	9,1 %	3,0 %	33,4	4,6 %	17,2 %	5,9 %	7
Sí	90,9 %	3,0 %	3,3	82,8 %	95,4 %	5,9 %	55

Notas

Nota 1: 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*1,96).
Nota 2: Los márgenes de calidad para las estimaciones tipo proporción se determinan en función del coeficiente de variación (CVE), así: < 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.

Soluciones Estratégicas en Información

COLOMBIA POTENCIA DE LA VIDA



Primer levantamiento de información sobre la percepción a los resultados del PND 2022-2026.

Población

Empleabilidad

Estadísticas descriptivas con datos expandidos

Pregunta	Media	Desviación estándar	Mínimo	Máximo	Tamaño de la población
402. ¿Cuántos miembros de su hogar son beneficiarios de la Alianza "Avancemos Bajo Cauca"? Son beneficiarios quienes participan o han participado en actividades apoyadas por la Alianza.	1,2	0,4	1,00	5,00	2012
309B. ¿Cuántos hijos?	2,3	1,3	1,00	8,00	1568
310. ¿Actualmente cuántas personas hacen parte de su hogar, incluyéndose usted?	4,0	1,7	1,00	11,00	2012
306. ¿Cuántos años cumplidos tiene?	36,1	11,5	18,00	78,00	2012
1016. ¿Cuánto tiempo duró consiguiendo este empleo?	10,0	13,2	1,00	60,00	1142
1015_2. ¿Cuánto tiempo duró en ese empleo?	8,3	15,7	1,00	120,00	194
1017_3. Podría indicarme por favor el cargo ocupado, tipo de contrato, el lugar en el que trabaja actualmente (empresa y ciudad) y el tiempo que lleva en su actual cargo. Tiempo laborado a la fecha en meses	27,6	34,4	1,00	120,00	1142
1012_3. ¿Cuál fue el último trabajo que tuvo, antes de convertirse en beneficiario de la	40,1	38,7	1,00	120,00	835

Soluciones Estratégicas en Información

COLOMBIA POTENCIA DE LA VIDA

Fuente: Elaboración propia, SEI, agosto 2023.

Ahora, con base en la batería de preguntas relacionadas con las variables incluidas en el instrumento o encuesta definidas para responder a los objetivos que el DNP determine para este levantamiento, se construyeron indicadores sobre la percepción general de estos temas. Los resultados son presentados como estimaciones de parámetros tipo Total o Razón (promedios, proporciones, etc.) y sus respectivas medidas de calidad (error estándar, error estándar relativo, límites de confianza al 95% y DEFF).

De acuerdo con esto, el parámetro tipo total tiene la siguiente expresión matemática:

$$t = \sum_{U} y_k$$

En donde U es el universo de estudio, y la variable y_k es una característica de la población observada en la unidad k.

Ahora, con base en este parámetro se define el de la razón (proporción o promedio), que tiene la siguiente expresión matemática:

$$R = \frac{t_y}{t_z} = \frac{\sum_U y_k}{\sum_U z_k}$$

ii) Estimadores a considerar junto con sus expresiones teóricas para estimar los parámetros de interés.

Al tener probabilidades de inclusión en cada una de las etapas de muestreo, se plantea el estimador de Horvitz-Thompson de manera general, el cual para estimar el total de una variable Y observada en la muestra, toma la siguiente expresión:

$$\hat{t}_{y,\pi} = \sum_{S} \frac{y_k}{\pi_k} = \sum_{S} d_k y_k$$

Fuente: (Gutiérrez, Estrategias de Muestreo. Diseño de Encuestas y Estimación de Parámetros, 2016)

En donde π_k es la probabilidad de inclusión del k-ésimo elemento, el cual está dado por las probabilidades de inclusión de las seis etapas de muestreo planteadas.

Ahora bien, dado que el diseño muestral es estratificado y multietápico, con selección de las unidades de muestreo a través del método M.A.S. (UPM, centros poblados, veredas, manzanas, segmentos, viviendas, hogares y personas), el estimador de Horvitz-Thompson, que es insesgado para el total poblacional, se expresa como:

$$\hat{t}_{y,\pi} = \sum_{r=1}^{33} \sum_{h=1}^{n_r} \sum_{k=1}^{n_{hr}} \frac{1}{\pi_{I,khr}} \sum_{z=1}^{3} \left[\sum_{j \in S_{II,z,k,h,r}} \frac{1}{\pi_{II,j}} \sum_{l \in S_{III,j,z,k,h,r}} \frac{1}{\pi_{III,l}} \sum_{m \in S_{IV,l,j,z,k,h,r}} \frac{1}{\pi_{IV,m}} \sum_{f \in S_{V,m,l,j,z,k,h,r}} \frac{1}{\pi_{V,f}} \sum_{p \in S_{VI,f,m,l,j,z,k,h,r}} \frac{y_p}{\pi_{VI,p}} \right]$$

Fuente: Con base (Gutiérrez, Estrategias de Muestreo. Diseño de Encuestas y Estimación de Parámetros, 2016)

En donde:

- r varía de 1 a 33, corresponde a las entidades territoriales del PND;
- n_r es el total de estratos[43] construidos en la entidad territorial r;
- h varía de 1 a n_r, corresponde a los estratos construidos en cada entidad territorial;
- n_{hr} es el total de UPM seleccionadas en el estrato h de la entidad territorial r;
- k varía de 1 a n_{hr}, corresponde a las UPM seleccionadas del estrato h en la entidad territorial r;
- $\pi_{I,khr}$ es la probabilidad de inclusión de la k-*ésima* UPM que pertenecen al estrato h de la entidad territorial r;
- z varía de a 1 a 3, corresponde a la zona urbana (1), zona rural en el nivel de centros poblados[44] (2) y veredas[45] (3);

[43] Los estratos están definidos como agrupaciones de UPM, se construyeron a partir del algoritmo de Lavallée, P. and Hidiroglou, M.A. (1988), dentro de cada entidad territorial.
[44] Conformada por los centros poblados (clase 2).
[45] Conformada por la zona rural dispersa (clase 3).

- $S_{II,z,k,h,r}$ es la muestra de USM[46] (veredas/centros poblados/área urbana[47]), de la zona z, de la k-*ésima* UPM, del estrato h en la entidad territorial r;
- $\pi_{II,j}$ es la probabilidad de inclusión de las USM (veredas/centros poblados/área urbana[48]) de la muestra $S_{II,z,k,h,r}$;
- $S_{III,j,z,k,h,r}$ es la muestra de las unidades de muestreo en la tercera etapa, de la USM j, de la zona z, de la k-*ésima* UPM, del estrato h en la entidad territorial r. Corresponde a la muestra de manzanas en los casos de la zona urbana y centros poblados, y corresponde a la muestra de segmentos en el caso de la zona rural dispersa;
- $\pi_{III,l}$ es la probabilidad de inclusión de las unidades de muestreo en la tercera etapa (manzanas/segmentos) de la muestra $S_{III,j,z,k,h,r}$;
- $S_{IV,l,j,z,k,h,r}$ es la muestra de viviendas en la cuarta etapa, de la (manzana/segmento) l, de la USM j, de la zona z, de la k-*ésima* UPM, del estrato h de la entidad territorial r;
- $\pi_{IV,m}$ es la probabilidad de inclusión de las unidades de muestreo en la cuarta etapa (viviendas) de la muestra $S_{IV,l,j,z,k,h,r}$;
- $S_{V,m,l,j,z,k,h,r}$ es la muestra de hogares en la quinta etapa, de la vivienda m, de la (manzana/segmento) l, de la USM j, de la zona z, de la k-*ésima* UPM, del estrato h de la entidad territorial r;
- $\pi_{V,f}$ es la probabilidad de inclusión de las unidades de muestreo en la quinta etapa (hogares) de la muestra $S_{V,m,l,j,z,k,h,r}$;
- $S_{VI,f,m,l,j,z,k,h,r}$ es la muestra de personas en la sexta etapa, del hogar f, de la vivienda m, de la (manzana/segmento) l, de la USM j, de la zona z, de la UPM k, del estrato h de la entidad territorial r;
- $\pi_{VI,p}$ es la probabilidad de inclusión de las unidades de muestreo en la sexta etapa (personas) en la muestra $S_{VI,f,m,l,j,z,k,h,r}$.

Con base en este estimador para un total, se plantea la estimación de razones de donde se derivan los promedios y porcentajes, las cuales son las variables de interés que priman en el desarrollo y la presentación de resultados de este estudio, y toman la siguiente forma:

$$\hat{R} = \frac{\hat{t}_{y,\pi}}{\hat{t}_{z,\pi}}$$

Fuente: (Särndal, Swensson, & Wretman, Model Assisted Survey Sampling., 2002)

[46] Unidades secundarias de muestreo.
[47] En el caso de la zona urbana, la muestra de USM corresponde a toda el área urbana de la UPM.
[48] En el caso de la zona urbana, $\pi_{II,j} = 1$, puesto que la muestra de USM corresponde a toda el área urbana de la UPM.

Adicional a la utilización del estimador de Horvitz-Thompson, también se plantea la utilización de estimadores de calibración, dependiendo de las variables auxiliares (X) de interés disponibles, por ejemplo, en la información oficial del censo de población y vivienda 2018 del DANE o el registro de usuarios residenciales con servicio de energía de la Superintendencia de Servicios Públicos (SSP); que permitirán garantizar el balance para características de la población que son de interés.

Fundamentalmente, el método de calibración consiste en calcular pesos $w_k$que sean muy cercanos a los inversos de la probabilidad de inclusión del $k-ésimo$ elemento (d_k), dada la siguiente restricción:

$$\sum_{k \in S} w_k X_k = t_x$$

Fuente: (Gutiérrez, Estrategias de Muestreo. Diseño de Encuestas y Estimación de Parámetros, 2016)

Particularmente, dependiendo de los resultados obtenidos y de la disponibilidad de información auxiliar relacionada con zona, región, sexo, grupo etario, etc., se podrá utilizar la técnica de calibración para el ajuste de los factores de expansión, así como estimadores de calibración para optimizar la estimación de la varianza del estimador.

iii) Definición de expresiones teóricas para estimar de forma insesgada o aproximadamente insesgada la varianza de los estimadores de los parámetros de caracterización de acuerdo con el diseño probabilístico definido.

Al utilizar para el total, el estimador de Horvitz-Thompson $(\pi - estimador)$, la expresión general de la varianza del estimador en un diseño en seis etapas está dada por:

$$Var(\hat{t}_{y,\pi}) = \sum_{q=1}^{6} V_q$$

Fuente: Con base en la expresión 7.4.3 (Gutiérrez, Estrategias de Muestreo. Diseño de Encuestas y Estimación de Parámetros, 2016)

Siendo V_q la varianza de la $q-ésima$ etapa de muestreo, de tal forma que el estimador insesgado de la varianza está dado por:

$$\widehat{Var}(\hat{t}_{y,\pi}) = \sum_{q=1}^{6} \hat{V}_q$$

Fuente: Con base en la expresión 7.4.4 (Gutiérrez, Estrategias de Muestreo. Diseño de Encuestas y Estimación de Parámetros, 2016)

Siendo $\hat{V}_q$ un estimador insesgado de V_q. En el caso del diseño muestral a implementar, el estimador insesgado de la varianza está dado por la siguiente expresión:

$$\widehat{Var}_{multietápico}(\hat{t}_{y,\pi}) = \sum_{r=1}^{33}\sum_{h=1}^{n_r}\left[\hat{V}_{UPM_{h,r}} + \sum_{k=1}^{n_{hr}}\frac{\hat{V}_{UPM_{k,h,r}}}{\pi_{I,k,h,r}}\right]$$

Con las varianzas asociadas a la primera etapa $\hat{V}_{UPM_{hr}}$ dadas por:

$$\hat{V}_{UPM_{hr}} = \frac{{N_{I,h,r}}^2}{n_{I,h,r}}\left(1 - \frac{n_{I,h,r}}{N_{I,h,r}}\right)S^2_{yS_{I,h,r}}$$

Siendo $S^2_{yS_{I,h,r}}$ la varianza muestral entre las UPM que pertenecen al estrato h de la entidad territorial r.

Las varianzas a nivel de las UPM están dadas por:

$$\hat{V}_{UPM_{k,h,r}} = \sum_{z=1}^{3}\left[\frac{{N_{z,k,h,r}}^2}{n_{z,k,h,r}}\left(1 - \frac{n_{z,k,h,r}}{N_{z,k,h,r}}\right)S^2_{yS_{II,z,k,h,r}} + \frac{N_{z,k,h,r}}{n_{z,k,h,r}}\sum_{j=1}^{n_{z,k,h,r}}\hat{V}_{USM_{j,z,k,h,r}}\right]$$

Con:

$$S^2_{yS_{II,z,k,h,r}} = \frac{1}{n_{z,k,h,r}-1}\sum_{j=1}^{n_{z,k,h,r}}\left(\hat{t}_j - \bar{\hat{t}}_{z,k,h,r}\right)^2$$

$$\hat{t}_j = \sum_{l=1}^{n_{j,z,k,h,r}}\frac{\hat{t}_l}{\pi_{III,l}}\,,\ \bar{\hat{t}}_{z,k,h,r} = \sum_{j=1}^{n_{z,k,h,r}}\left(\frac{\hat{t}_j}{n_{z,k,h,r}}\right)$$

Las varianzas a nivel de las USM (CP/veredas/área urbana) están dadas por:

$$\hat{V}_{USM_{j,z,k,h,r}} = \frac{{N_{j,z,k,h,r}}^2}{n_{j,z,k,h,r}}\left(1 - \frac{n_{j,z,k,h,r}}{N_{j,z,k,h,r}}\right)S^2_{yS_{III,j,z,k,h,r}} + \frac{N_{j,z,k,h,r}}{n_{j,z,k,h,r}}\sum_{l=1}^{n_{j,z,k,h,r}}\hat{V}_{III_{l,j,z,k,h,r}}$$

Con:

$$S^2_{yS_{III,j,z,k,h,r}} = \frac{1}{n_{j,z,k,h,r}-1}\sum_{l=1}^{n_{j,z,k,h,r}}\left(\hat{t}_l - \bar{\hat{t}}_{j,z,k,h,r}\right)^2$$

$$\hat{t}_l = \sum_{m=1}^{n_{l,j,z,k,h,r}}\frac{\hat{t}_m}{\pi_{IV,m}}\,,\ \bar{\hat{t}}_{j,z,k,h,r} = \sum_{l=1}^{n_{j,z,k,h,r}}\left(\frac{\hat{t}_l}{n_{j,z,k,h,r}}\right)$$

Las varianzas a nivel de las manzanas/segmentos están dadas por:

$$\hat{V}_{III_{l,j,z,k,h,r}} = \frac{N_{l,j,z,k,h,r}{}^2}{n_{l,j,z,k,h,r}}\left(1-\frac{n_{l,j,z,k,h,r}}{N_{l,j,z,k,h,r}}\right)S^2_{yS_{IV,l,j,z,k,h,r}} + \frac{N_{l,j,z,k,h,r}}{n_{l,j,z,k,h,r}}\sum_{m=1}^{n_{l,j,z,k,h,r}}\hat{V}_{IV_{m,l,j,z,k,h,r}}$$

Con:

$$S^2_{yS_{IV,l,j,z,k,h,r}} = \frac{1}{n_{l,j,z,k,h,r}-1}\sum_{m=1}^{n_{l,j,z,k,h,r}}\left(\hat{t}_m - \bar{\hat{t}}_{l,j,z,k,h,r}\right)^2$$

$$\hat{t}_m = \sum_{f=1}^{n_{m,l,j,z,k,h,r}}\frac{y_f}{\pi_{V,f}},\ \bar{\hat{t}}_{l,j,z,k,h,r} = \sum_{m=1}^{n_{l,j,z,k,h,r}}\left(\frac{\hat{t}_m}{n_{l,j,z,k,h,r}}\right),\ y_f = y_p,\ \text{Tal que } p \in S_{VI,f,m,l,j,z,k,h,r}$$

Las varianzas a nivel de las viviendas están dadas por:

$$\hat{V}_{IV_{m,l,j,z,k,h,r}} = \frac{N_{m,l,j,z,k,h,r}{}^2}{n_{m,l,j,z,k,h,r}}\left(1-\frac{n_{m,l,j,z,k,h,r}}{N_{m,l,j,z,k,h,r}}\right)S^2_{yS_{V,m,l,j,z,k,h,r}}$$

Con:

$$S^2_{yS_{V,m,l,j,z,k,h,r}} = \frac{1}{n_{m,l,j,z,k,h,r}-1}\sum_{f=1}^{n_{m,l,j,z,k,h,r}}\left(y_f - \bar{\hat{t}}_{m,l,j,z,k,h,r}\right)^2$$

$$\bar{\hat{t}}_{m,l,j,z,k,h,r} = \sum_{f=1}^{n_{m,l,j,z,k,h,r}}\left(\frac{y_f}{n_{m,l,j,z,k,h,r}}\right)$$

Dado que las fórmulas anteriores son usadas para la estimación de la varianza en el caso de un total, y teniendo en cuenta que los principales resultados del estudio provienen de la estimación de porcentajes o razones, se hace necesario, para el caso de la estimación de varianza para razones, desarrollar la aproximación lineal de la razón vía series de Taylor, de manera tal que según Särndal et.al (2002), sección 5.6., se debe usar la variable e_k definida como:

$$e_k = (y_k - \hat{R}z_k)\frac{1}{\hat{t}_{z,\pi}}$$

Fuente: (Särndal, Swensson, & Wretman, Model Assisted Survey Sampling., 2002)

Así, con base en la expresión anterior y reemplazando la variable Y en las fórmulas de la varianza por la variable anterior, la aproximación de la varianza queda expresada en función de la variable e, con lo cual se llega finalmente a la estimación de la varianza del estimador de una razón.

Es relevante mencionar que se implementaron procedimientos alternativos (robustos) de estimación de los parámetros de interés, mediante la versión 15 de STATA. Este software cuenta con un módulo de comandos que permite hacer análisis cuando los datos provienen de una muestra que fue recolectada a partir de un diseño probabilístico multietápico, permitiendo definir la información más importante del diseño de muestreo implementado.

Esto hace que la varianza estimada sea calculada teniendo en cuenta todas las etapas, estratos y dominios de análisis, permitiendo así, una buena aproximación. Específicamente, el módulo de comandos SVY tiene programados para la estimación de la varianza las ecuaciones matemáticas de la aproximación lineal de Taylor[49].

1.3.13. Tamaño de muestra.

i) Parámetros trazadores para el cálculo de tamaño de muestra

Como ya se mencionó, los resultados entendidos como indicadores, específicamente, de tipo proporción, para análisis inferencial a partir de la información recopilada a través del cuestionario serán presentados como estimación de parámetros que pueden ser expresados como totales, proporciones, medias, razones o coeficientes de regresión.

Específicamente, el estudio cuenta con información primaria que hace referencia, a datos recopilados mediante un trabajo de campo con visita a personas de 18 años o más en sus hogares, los cuales son validados para garantizar su integridad y procesados para generar las estimaciones, que respondan al alcance inferencial de indicadores y/o cuadros de salida, dando de esta forma respuesta a los requerimientos de información. Por otra parte, para definir el tamaño mínimo de muestra se acogieron los lineamientos definidos en el anexo técnico del contrato con relación a: niveles de desagregación, el margen de error, el nivel de confianza y el parámetro trazador.

ii) Expresiones para el cálculo de tamaño de muestra

Para validar que el tamaño de muestra propuesto (y su distribución) satisface los requerimientos de precisión y confiabilidad para el parámetro trazador, se realizó un análisis comparativo bajo el siguiente enfoque. Mediante el uso de la librería samplesize4surveys, del software estadístico R versión 4.1.1 (Gutiérrez, 2018), es posible identificar bajo diferentes escenarios, el tamaño óptimo que garantice las

[49] https://www.stata.com/manuals13/svyvarianceestimation.pdf#svyvarianceestimation, páginas 6-8, *"The ratio estimator"*, *"Linearized/robust variance estimation"*.

medidas de calidad descritas, tanto para parámetros de caracterización como de cuantificación.

Al definir qué valor del margen de error (ε) se desea, así como el DEFF y, de otra parte, conociendo el tamaño de la población (N) y la proporción estimada (parámetro de interés - P), el cálculo del tamaño de muestra estaría dado por:

$$n = \frac{n_0}{1 + \frac{n_0}{N}}$$

donde $n_0 = \frac{Z_{1-\frac{\alpha}{2}}^2 S^2}{\varepsilon^2}$ y $S^2 = P(1-P)DEFF$.

Ahora, si el cálculo del tamaño de muestra se realiza teniendo un coeficiente de variación (cve) deseable, la fórmula a utilizar es:

$$n = \frac{S^2}{P^2 cve^2 + \frac{S^2}{N}}$$

Este enfoque se implementará más adelante para validar los resultados de la simulación.

iii) Tamaño de muestra según etapa de muestreo

Para el diseño muestral probabilístico propuesto: multietápico, estratificado (Nivel I: entidad territorial, es decir, departamento 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), el tamaño de muestra que responde a los requerimientos mínimos del anexo técnico es de:

- 238 UPM de las cuales 33 de ellas (capitales o áreas metropolitanas) son de inclusión forzosa y representan a 57 municipios. Las 205 UPM restantes representan a 205 municipios.
- 13.400 encuestas a personas de 18 años o más, con una distribución aproximada del 30 % en la zona rural y 70 % en la zona urbana. Esto representa un incremento del 4,2 % frente al tamaño del anexo técnico, es decir, 542 encuestas adicionales a las mínimas establecidas.
- Como se menciona anteriormente, los tamaños óptimos para las etapas de muestreo posteriores a la primera se definen a través de la simulación y dependerán de las UPM que queden finalmente incluidas en la muestra.

iv) Estrategia orientada a disminuir el efecto de diseño (DEFF).

Siendo el efecto de diseño (DEFF) una medida utilizada para comparar la eficiencia entre diseños probabilísticos multietápico e identificar la ganancia o pérdida de precisión en comparación con un diseño aleatorio simple sin reemplazo (Gutiérrez, 2016), para este caso, se determinó implementar diferentes estrategias técnicas que fueron definiéndose con base en las conclusiones que se obtuvieron analizando las medidas de calidad del ejercicio de simulación, para los principales dominios de análisis geográficos.

Sin embargo, como estrategia para mejorar la eficiencia del diseño probabilístico, principalmente, se incrementaron los tamaños mínimos de muestra de UPM y UFM sugeridos en el anexo técnico, lo cual tiene una incidencia directa (el aumentar el tamaño de muestra, particularmente, en la primera etapa que según la teoría es la que aporta más varianza) en la reducción del nivel de error (m.e. o c.v.e).

En síntesis, las acciones más importantes para controlar el DEFF (o minimizar la varianza del diseño muestral probabilístico definido) fueron:

- Los distintos niveles de estratificación fueron definidos para garantizar estratos homogéneos entre sí y heterogéneos entre ellos. En el caso del Nivel II de estratificación se tienen UPM de inclusión forzosa; el Nivel IV (manzanas en cabecera) permitió construir una cantidad óptima de estratos dentro de cada UPM;
- El algoritmo de estratificación utilizado permitió determinar la cantidad de estratos y UPM a incluir en la muestra de manera óptima (tamaños similares en los estratos) para garantizar los niveles de error exigidos para los principales dominios geográficos de análisis;
- Se definieron UPM como áreas metropolitanas (por ejemplo, Medellín hace parte del área metropolitana del Valle de Aburra, la cual está conformada por 10 municipios) para disminuir el c.v.e o margen de error para las entidades territoriales;
- La asignación proporcional de la muestra de unidades finales se ajustó para incrementar el tamaño de muestra asignado en la zona rural, con el fin de obtener en la simulación un c.v.e o error estándar relativo (EE) máximo del 5% con un nivel de confiabilidad del 95 %;
- Se estableció una cantidad óptima de centros poblados o veredas a incluir en la muestra por UPM;
- La cantidad de viviendas a seleccionar por manzana definida fue de 6, permitiendo maximizar la cantidad de estas unidades a seleccionar en la zona urbana de cada UPM;
- Se definió solo un hogar por vivienda y solo una persona por hogar para controlar la correlación intraclásica $(rho = \rho)$ en las últimas etapas de muestreo. Es decir,

se buscó mayor heterogeneidad al interior de las manzanas para minimizar la varianza en esta etapa de muestreo, entre manzanas y entre estratos de manzanas, en el nivel de cada U.P.M;

Por último, es relevante mencionar que se utilizará la transformación Logit para la construcción de los límites del intervalo de confianza.

1.3.14. Precisión de los resultados.

i) Simulación de escenarios de coeficientes de variación y margen de error que se acerquen al error mínimo esperado de la consultoría.

A partir del marco de UPM construido, el marco de manzanas y el marco de centro poblados, construidos con base en el censo 2018, se realizó la simulación, cuyo procedimiento a continuación se detalla:

Se generó una variable aleatoria con distribución Bernoulli de parámetro 0.5 que permitiera obtener la estimación del error esperado para una razón o proporción de alrededor del 50%. Esto con el fin de maximizar el tamaño de muestra para la población objetivo en los niveles: nacional, zona y entidad territorial; ya que cuando se estiman porcentajes cercanos al 50% se presentan los errores muestrales más altos y, porque el parámetro principal se definió como una razón, específicamente, una proporción.

Con el fin de realizar la simulación implementando el diseño muestral propuesto, se realizaron inicialmente diferentes escenarios para determinar la cantidad de subestratos a utilizar y de UPM a entrar como inclusión forzosa en cada entidad territorial (estrato).

Para esto, se escogió el algoritmo propuesto por Lavallée, P. and Hidiroglou, M.A. (1988), desarrollado por Kozak, M. (2004) [50] e implementado en el software libre R por Sophie Baillargeon and Louis-Paul Rivest (2014)[51], el cual define los estratos de tal forma que el coeficiente de variación (CV) sea menor a un valor determinado; en este caso la intención era obtener un error de muestreo inferior al 7 % en cada entidad territorial para la estimación de los parámetros de interés.

Posteriormente, al interior de los subestratos definidos por el algoritmo se realizó la selección de UPM por réplica, dado la implementación de la técnica de Monte Carlo, se asignó la cantidad mínima de encuestas a realizar por estrato proporcionalmente al

[50]Kozak, M. (2004). Optimal stratification using random search method in agricultural surveys. *Statistics in Transition*, 6(5), 797-806. n. 2.2-5. https://CRAN.R-project.org/package=stratification. Usando la función strata. LH

tamaño poblacional proyectado, y se fue ajustando por zona para garantizar los márgenes de error requeridos según el anexo técnico.

Teniendo como referente lo anterior, para la simulación se implementó el diseño muestral planteado: estratificado, multietápico, M.A.S. de UPM y M.A.S. en las siguientes etapas. Se estimó el parámetro (proporción) y las medidas de calidad asociadas (coeficiente de variación y margen de error) a través de la técnica *Bootstrap*, con lo cual se llega a una aproximación que generalmente sobreestima la varianza y aun así, se obtuvieron los siguientes márgenes de error o coeficientes de variación en promedio: inferiores al 5 % para el nivel nacional, alrededor del 7 % para las entidades territoriales de Bogotá, Antioquia, Atlántico, Cundinamarca, Magdalena, Santander y Valle y alrededor del 10 % para las entidades restantes e inferiores al 5 % por zona (urbana o rural).

En el anexo estadístico (Anexo 3), se encuentra el código utilizado y desarrollado en el software libre estadístico R (R Core Team, 2015), para la simulación antes descrita.

A continuación, se realiza una breve descripción del método de *Bootstrap*, según lo planteado por Särndal et al (2002). Suponiendo que una muestra probabilística S es seleccionada de la población objetivo U, bajo un diseño muestral arbitrario sin remplazamiento, el parámetro poblacional de interés θ es estimado por $\hat{\theta}$, bajo la técnica de *Bootstrap,* la cual funciona así:

1. A partir de la muestra S construir una población artificial U^*.
2. Seleccionar una serie de muestras *(Bootstrap*) independientes de U^* mediante el mismo diseño probabilístico usado en la selección de la muestra S. Cada muestra *Bootstrap* debe ser remplazada en U^* antes de realizar la siguiente selección. Para cada muestra calcular el estimador $\widehat{\Theta}_a^*$ $(a = 1, \ . \ . \ . \ , \ A.)$
3. La distribución observada de $\widehat{\Theta}_1^*, \ldots, \widehat{\Theta}_A^*$. es considerada una "estimativa" de la distribución muestral del estimador $\hat{\theta}$ y la varianza $V(\hat{\theta})$ es estimada por:

$$\hat{V}_{BS} = \frac{1}{A-1}\sum_{a=1}^{A}(\widehat{\Theta}_a^* - \widehat{\Theta}^*)^2. \text{ , con } \widehat{\Theta}^* = \frac{1}{A}\sum_{a=1}^{A}\widehat{\Theta}_a^*$$

De esta manera, con base en una simulación de 1.000 réplicas implementada a través del método de *Bootstrap*, se obtuvo una muestra de 13.400 personas de 18 años o más en 238 UPM, con una distribución aproximada del 70 % en la zona urbana y un 30 % en la zona rural.

Los tamaños óptimos para las etapas de muestreo posteriores a la primera se definieron a través de la simulación y depende de las UPM que quedaron finalmente incluidas en la muestra.

La distribución de la muestra de personas de 18 años o más por entidad territorial y zona se incluyó en el anexo estadístico (Anexo 3), así como los resultados relacionados con los márgenes de calidad para variaciones del parámetro tipo proporción entre el 10 % y el 50 %, para los principales dominios de análisis geográficos (nacional, zona y 33 entidades territoriales).

En la Tabla 3, se presentan resultados relacionados con los márgenes de calidad para un parámetro trazador del 50 % y un nivel de confianza del 95 %, para el total nacional por zona y para cada una de las 33 entidades territoriales; así como el tamaño de muestra y sus márgenes de calidad.

- Para nivel nacional (total de la muestra) se anticipa un c.v.e.[52] de aproximadamente 4 % cuando lo esperado es máximo del 5 %;
- Para la zona urbana se anticipan c.v.e. de aproximadamente 4 % cuando lo esperado es máximo del 5 %;
- Para la zona rural se anticipan c.v.e. de aproximadamente 5 % cuando lo esperado es máximo del 5 %;

Tabla 3. Resultados de la simulación para determinar el nivel de precisión con 1.000 réplicas, a nivel nacional, zona y municipios considerando una proporción del 50 % con un nivel de confianza del 95 %.

Departamento	Margen de error	Simulación		
		Urbano	Rural	Total
Bogotá, D.C.	5.5%	400	64	464
Antioquia	5.6%	443	166	609
Valle del Cauca	5.5%	432	154	586
Atlántico	5.9%	481	99	580
Cundinamarca	7.0%	450	167	617
Santander	5.4%	466	142	608
Magdalena	6.0%	333	384	718
Archipiélago de San Andrés	7.3%	195	170	365
Amazonas	9.0%	216	73	289
Arauca	8.5%	230	171	401
Bolívar	7.4%	287	150	437
Boyacá	8.3%	322	94	416
Caldas	7.6%	239	195	434
Caquetá	8.1%	249	165	414

[52] Coeficiente de variación estimado.

Casanare	9.5%	244	180	424
Cauca	7.9%	275	120	395
Cesar	7.5%	260	125	385
Chocó	7.8%	285	100	385
Córdoba	7.1%	274	124	398
Guainía	10%	130	70	200
Guaviare	6.6%	300	55	355
Huila	8.3%	233	158	391
La Guajira	7.6%	268	114	382
Meta	6.7%	309	50	359
Nariño	7.4%	296	91	387
Norte de Santander	8.3%	275	87	362
Putumayo	8.1%	250	90	340
Quindío	8.0%	249	74	323
Risaralda	8.4%	235	100	335
Sucre	7.8%	198	102	300
Tolima	9.4%	217	73	290
Vaupés	8.6%	164	60	224
Vichada	8.1%	178	50	228
Margen de error, agregado 33 entidades territoriales, zona urbana y zona rural.	**1.8%**	**1.9%**	**4.2%**	
Total		9.383 70.0%	4.017 30.0%	13.400

Fuente: Cálculos propios

Con respecto al tamaño de muestra definido, en el caso de las entidades territoriales, la Tabla 4 presenta la distribución de UPM de inclusión forzosa por departamento.

Tabla 4. Distribución de los municipios de inclusión forzosa por departamento[53].

Departamento	Municipios de IF
Amazonas	Leticia, Puerto Nariño
Antioquia	Medellín, Barbosa, Bello, Caldas, Copacabana, Envigado, Girardota, Itagüí, La estrella, Sabaneta

[53] Para el departamento de Guaviare no quedó un estrato de inclusión de forzosa, por lo tanto el municipio de San José del Guaviare no quedó dentro de la muestra.

Arauca	Arauca, Arauquita, Cravo Norte, Fortúl, Puerto Rondón, Saravena, Tame
Archipiélago de San Andrés	San Andrés, Providencia
Atlántico	Barranquilla, Galapa, Malambo, Puerto Colombia, Sabanalarga, Soledad
Bogotá, D.C.	Bogotá, D.C.
Bolívar	Cartagena, Arjona, El Carmen de Bolívar, Magangué
Boyacá	Tunja, Chiquinquirá, Duitama, Moniquirá, Paipa, Puerto Boyacá, Sogamoso
Caldas	Manizales, La Dorada
Caquetá	Florencia, San Vicente del Caguán
Casanare	Yopal, Aguazul, Paz de Ariporo, Tauramena, Villanueva
Cauca	Popayán, El Tambo, Santander de Quilichao
Cesar	Valledupar, Aguachica, Agustín Codazzi
Chocó	Quibdó, Alto Baudó, Bajo Baudó, Istmina, Riosucio
Córdoba	Montería, Cereté, Lorica, Sahagún
Cundinamarca	Chía, Facatativá, Fusagasugá, Madrid, Mosquera, Soacha, Zipaquirá
Guainía	Inírida, Barranco Minas
Huila	Neiva, Garzón, La Plata
La Guajira	Riohacha, Maicao, Manaure
Magdalena	Santa Marta, Ciénaga
Meta	Villavicencio, Acacías, Granada
Nariño	Pasto, Barbacoas, Ipiales, San Andrés de Tumaco, Túquerres
Norte de Santander	Cúcuta, Los Patios, Ocaña
Putumayo	Mocoa, Puerto Asís
Quindío	Armenia, Calarcá
Risaralda	Pereira, Dosquebradas, Santa Rosa de Cabal
Santander	Bucaramanga, Floridablanca, Girón, Piedecuesta
Sucre	Sincelejo, Corozal
Tolima	Ibagué, Chaparral
Valle del Cauca	Cali, Guadalajara de Buga, Candelaria, Jamundí, Palmira, Tuluá, Yumbo
Vaupés	Mitú, Pacoa
Vichada	Puerto Carreño, La Primavera, Santa Rosalía, Cumaribo

Fuente: Cálculos propios

ii) Medidas de precisión de los estimadores de los parámetros

Sobre las medidas de precisión o niveles máximo de error anticipados, como ya se analizó en el ítem "*Simulación de escenarios de coeficientes de variación y margen de error que se acerquen al error mínimo esperado de la consultoría*.", los resultados de la simulación permitieron encontrar tamaños mínimos de muestra que garantizaran para los principales dominios de análisis, márgenes de error o c.v.e por debajo del máximo establecido en el anexo técnico. En consecuencia, el tamaño propuesto garantiza los márgenes de precisión exigidos para el estudio.

Ahora, haciendo uso del enfoque de validación mencionado, para proporciones del 50 % (cuando se estiman porcentajes cercanos al 50 % se presentan los errores muestrales más altos), y considerando diferentes valores[54] de DEFF[55] y márgenes de error a partir del 4 %, para el grupo de las 7 entidades territoriales principales (Bogotá, Antioquia, Atlántico, Cundinamarca, Magdalena, Santander y Valle) se identifica que el tamaño mínimo promedio de encuestas para cada una de estas entidades territoriales debe estar alrededor de 530 para garantizar un margen de error máximo del 7 % (Figura 6) con DEFF esperados entre 1.7 y 1.9.

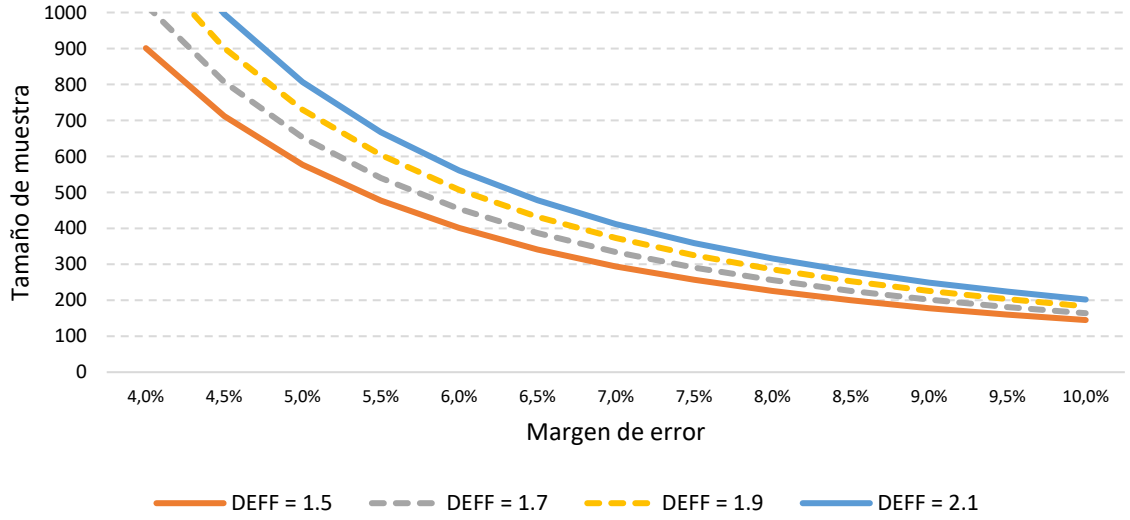
Específicamente, para Bogotá alrededor de 460 encuestas, 700 para Magdalena y aproximadamente 580 encuestas, en promedio, para las 5 entidades territoriales restantes, tal como se hizo en la simulación.

Bajo los parámetros anteriores, de manera análoga se tiene para el grupo de las 26 entidades territoriales restantes, el tamaño mínimo de encuestas, en promedio, para estas entidades territoriales debe ser aproximadamente de 372 para garantizar un margen de error máximo del 10 % (Figura 7) con DEFF esperados entre 1.7 y 1.9. Estos tamaños, con una asignación especifica por entidad y zona, fueron definidos dentro del proceso de simulación.

Finalmente, bajo este mismo enfoque de validación para proporciones del 50 %, considerando márgenes de error entre el 3 % y 5 % y diferentes valores del DEFF, para la zona rural se identifica que el tamaño mínimo de encuestas (30 %) debe ser de 4.000 (Figura 8).

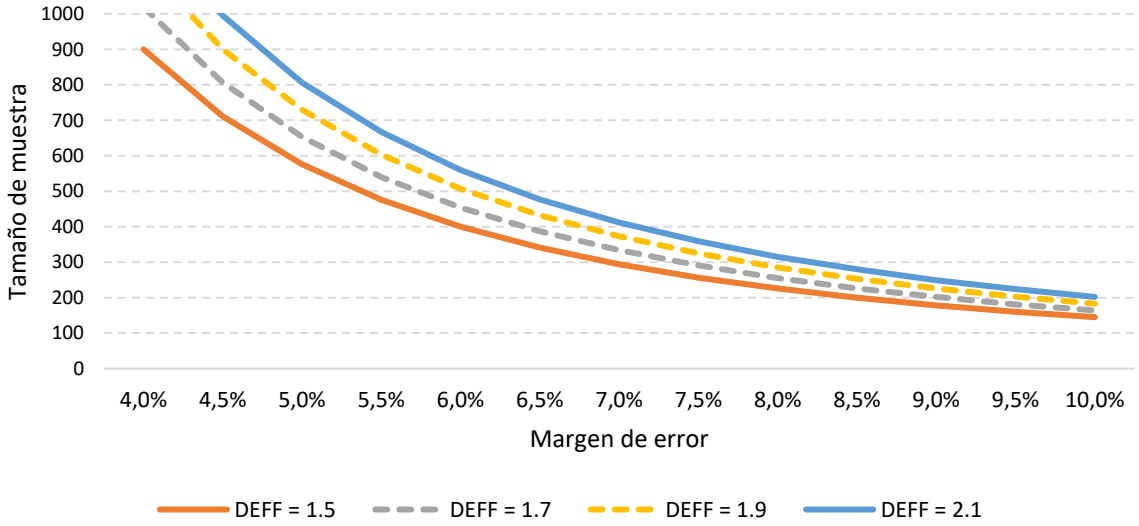
[54] Los valores del DEFF fueron definidos entre 1.5 y 2.1 con base en lo observado en los levantamientos anteriores de la EPC PND.
[55] Naciones Unidas recomienda que el deff que se defina no sea inferior a 2 para estudios de la complejidad como el que aquí se plantea. Ver: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://unstats.un.org/unsd/publication/seriesf/seriesf_98s.pdf, página 88 de 252.

Figura 6. Variación del tamaño de muestra para diferentes márgenes de error, en proporciones del 50%, para las 7 entidades territoriales principales



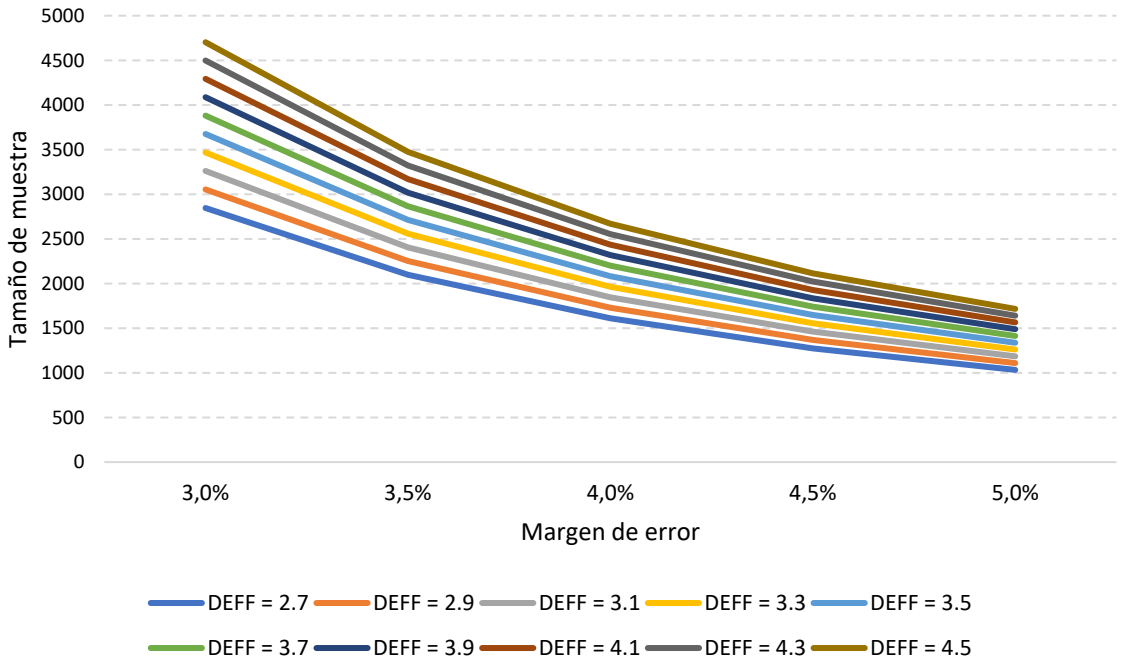
Fuente: Cálculos propios.

Figura 7. Variación del tamaño de muestra para diferentes márgenes de error, en proporciones del 50%, para resto de entidades territoriales



Fuente: Cálculos propios.

Figura 8. Variación del tamaño de muestra para diferentes márgenes de error, en proporciones del 50%, para la zona rural



Fuente: cálculos propios

1.3.15. Selección de la muestra.

Con base en el diseño muestral definido y a partir de la semilla “210923” se seleccionaron las UPM y los municipios para el levantamiento de información primaria a través de la encuesta de percepción ciudadana. El detalle de este proceso se puede observar en los códigos proporcionados en el anexo estadístico (Anexo 3).

1.3.16. Ajuste de cobertura.

i. Porcentaje de sobremuestra sobre el tamaño de muestra inicial

Con respecto a las UPM (municipios o áreas metropolitanas), la exclusión de estos de la muestra solo está condicionada a la situación de orden público, inseguridad o dificultades de acceso originados por desastres naturales. En caso de ocurrir, se documentará y presentará al CTS y al supervisor del contrato, las causas que originaron tal situación y se asignará el siguiente municipio, según el orden establecido en la selección con la semilla descrita anteriormente.

En el nivel de los centros poblados o veredas, se sigue el protocolo para justificar las razones de pérdida y con respecto a la inclusión del reemplazo, se selecciona la siguiente unidad en el orden de inclusión aleatoria definido para la selección de la muestra inicial.

En el caso de las manzanas cartográficas, se tiene en cuenta los resultados de cobertura de los levantamientos relacionados con el PND 2014-2018 y el PND 2018-2022 para estimar la pérdida potencial, la cual fundamentalmente se encuentra asociada a defectos de la información cartográfica (desactualización del marco construido con base en el censo de 2018), las dificultades de acceso en áreas de propiedad horizontal y problemas de inseguridad en áreas de estrato bajo, para incrementar la muestra en la respectiva etapa. Los reemplazos, para enfrentar los retos que lo anterior implica, se realizan considerando el "ordenamiento" aleatorio de cada marco realizado con base en el algoritmo Coordinado Negativo.

Para las viviendas, se socializa con el CTS y el supervisor de contrato, un protocolo para la selección proporcional y asignación del tamaño de viviendas en función de la cantidad de unidades identificadas en el recuento, así como del estrato, diferenciando en las UPM por tamaño poblacional. Lo anterior, con el fin de mitigar la pérdida, así como el sesgo relacionado con la subcobertura de personas de 18 años o más de los estratos 4, 5 y 6.

En cuanto a los hogares y personas seleccionadas mediante el algoritmo Coordinado Negativo incorporado en el Dispositivo Móvil de Captura (DMC), no se realiza reemplazo. Es decir, si el informante de un hogar rechaza la encuesta, esto no implica inmediatamente seleccionar un reemplazo del hogar. Lo mismo aplica para el caso de la persona seleccionada. Lo anterior, porque se considera que esto resulta perjudicial en términos del sesgo que se genera, dado que el encuestador podría terminar abordando a las personas de hogares con mayores facilidades de acceso, de tal manera que no queden incluidos (representados) en la muestra, personas con determinados perfiles, como por ejemplo aquellas ocupadas laboralmente.

En relación con el tema anterior, es relevante mencionar que para cada unidad final de observación se realizan hasta tres visitas en hora y días diferentes, antes de asignarle el resultado de encuesta definitivo. No obstante, se incrementa en un 30% la muestra de personas de 18 años o más para mitigar la pérdida.

Excepto en circunstancias especiales, se considera que la estrategia propuesta es la más conveniente en términos de garantizar la mejor cobertura, sin afectar la rigurosidad técnica con la que deben implementarse los diseños probabilísticos en terreno. De otra parte, el desarrollar el algoritmo de selección en los DMC, permite tener

el control total de las unidades seleccionadas, eliminando el sesgo asociado a este procedimiento, como se indicó anteriormente.

Asimismo, el aplicativo incorporado en los DMC permite tener acceso a la información y por ende facilita el control y seguimiento de la cobertura. Para esto, la compañía desarrolla reportes para los distintos actores (encuestadores, supervisores), con el fin de evaluar el desempeño, analizar las causas de pérdida para tomar acciones preventivas y garantizar el cumplimiento de las metas (tamaños mínimos de muestra).

Con respecto a los reportes para control y seguimiento, en su construcción se tienen en cuenta las unidades de muestreo, las unidades de información, las unidades de seguimiento y los resultados de contacto a nivel de las unidades de muestreo que se definan.

Las unidades de seguimiento están relacionadas con el nivel geográfico (región, departamento, municipio, zona y estrato socioeconómico), el recurso humano (ruta de trabajo, supervisor y encuestador) y los resultados de contacto definidos; por ejemplo, en el nivel de las personas seleccionadas, lo antes descrito permite establecer la razón especifica de no respuesta (rechazo, ausente, ocupado, acceso difícil por circunstancias de orden público, etc.).

Finalmente, se informará oportunamente al supervisor de contrato y al CTS sobre el avance, generando informes quincenales, conforme lo indica la entidad en el anexo técnico, y adicionalmente se contará con un reporte en línea (tablero visualizador), en el que el DNP podrá consultar el avance del trabajo de campo.

ii. Descripción del plan muestral para la elección de los posibles reemplazos en el marco de muestreo

Para la selección probabilística de las unidades de muestreo se desarrolló un código computacional en R, ajustado a los lineamientos (métodos y algoritmos) definidos en la propuesta y se programaron en el aplicativo de captura los algoritmos correspondientes para la selección de viviendas, hogares y personas (M.A.S.).

La muestra inicial, es decir, la definida antes de la salida para realizar la recolección de las encuestas en los municipios de la muestra, se entrega al CTS junto con el código desarrollado y la semilla utilizada para la selección. Con esto, se demuestra el rigor técnico y objetividad en la implementación del diseño probabilístico comprometido, dado que, si el supervisor de contrato o el CTS lo estima conveniente, replicará la selección y deberá encontrar los mismos resultados.

En caso de que el análisis de la cobertura, durante el desarrollo del operativo (informes de avance), proporcione indicios para concluir que se requiere ampliar la

muestra en determinadas unidades de muestreo (excepto municipios) para garantizar las metas de cobertura, se utilizará el mismo código computacional y semilla inicial, para generar los reemplazos respectivos, los cuales en el caso del M.A.S. quedan determinados desde la selección de la muestra inicial.

Es de recordar que el Sistema de Gestión de Calidad (SGC) bajo el cual se adelanta cualquier estudio relacionado con el levantamiento de información mediante encuestas (procesos de recolección, captura, validación y procesamiento y, supervisión de campo), permite una comunicación eficiente y la articulación de los distintos roles, entre los actores involucrados, propendiendo de esta manera por la búsqueda de los objetivos o metas en un escenario propositivo y colaborativo.

Lo anterior para advertir que la implementación del diseño probabilístico se desarrolla con el rigor estadístico que lo exige y todas las acciones, así como causas que las originaron estarán documentadas para su respectiva verificación por parte del supervisor de contrato y el CTS.

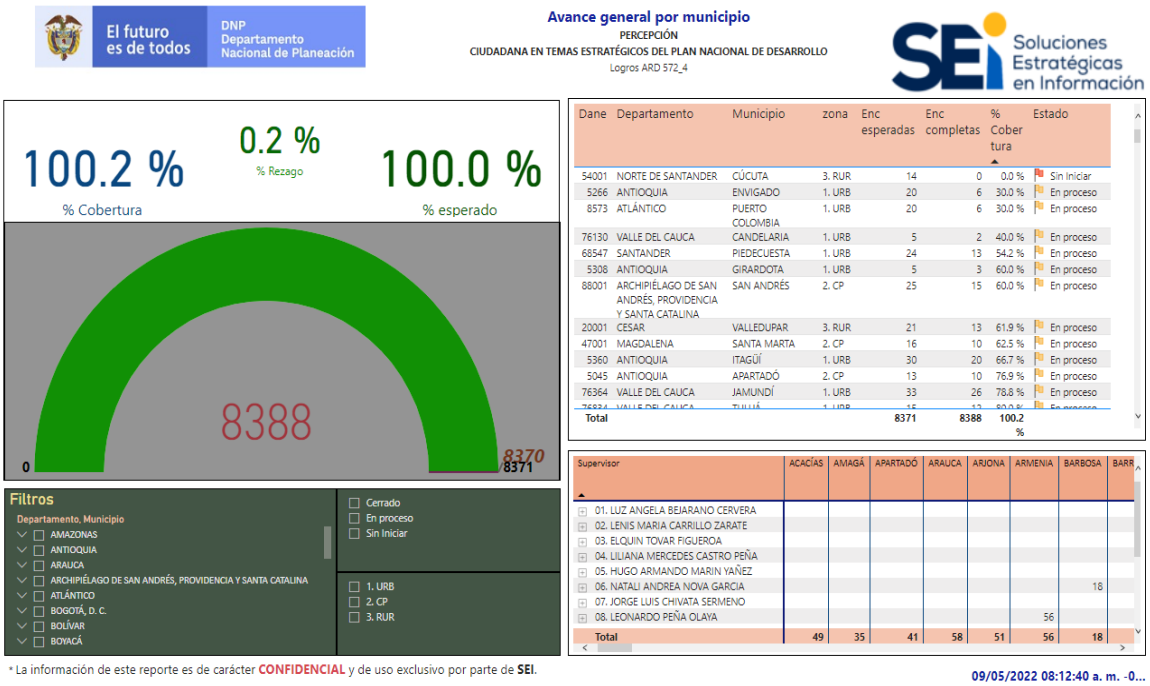
iii. Análisis de la no respuesta

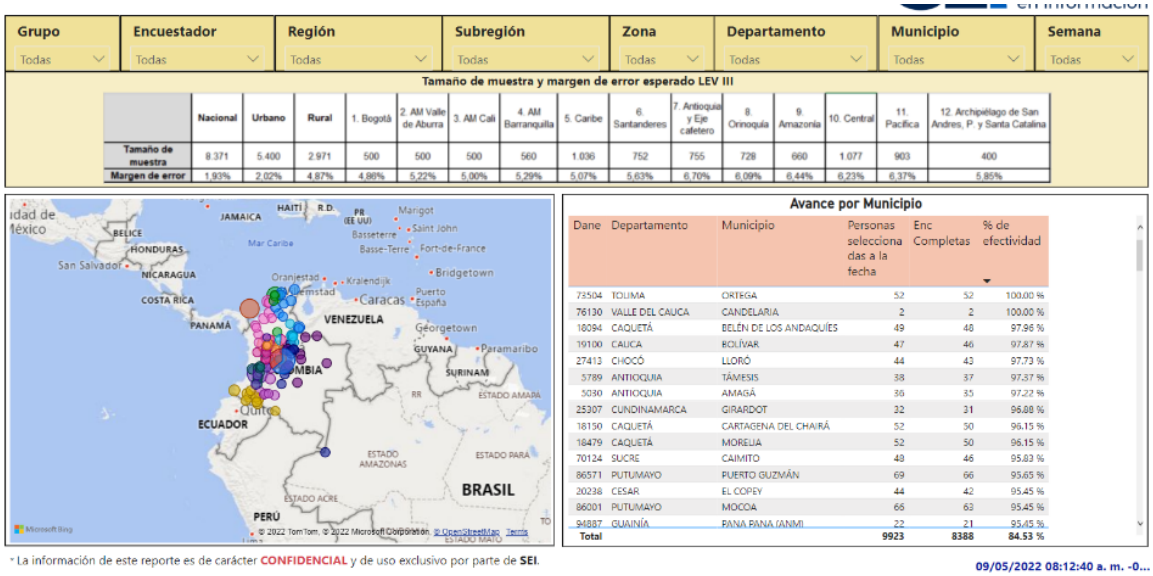
De acuerdo con los resultados de cobertura que se obtengan por unidad de muestreo, se realiza un análisis del comportamiento de la no respuesta en el nivel que corresponda o en cada etapa de muestreo.

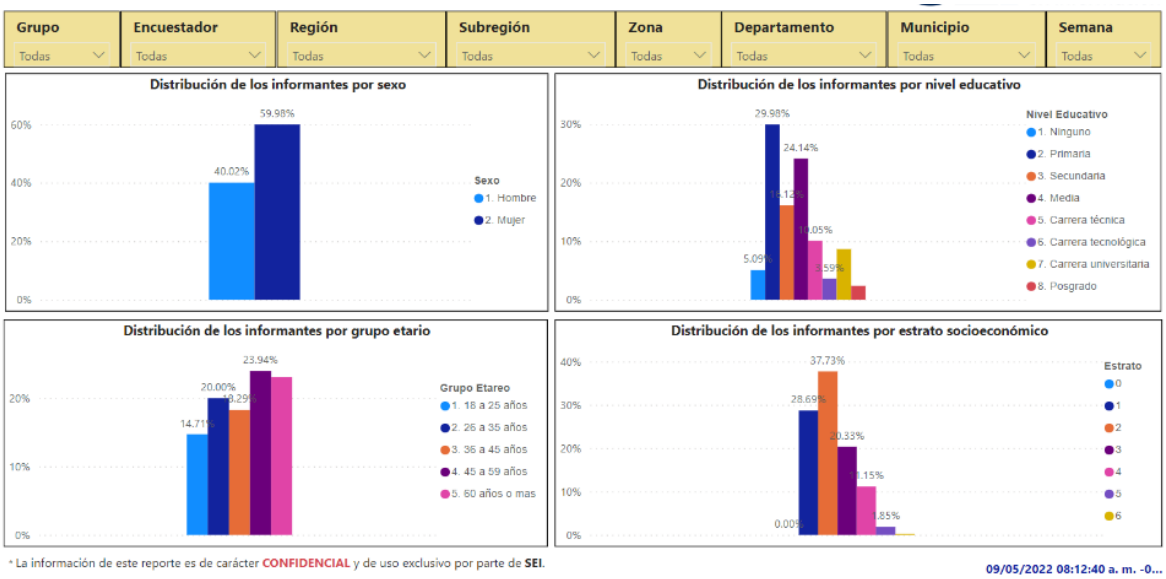
Al respecto, el seguimiento a las metas de cobertura se realiza a través de un reporte en la herramienta Power BI versión online, que se divide en hojas que permiten observar para distintos dominios de análisis el avance.

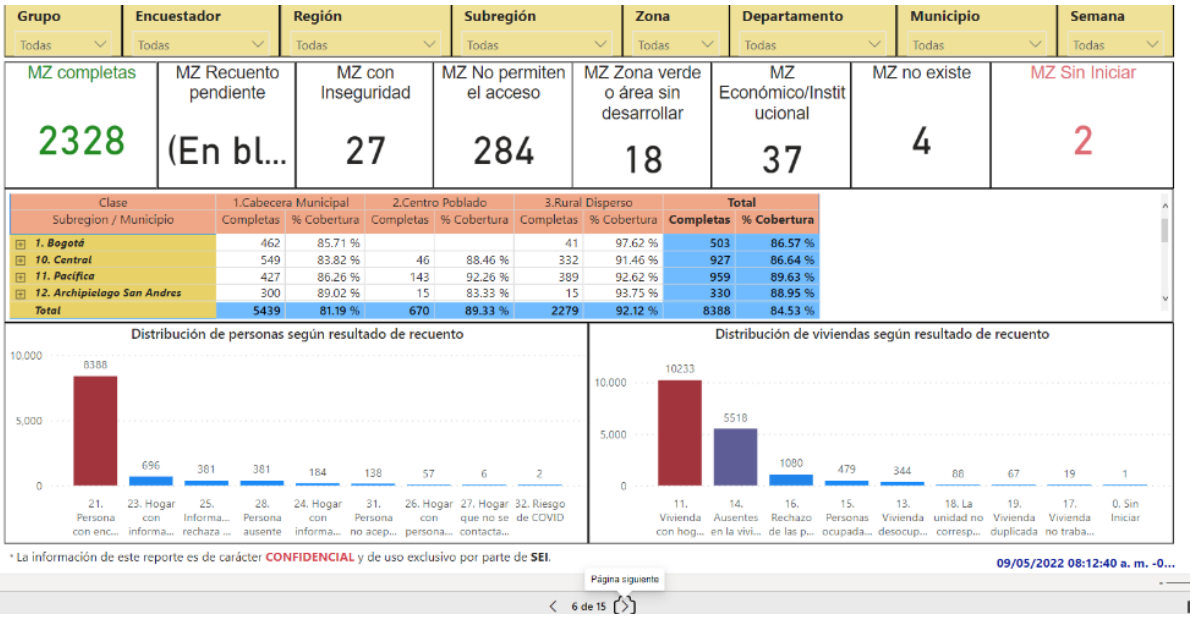
En la Figura 9 se presentan algunas salidas que evidencian lo antes mencionado y que hacen parte del reporte del levantamiento IV de la EPC PND 2018-2022.

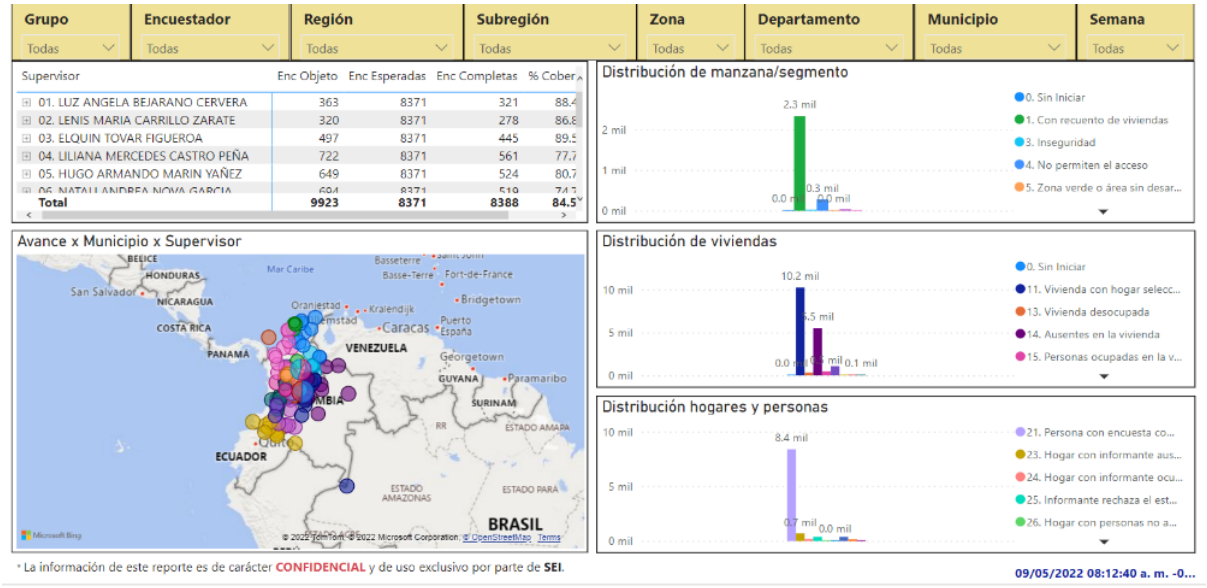
Figura 9. Avance por dominios de interés.






Grupo
Encuestador
Región
Subregión
Zona
Departamento
Municipio
Semana
Todas
Distribución de los informantes por sexo
59.98%
40.02%
Sexo
1. Hombre
2. Mujer
Distribución de los informantes por nivel educativo
29.98%
24.14%
18.12%
10.05%
5.09%
3.59%
Nivel Educativo
1. Ninguno
2. Primaria
3. Secundaria
4. Media
5. Carrera técnica
6. Carrera tecnológica
7. Carrera universitaria
8. Posgrado
Distribución de los informantes por grupo etario
23.94%
20.00%
19.29%
14.71%
Grupo Etareo
1. 18 a 25 años
2. 26 a 35 años
3. 36 a 45 años
4. 45 a 59 años
5. 60 años o mas
Distribución de los informantes por estrato socioeconómico
37.73%
28.69%
20.33%
8.15%
1.85%
0.00%
Estrato
0
1
2
3
4
5
6
La información de este reporte es de carácter CONFIDENCIAL y de uso exclusivo por parte de SEI.
09/05/2022 08:12:40 a. m. -0...


Grupo
Encuestador
Región
Subregión
Zona
Departamento
Municipio
Semana
Todas
MZ completas
2328
MZ Recuento pendiente
(En bl...
MZ con Inseguridad
27
MZ No permiten el acceso
284
MZ Zona verde o área sin desarrollar
18
MZ Económico/Institucional
37
MZ no existe
4
MZ Sin Iniciar
2
Clase
Subregion / Municipio
1.Cabecera Municipal
2.Centro Poblado
3.Rural Disperso
Total
Completas
% Cobertura
1. Bogotá 462 85.71 % 41 97.62 % 503 86.57 %
10. Central 549 83.82 % 46 88.46 % 332 91.46 % 927 86.64 %
11. Pacífica 427 86.26 % 143 92.26 % 389 92.62 % 959 89.63 %
12. Archipiélago San Andres 300 89.02 % 15 83.33 % 15 93.75 % 330 88.95 %
Total 5439 81.19 % 670 89.33 % 2279 92.12 % 8388 84.53 %
Distribución de personas según resultado de recuento
8388
696
381
381
184
138
57
6
2
21. Persona con enc...
23. Hogar con informa...
25. Informa... rechaza ...
28. Persona ausente
24. Hogar con informa...
31. Persona no acep...
26. Hogar con persona...
27. Hogar que no se contacta...
32. Riesgo de COVID
Distribución de viviendas según resultado de recuento
10233
5518
1080
479
344
86
67
19
1
11. Vivienda con hog...
14. Ausentes en la vivi...
16. Rechazo de las p...
15. Personas ocupada...
13. Vivienda desocup...
18. La unidad no corresp...
19. Vivienda duplicada
17. Vivienda no traba...
0. Sin Iniciar
La información de este reporte es de carácter CONFIDENCIAL y de uso exclusivo por parte de SEI.
09/05/2022 08:12:40 a. m. -0...
Página siguiente
6 de 15



Fuente: Elaboración propia.

Por otra parte, sobre la no respuesta observada, un escenario probable es construir un modelo de "propensión" para respondientes con base en los no respondientes. Específicamente, se ajustan diferentes modelos de clasificación con el fin de determinar si la no respuesta puede asumirse aleatoria. Un primer escenario consiste en un modelo multinivel[56] o jerárquico (modelo 1), donde la variable respuesta se define como la probabilidad de participar en la encuesta (éxito: encuesta efectiva), y tomando como variables explicativas algunas variables socioeconómicas[57], los niveles considerados para la jerarquía pueden ser: entidad territorial, municipio y zona de localización de la vivienda.

Así, la expresión matemática del modelo es:

$$\Pr\left(Y_{ijk}=1\middle|\mathbf{X},\mathrm{u}_i,u_{ij},u_{ijk}\right)=H\left(\boldsymbol{X'\beta}+\mathrm{u}_i+u_{ij}+u_{ijk}\right)=\frac{\exp\left(\boldsymbol{X'\beta}+\mathrm{u}_i+u_{ij}+u_{ijk}\right)}{\mathbf{1}+\exp\left(\boldsymbol{X'\beta}+\mathrm{u}_i+u_{ij}+u_{ijk}\right)},$$

Con $\mathrm{u}_i \sim N(0,\sigma^2_{et})$, $\mathrm{u}_{ij} \sim N(0,\sigma^2_{mpio})$ y $\mathrm{u}_{ijk} \sim N(0,\sigma^2_{zona})$ efectos aleatorios asociados a la entidad territorial, municipio y zona respectivamente; donde, $\boldsymbol{X}$ es el vector de variables explicativas y $\boldsymbol{\beta}$ el vector de coeficientes del modelo.

[56] Esto con el fin de tener en cuenta el diseño muestral multietápico implementado.
[57] Las cuáles serán seleccionadas de común acuerdo con el CTS y corresponden a variables observadas en todos los levantamientos.

El segundo escenario es un modelo lineal generalizado (MLG), considerando como variable respuesta la probabilidad de participar en la encuesta y variables explicativas la entidad territorial, la zona, el estrato socioeconómico, entre otras. De acuerdo con lo anterior, la expresión matemática del modelo anteriormente presentado queda expresada por la siguiente ecuación:

$$\Pr(\mathbf{Y} = 1|\mathbf{X}) = \frac{\exp\left(\boldsymbol{X}'\boldsymbol{\beta}\right)}{\mathbf{1} + \exp\left(\boldsymbol{X}'\boldsymbol{\beta}\right)}$$

Donde, $\boldsymbol{X}$ es el vector de variables explicativas y $\boldsymbol{\beta}$ el vector de coeficientes del modelo.

Para comparar los diferentes modelos que se planteen, se tendrá en cuenta el criterio de AIC. También se realiza un ANOVA para observar otro criterio de comparación entre los modelos, para así, finalmente seleccionar un modelo y estimar la propensión de respuesta.

Por otra parte, considerando lo anterior y como es sugerido en Valliant et al. (2018), es posible implementar un ajuste de no respuesta a partir del *propensity score*, lo que consiste en ajustar los pesos muestrales por el factor $1/\hat{\phi}_i$, donde $\hat{\phi}_i$, es la propensión de respuesta estimada para la unidad i, calculada a través del modelo seleccionado.

Así, el factor de expansión ajustado finalmente será dado por:

$$Fex_{p,ajust} = \frac{Fex_{final_p}}{\hat{\phi}_i}$$

Una vez se cuenta con el factor de expansión ajustado (por no respuesta) se implementará la técnica de calibración. Específicamente, se construirán pesos w_k con base en los pesos iniciales $Fex_{p,ajust} = f_k$ y un vector de información (v_k) construido con base en la información secundaria disponible. Lo anterior, se expresa como:

$$w_k = f_k v_k,$$

Donde $v_k = 1 + \lambda'_r y_k$ y $\lambda'_r = (Y - \sum_r f_k y_k)'(\sum_r f_k y_k y'_k)^{-1}$, con y_k como el valor de la variable observada a través de la encuesta y Y como el vector de totales poblacionales de las variables auxiliares y el vector λ_r es determinado usando la ecuación de calibración: $\sum_s w_k x_k$. Lo anterior, de acuerdo con lo propuesto por Estevao y Särndal (2000).

Por último, se implementa 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.

i) Descripción de una estrategia estadística en caso de llegar a presentarse ausencia de respuesta no ignorable

Es posible que se presenten dos tipos de "no respuesta". La primera se tipifica cuando las personas seleccionadas se niegan a responder algunas de las preguntas de la encuesta. Otra, cuando no es posible encuestar a la persona seleccionada después de agotar todos los esfuerzos posibles.

Cuando la no-respuesta se presenta en algunas preguntas, existen métodos de imputación para simular los valores faltantes con modelos basados en la información disponible de los demás encuestados. Särndal & Lundström (2005) recomiendan no olvidar que la imputación de un dato faltante, por definición "datos artificiales", tiene un error asociado, el cual se denomina "error de imputación".

Softwares estadísticos como STATA, SAS, R y SPSS, entre otros, incluyen, técnicas como la imputación múltiple (IM) (Rubin (1987)). El término imputación hace referencia a que valores faltantes son reemplazados con un conjunto de valores plausibles a partir de la información que se pudo recolectar en los otros registros de la muestra, incluso, usando información existente en el universo. La técnica de imputación múltiple consta de 3 pasos:

1. Imputación.
2. Análisis de datos completos (Estimación).
3. Agrupación.

Esta técnica busca solucionar los problemas de la imputación simple, relacionados con la introducción de sesgos y la distorsión de la potencia de las pruebas de hipótesis (Medina & Galván (2007)). La técnica de IM se basa en métodos bayesianos, a través de los cuales, utiliza procesos de simulación de Monte Carlo para así sustituir los datos faltantes (paso 1). En cada simulación se analiza la matriz de datos completos a partir de metodologías estadísticas convencionales (paso 2), y finalmente, se combinan los resultados de las dos etapas anteriores, con el fin de producir estimadores robustos, junto a su error estándar, intervalos de confianza y otros procesos inferenciales.

Es importante hacer una revisión de los datos y de las tasas de no respuesta para evitar que se presenten mayores problemas con la imputación. Adicional a esto, se asumen los siguientes supuestos: (i) el esquema de datos faltantes es aleatorio, lo que significa que el patrón no depende de la misma variable a imputar, (ii) el modelo utilizado para generar los datos imputados es adecuado, en el sentido de que existe correlación

entre la variable a imputar y el vector de covariables, y (iii) el modelo de análisis guarda relación con el utilizado para llevar a cabo la imputación. Aunque se describen las diferentes técnicas de imputación que se podrían utilizar, es importante mencionar que la experiencia que se tiene en el desarrollo de aplicativos robustos para captura y validación, permite garantizar la integridad de la información, haciendo que la posibilidad de que se necesite realizar imputación sea baja.

En el otro escenario, cuando la sobre muestra no logra cubrir el fenómeno de ausencia de respuesta, es necesario llevar a cabo una calibración de los factores de expansión. La calibración se basa en dos principios básicos, por un lado, (1) ser muy similares a los pesos muestrales planteados durante el diseño muestral, y por el otro lado, (2) utilizar información auxiliar para lograr que los resultados del muestreo puedan ser expandidos a la población de estudio. Los dos objetivos principales que se persiguen con la calibración son, en primer lugar, la reducción de la varianza de los estimadores, y, en segundo lugar, la reducción del sesgo debido a la no respuesta.

Si lo que se busca es estimar el total $t_y = \sum_U y_k$, el estimador de Horwitz Thompson de respuesta completa estará dado por:

$$\hat{t}_{yRC} = \sum_s f_k y_{*k} = \sum_s f_k y_k + \sum_{s-r_i} f_k \hat{y}_k$$

Donde $y_{*k} = \begin{cases} y_k \; para \; k \in r_i \\ \hat{y}_k \; para \; k \in s - r_i \end{cases}$ son el resultado de los valores de la variable de la encuesta, y_k son los valores observados y $\hat{y}_k$ son los valores imputados en para los registros faltantes.

Para implementar lo anterior, se utilizan los siguientes comandos en Stata: Calibrate y Calibest. Este último, es un estimador que tiene en cuenta la calibración realizada con el comando Calibrate.

ii) Presentación de los factores de expansión ajustados por la ausencia de respuesta

El problema de la no respuesta siempre se presenta en estudios poblacionales de magnitud similar al que nos ocupa. Básicamente, la pérdida de muestra si no se ajusta, no permite restituir el tamaño absoluto del universo ideal de estudio. Asimismo, la no respuesta puede no tener un comportamiento aleatorio, lo que conduce a tener estimaciones sesgadas por no contemplar esta pérdida diferencial. Por lo anterior, y considerando que se implementará un modelamiento para el ajuste de la no respuesta, resulta imprescindible implementar un ajuste a los factores básicos de expansión (inverso multiplicativo de las probabilidades de selección o inclusión de las unidades de muestreo seleccionadas), con el propósito de obtener estimaciones puntuales ajustadas al

universo de referencia, buscando reducir los sesgos que son inducidos por la presencia de la no respuesta.

Además del ajuste de la no respuesta a partir de un modelo de clasificación, será necesario implementar la técnica de calibración, con lo cual se busca afectar los pesos muestrales en la menor medida, pero que al final, la restitución de los tamaños absolutos se realice sobre la población objetivo de referencia, con base en información auxiliar disponible.

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 con los pesos iniciales f_k ajustados a partir de un vector de información. Estos estarán dados por:

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

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

Adicionalmente, se propone considerar el ajuste clásico de no respuesta a los factores básicos de muestreo para determinadas etapas. Este ajuste, permite mejorar la validez de las estimaciones, en el sentido de que permite corregir defectos del marco de muestreo y la no respuesta por razones de diversa índole, por ejemplo: rechazo, informante idóneo ocupado, ausente u otro.

Este ajuste es factible si es sustentable la hipótesis de que los resultados de unidades no encuestadas son, en promedio, similares a los de las encuestadas, dentro de una misma unidad de muestreo (UPM, USM o UFM). Para ello, se consideran las metodologías propuestas por (Bautista, 1998) y por (Valliant, 2018).

Al respecto, el factor de ajuste básico propuesto por (Bautista, 1998) por No Cobertura o No Respuesta tiene la siguiente expresión matemática genérica:

$$f_{ajuste} = \frac{n + n_{ad} - n_{fu}}{n + n_{ad} - n_{fu} - n_r}$$

Donde:

- n : cantidad esperada de la muestra.
- n_{ad}: cantidad adicional de elementos o unidades observadas que aparecen por defectos en el marco.
- n_{fu} : cantidad de elementos o unidades fuera del universo.
- n_r : cantidad de rechazos o no respuesta propiamente dicha (rechazos o ausentes).

Una vez determinado este factor se multiplica por el factor básico final, para generar el factor de expansión ajustado por la no cobertura. De otra parte, y como propuesta técnica para el manejo de los "outliers", se sugiere suavizar el efecto de los valores atípicos observados en los factores de expansión calibrados, a través de la metodología propuesta por Plotter (1990, 1993), la cual se caracteriza por:

- Establecer los límites inferiores (*L*) y superiores (*U*) para los factores de expansión. Este paso es arbitrario y depende del contexto y experiencia del investigador.

- Cualquier valor superior al límite propuesto debe recalcularse. Es decir,

$$w_{k,trim} = \{U \; si \;\; w_k \geq U, \quad w_k si \;\; L < w_k < U, \qquad L \; si \;\; w_k \leq L$$

  Donde $w_{k,trim}$ es el conjunto de pesos suavizados.

- Se calcula el valor total neto de peso perdido por el proceso de suavizamiento.

$$K = \sum_{k \in S} |w_k - w_{k,trim}|$$

- Se distribuye *K* uniformemente entre las unidades cuyos factores de expansión (i) no fueron recortados al evaluar el límite superior, o (ii) entre las unidades que tienen factores de expansión menores al límite superior, y que se ha definido inicialmente tanto *L* como *U*.
- Los pasos anteriores se repiten hasta que no haya pesos fuera de los límites.

La metodología anterior se encuentra implementada en la función trimWeights del software libre estadístico R versión 4.1.1 del paquete survey (Lumley, 2017).

1.4. Diseño del operativo de recolección en campo

Los municipios seleccionados se agruparán en rutas de acuerdo con su proximidad y vías de comunicación entre ellos y con la cantidad de labor representada en el número de encuestas a realizar, de manera que se obtengan cargas de trabajo y tiempos de vinculación equivalentes y así garantizar que todos los grupos estén terminando casi al mismo tiempo.

En el Anexo 4, se presentan las rutas de campo diseñadas para la recolección. Se tiene previsto el esquema de un supervisor por cada cuatro o cinco encuestadores, dependiendo de la ruta. Cada supervisor reportará al coordinador de trabajo de campo el avance, y es así como la logística requerida para adelantar el trabajo de campo se puede resumir de manera general así:

1. Los supervisores contarán con la cartografía, tanto de la cabecera municipal como de los centros poblados y veredas seleccionados, con la ubicación de las manzanas seleccionadas desde la oficina central para el caso de los dos primeros. Para la zona rural dispersa, será necesario hacer el levantamiento del marco muestral de las veredas seleccionadas en la muestra, para que a partir de allí se pueda realizar la selección de hogares. Esta labor requiere de gestión previa con las alcaldías de los municipios para obtener la información más reciente en cuanto a cartografía rural se refiere, además de la construcción de la distribución de viviendas y hogares en la vereda, lo cual se hace a través de cartografía social con el apoyo de representantes de la comunidad que pueden ser presidentes de las JAC, lideres veredales o personas conocedoras de cada vereda de la muestra.
2. Previo al operativo de campo se adelantará un recorrido en los conjuntos cerrados y edificios de las cuatro principales ciudades (Bogotá, Cali, Medellín y Barranquilla) con el fin de recoger información sobre el nombre de los conjuntos, administrador, teléfonos de contacto, etc. Lo anterior para agilizar el trámite de las cartas para la sensibilización de la comunidad previamente al operativo de campo.
3. En el DMC del supervisor se cargará la información necesaria para el control de cobertura y se instalará además un aplicativo de supervisión para registrar las técnicas empleadas con cada miembro de su equipo encuestador, para garantizar el seguimiento a su labor.
4. El sistema asignará de manera automática a cada encuestado un Número de Identificación Muestral (NIM) y registrará el código del encuestador a quien se le asigne cada unidad de observación.
5. Como previsión para los casos en que sea imposible la utilización de los equipos (por pérdida o daño de estos, o porque las condiciones de seguridad de la zona

donde se hará la aplicación no hacen viable el uso de DMC), se llevará una reserva de formularios impresos, los cuales serán diligenciados por el encuestador, revisados por el supervisor y digitados posteriormente (en terreno).

6. Diariamente cada encuestador deberá asegurarse de sincronizar la información, lo cual sucede tan pronto como tenga conexión estable a internet. En este proceso se descargan automáticamente las encuestas en el servidor de la firma, y mientras esto sucede, cada encuestador deberá asegurarse de hacer copia de respaldo de la información, conforme a las instrucciones impartidas durante la capacitación.
7. SEI proveerá el soporte técnico que sea necesario durante todo el desarrollo del proyecto de recolección, con personal de su planta, experto en sistemas, que será asignado desde un comienzo a este proyecto.

Se definirán estrategias orientadas a sortear las dificultades relacionadas con el orden público y otras eventualidades que pueden tener impacto en el proyecto, como es el caso del abordaje de los estratos altos o las zonas de difícil acceso por orden público o seguridad.

Esquema operativo

Considerando todo el contexto del estudio, el desarrollo del trabajo en cada ruta se resume de manera general, así:

1. El DNP enviará comunicaciones a las administraciones municipales seleccionadas en la muestra para informar del desarrollo del estudio y solicitar su colaboración para promover la participación de la población. Asimismo, entregará cartas personalizadas para la identificación del personal encuestador, las cuales serán portadas para mostrar a quien lo solicite.
2. El supervisor recibirá un cronograma donde se le indica el orden y las fechas estimadas en que debe abordar cada municipio.
3. De ser posible, previamente a la llegada al municipio el supervisor se contactará telefónicamente con las autoridades municipales para informar sobre el arribo del grupo, aclarar los objetivos del estudio, indagar sobre la situación de orden público y presentar a la firma. La información que se obtenga sobre cada vereda o zona será evaluada para definir la viabilidad de que el equipo ingrese allí. Cuando la situación lo amerite el Coordinador de Trabajo Campo informará a la dirección del proyecto y al CTS para evaluar situaciones específicas y tomar decisiones al respecto.
4. Al llegar al municipio, el supervisor se presentará a las autoridades locales pertinentes para socializar el proyecto y solicitar su colaboración. Como parte de los protocolos de seguridad de la compañía, el supervisor indagará con las

autoridades competentes la situación actual de orden público, principalmente en los centros poblados y veredas a visitar dentro del municipio

5. El supervisor se encargará de organizar el trabajo de su grupo distribuyendo entre sus encuestadores las manzanas y veredas a visitar, de tal forma que se optimice el uso de los recursos como son tiempo, transportes, etc.
6. Tanto encuestadores como supervisores portarán además chaleco y carné con distintivos de SEI. La identidad de cada una de las personas de campo, y por ende la veracidad del estudio, puede ser consultada a través de la página www.sei-consultores.com.co o en la sede central, cuando alguien así lo requiera, donde encontrarán el botón denominado "verificación de personal"; al dar clic allí se podrá digitar el número de cédula de la persona que se requiere confirmar. Si no fuera posible el acceso a internet, estará disponible el número telefónico de un delegado para brindar información sobre el estudio, quien podrá ser consultado durante el tiempo que dure el operativo de recolección.
7. Cada encuestador realizará el recuento de viviendas y hogares de acuerdo con la metodología diseñada para tal fin.
8. Cada encuestador visitará los hogares y abordará a la persona seleccionada de 18 años o más dentro del hogar, para aplicar el formulario (digital) de acuerdo con las normas suministradas durante la capacitación y al manual del encuestador (Anexo 6) y el protocolo de seguridad.
9. Antes de cualquier aplicación deberá hacer la lectura del consentimiento informado al informante, el cual será grabado, en todos los casos, como constancia de la aceptación y participación voluntaria en el estudio.
10. Cada supervisor deberá aplicar las técnicas programadas para verificar la calidad de la información recolectada por su grupo.
11. Frecuentemente (por lo menos una vez al día), cada encuestador deberá sincronizar[58] para enviar al servidor la información recopilada en su jornada.
12. Como medida preventiva para evitar la pérdida de información en caso de robo o pérdida, el encuestador deberá hacer una copia en un dispositivo externo, cada vez que finalice una encuesta. De este modo la información que no haya podido sincronizarse aún, estará salvaguardada ante cualquier eventualidad.
13. Periódicamente el supervisor enviará los materiales de muestra actualizados, los formularios y formatos que se hayan diligenciado en físico y el listado de muestra con los resultados registrados en el sistema por unidad de observación.

[58] La sincronización está configurada dentro del aplicativo de recolección para ser usada de manera controlada por los encuestadores, pues son ellos los que pueden garantizar la conexión a redes wifi o móviles. La información se captura de forma offline y puede ser sincronizada las veces que el encuestador tenga a disposición conectividad. La información capturada es encriptada en todo momento cuando sincroniza se transmite directamente a servidores de SEI los cuales almacenan todos los archivos de información (datos, imágenes, grabaciones, puntos GPS, etc.) para ser procesados directamente por personal autorizado. De esta forma se garantiza en todo momento, la seguridad de los datos y que nadie ajeno al equipo de trabajo de recolección tenga acceso a ellos.

Debido a que se ha vuelto una situación común que la población en general se encuentre prevenida para atender a los encuestadores en sus hogares, es de interés de SEI lograr un acercamiento adecuado con las comunidades que se visitarán con información clara y precisa que no genere falsas expectativas (por ejemplo: que no tendrán beneficios individuales ni contraprestaciones por colaborar en el estudio) pero que además contrarreste los efectos negativos que produce la desinformación y la difusión de noticias falsas por las redes sociales o por el voz a voz, con el ánimo de generar pánico, o simplemente como consecuencia de la conducta de personas irresponsables que no son conscientes del daño que causan con esas acciones.

Para el caso en que se divulgue una noticia falsa que pueda impactar negativamente el estudio, se tendrá previamente establecido el procedimiento a seguir, que incluirá entre otros aspectos, la definición de un comité para analizar la situación y orientar las acciones a emprender. Dicho comité incluirá a un responsable de comunicaciones, a la dirección del proyecto y a la coordinación de trabajo de campo.

1.4.1. Método, estrategia de recolección y dispositivo móvil de captura a usar.

La información será obtenida por entrevista directa, cara a cara, seleccionando una muestra probabilística representativa para toda la población de 18 años o más del territorio nacional, utilizando, mientras las condiciones lo permitan, Dispositivos Móviles de Captura (DMC), específicamente **smartphones**[59], en lugar de formularios impresos, todo ello manejado bajo un robusto esquema de automatización que facilita el ingreso de información, el control logístico del operativo y la generación de datos.

Los smartphones permiten almacenar grandes cantidades de información y procesarla de forma eficaz, garantizando conectividad con redes que permiten acceder a internet por varios medios (WIFI-Ethernet-Móvil), lo que facilita la transmisión de datos sin importar su ubicación geográfica.

Los smartphones que utiliza actualmente SEI son de última generación, con especificaciones técnicas de alta gama, que permiten un óptimo rendimiento en operativos de recolección intensivos como el que nos ocupa, donde se requiere larga duración de las baterías y rapidez en el procesamiento, cumpliendo siempre con las normas de seguridad y manejo de los datos recolectados. Las especificaciones técnicas que ofrecen este tipo de dispositivos son las siguientes:

- Pantalla de 5.7”, 1440 x 720 pixeles

[59] El smartphone corresponde a un teléfono inteligente que combina las funciones de un teléfono celular tradicional (móvil) con las de un ordenador, por su capacidad de almacenamiento y velocidad de procesamiento, sumado a la alta eficiencia de conectividad respecto a otros tipos de dispositivos móviles. Consultado en: http://www.informeticplus.com/que-es-un-smartphone

- Sistema operativo Android v8.0 y 8.1 Oreo MIUI 9.5
- Batería de 3000 MAH
- Memoria RAM de 3 GB
- Memoria Rom de 32 GB
- Slot Memoria: Micro SD hasta de 256 GB
- Cámara trasera de 13 MP y delantera de 8 MP
- CPU de 1.4 GHZ OCTA-CORE y Media Tek Helio P22 OCTA-CORE
- Peso aproximado de 146 gramos

El diligenciamiento de la encuesta se realizará mediante entrevistas individuales y presenciales, que tendrán la característica de ser dirigidas; se desarrollarán cara a cara entre un encuestado y un encuestador, y permitirán conocer la opinión y los puntos de vista acerca de varios temas.

El formulario de la encuesta será aplicado de manera estricta, respetando el fraseo de cada pregunta y de cada opción de respuesta, con el fin de evitar inducir la respuesta del encuestado. Este tipo de cuestionarios estructurados no permiten profundizar en otros temas, dado que no tendrá preguntas con opción de respuesta abierta, para garantizar la comparabilidad con las mediciones históricas y facilitar el procesamiento de la información.

Tampoco será posible que el encuestador sugiera o aplique nuevas preguntas que surjan durante la interacción de la encuesta, ya que deberá ceñirse a la aplicación rigurosa de la secuencia de preguntas conforme se haya diseñado el cuestionario. De este modo, las respuestas a cada pregunta generalmente son muy breves, dado que en muchas de ellas se aclara inicialmente cuáles serán las posibles opciones.

Aplicar encuestas a través de esta metodología es muy útil cuando se hace necesario abordar una muestra grande con muchas encuestas, porque al tener un cuestionario estructurado se minimiza el riesgo de subjetividad por parte del encuestador. En función de la longitud del formulario es posible encontrar un equilibrio entre una visión racionalizada y optimizada del tiempo del encuestador y una visión flexible y humana del tiempo del informante[60].

El uso de DMC y formularios digitales garantiza un proceso de captura que no se limita a convertir en digital las respuestas del encuestado, sino que logra que la información sea consistente y que el operativo de recolección se apoye en sus bondades

[60] Con base en consulta de: *http://ec.europa.eu/europeaid/evaluation/methodology/examples/too_itw_res_es.pdf*

para controlar aspectos como asignación de la muestra, cobertura, generación de reportes.

Entre las razones que promueven el uso de dispositivos móviles y formularios digitales se encuentran[61]:

- Control de implementación del diseño muestral cuando se requiera, garantizando aleatoriedad en la selección.
- Control de información detallada sobre cada unidad de observación con datos como resultado o estado.
- Menor tiempo de duración de la encuesta, siempre que el diseño del formulario así lo considere, dado que las preguntas tipo matriz podrían llegar a hacer lento el sistema.
- Mejor aceptación por parte de los encuestados y encuestadores por considerar atractivo el uso de la tecnología.
- Pérdida de datos evitable, siempre y cuando el personal siga estrictamente las instrucciones para hacer copias de seguridad.
- Disminución de errores en el registro de información, porque se incorpora al sistema una malla de validación[62] que controla la consistencia de los datos.
- Posibilidad de conversión inmediata de los datos a formatos como SPSS, SAS, STATA y archivos planos delimitados.

Una vez aprobado el diseño definitivo del cuestionario y teniendo en cuenta los resultados de la prueba piloto, luego de los ajustes sugeridos y revisados por el DNP asimismo se ajustará el aplicativo a la medida para la captura inteligente del cuestionario (que se probará también en la prueba piloto), utilizando un software robusto (CSPro), de manera que cumpla exactamente los requerimientos del estudio. Para ello se diseñará una interfaz amigable que muestre las preguntas (una cada vez) en la pantalla del dispositivo, y permita ingresar las respuestas mientras controla de manera simultánea la consistencia de la información ingresada.

El sistema de procesamiento de encuestas (CSPro) es un paquete de software de dominio público utilizado para ingresar y editar datos provenientes de encuestas. CSPro

[61] Consultado en http://ccp.ucr.ac.cr/revista/. Mejora de la calidad y disminución de costos de censos y encuestas utilizando computadores de mano (DMC) de la Revista Electrónica "Población y Salud en Mesoamérica" - Volumen 5, número 1, informe técnico 2, jul - dic 2007 Número especial CRELES- Costa Rica: Estudio de Longevidad y Envejecimiento Saludable. Julio - diciembre, 2007 Publicado 1 de julio, 2007

[62] El equipo técnico construirá mallas de validación, consistentes en matrices que muestran las relaciones lógicas entre los diferentes capítulos y variables del formulario, establecen la secuencia de las preguntas y los códigos válidos, y permiten identificar inconsistencias y errores en los datos, partiendo de normas y rangos de valores aceptables. Estas mallas servirán de guía para programar las validaciones en el formato digital.

está diseñado con el fin de que sea muy fácil de usar, pero también para que sea lo suficientemente potente como para manejar las aplicaciones más complejas. CSPro admite la recopilación de datos en dispositivos Android (teléfonos y tabletas) a través de la aplicación de Android CSEntry funciona en colaboración con la versión de escritorio de CSPro la cual es descargable de la Play Store de Google.

Por sugerencia del CTS y como resultado de las discusiones sobre el formulario, se acordó agrupar los capítulos en los siguientes paquetes:

- Primer paquete: Sociodemográficos, Conectividad e Internet, Gobierno y Generales, Transparencia, Pobreza y Desigualdad (Bloque fijo)
- Segundo paquete: Seguridad, Niños, Niñas y Adolescentes, Situación del campo, Justicia, (bloque que se rotará).
- Tercer paquete: Educación, Salud, Trabajo, Cultura y Recreación, Infraestructura y Transporte (bloque que se rotará).
- Cuarto paquete: Minas y Energía, Barrios, Viviendas y Servicios Públicos, Ambiente, Paz y Migraciones, Otros Temas Estratégicos del Gobierno Nacional (bloque fijo).

Lo anterior con el fin de rotarlos durante la aplicación del cuestionario a lo largo del operativo de campo, de la misma manera se rotarán dentro de algunas preguntas las opciones de respuesta, para el caso en que el encuestador debe leerlas, como ya se mencionó anteriormente.

1.4.2. Plan de prueba piloto.

La prueba piloto es un evento previo a la adopción definitiva del formulario y de los procedimientos diseñados para el levantamiento de la información, bajo condiciones reales y con fines de evaluación. Para evitar sesgos en la evaluación de los resultados, la prueba piloto replicará las condiciones propias de la recolección de información en lo que tiene que ver con el instrumento que se utilice, el diseño muestral, y la metodología de identificación, ubicación y selección de las personas mayores de 18 años, todo lo cual debe reflejar las limitaciones y diversidad de situaciones que se encontrarán en el desarrollo del proyecto como tal.

Para la prueba piloto se organizará la capacitación de manera presencial y su organización debe ser similar a la que se utilizará en el proyecto, las temáticas a dictar son la presentación de las consideraciones generales del estudio, el formulario de encuestas, además, debe incluir los aspectos relacionados con la metodología de la prueba piloto y los formatos de esta para la evaluación de resultados.

Al siguiente día de terminada la capacitación, el grupo de campo se dirigirá al primer municipio de la muestra y al llegar se presentará a las autoridades municipales para explicar los objetivos del estudio y solicitar colaboración.

El primer día del operativo, el supervisor se encargará de asignar a los encuestadores las manzanas seleccionadas desde la oficina principal, y los encuestadores realizarán los recuentos de viviendas y hogares de acuerdo con la metodología establecida y como resultado de esta actividad, se seleccionarán las personas a encuestar dentro de cada una de las manzanas. Esta operación se hará tanto en la cabecera municipal como en los centros poblados seleccionados.

En el mismo día los encuestadores se dirigirán a los hogares y personas seleccionadas, para iniciar las respectivas encuestas. En el segundo día y así sucesivamente, continuarán con las encuestas programadas hasta alcanzar la meta propuesta. Al finalizar cada jornada de trabajo se revisará y evaluará la totalidad de formularios diligenciados, y se diligenciarán los formatos diseñados para hacer la observación.

Los principales aspectos por revisar en la prueba piloto serán los siguientes:

- Evaluar la metodología para el levantamiento de la información, específicamente la utilización y funcionalidad del formulario y demás instrumentos dirigidos a personas mayores de 18 años.
- Estimar los tiempos de diligenciamiento. La prueba servirá para establecer tiempos y rendimientos promedio para la aplicación de la encuesta.
- Evaluar los procedimientos de campo.
- Evaluar el impacto que puedan tener las dificultades de orden público y la percepción de conflicto en el desarrollo del operativo.
- Estimar ventajas y dificultades, que se podrían presentar durante el operativo, respecto al manejo de la muestra.
- Establecer la facilidad de localizar y gestionar las encuestas con los informantes idóneos en los hogares seleccionados.

1.4.2.1. Evaluación de resultados de la prueba piloto

Se efectuará una evaluación cualitativa, que se realizará mediante discusiones en grupo focal con todo el personal participante de la prueba piloto (observadores y personal de campo), con base en los siguientes formatos:

a. Anotaciones del observador.

b. Reporte de problemas y sugerencias del formulario.
c. Reporte de problemas y sugerencias de los manuales.
d. Registro de tiempos de encuesta.
e. Registro de tiempos de los recuentos y otros como tiempos de espera, desplazamientos, gestión, etc.
f. Registro de problemas de orden logístico y de procedimientos.

La evaluación será soportada por la información cuantitativa, debido a que las encuestas serán aplicadas en Dispositivos Móviles de Captura.

Si los observadores y otras personas responsables indican que la actividad en general avanzó fácilmente, es decir, que encontraron poca dificultad en la aplicación de los procedimientos y los encuestados en general comprendieron las preguntas, entonces se puede aceptar que estos instrumentos y procedimientos son adecuados para el estudio. Si, por el contrario, las observaciones indican que hubo problemas, entonces será necesario efectuar modificaciones.

La evaluación subjetiva no implica la aceptación de todas las observaciones, sino su discusión y análisis para llegar a un consenso sobre las causas probables y el procedimiento indicado para eliminarlas. Para agilizar la discusión, cada participante debe llevar registros o anotaciones de sus impresiones sobre estos aspectos, con el fin de que la discusión se realice de una forma dinámica y centrada sobre los temas de interés.

Las discusiones de grupo cubrirán todos los aspectos sometidos a evaluación, con el fin de llegar a conclusiones o recomendaciones para el mejoramiento de los instrumentos y/o procedimientos; se discutirán todas las observaciones presentadas por los involucrados, así como las causas y posibles soluciones a los hallazgos.

Para esta fase se dispondrá de cuadros de resumen con los resultados de encuesta, tasas de omisión para preguntas específicas, así como opciones de respuesta frecuentes que no se encontraban en el listado inicial, además de un documento borrador en el cual se consolidarán las observaciones de todos los participantes.

Las conclusiones que se adopten serán el producto de discusiones entre los funcionarios de la compañía y el DNP, en el contexto de la metodología y los requerimientos del estudio, se hará una evaluación de tipo consensual, es decir, por acuerdo entre los participantes acerca de las conclusiones individualmente presentadas en las reuniones de discusión, la cual estará soportada en cifras o mediciones de tipo cuantitativo hasta donde sea posible.

El producto de estas discusiones será un informe final de conclusiones y recomendaciones que se presentará al DNP para que este defina las acciones a tomar para el ajuste de procedimientos o formulario.

1.4.2.2. Muestra para la prueba piloto

La muestra de municipios para la prueba piloto del levantamiento fue propuesta por SEI y avalada por el CTS. La propuesta se basó en la experiencia de la firma dado que la mayoría de preguntas del formulario han sido piloteadas en levantamientos previos de información y de ahí surge la recomendación de aplicar alrededor de 45 encuestas distribuidas en zona urbana y rural, para validar el lenguaje y estructura de las preguntas nuevas, aunque durante la prueba se aplicará la totalidad del cuestionario

Ahora bien, respecto a la distribución urbano-rural se tuvo en consideración el porcentaje nacional para cada zona. Los municipios escogidos se acordaron con base en la cercanía a la capital, además de variedad cultural.

En la tabla 5 se presenta la muestra de municipios y cantidad de manzanas seleccionadas.

Tabla 5. Municipios, cantidad de encuestas y número de manzanas para prueba piloto

Grupo	Municipio	Tipo de muestra	Cantidad encuestas	Número de manzanas
1	Bogotá	Cabecera	15	4 manzanas
2	Guamo – Tolima	Cabecera y Centro poblado	15	4 manzanas
3	Tunja - Boyacá	Cabecera y centro poblado	15	4 manzanas
Total 3 municipios			45	

Fuente: SEI

1.4.3. Cronograma de recolección.

En el Anexo 5 se presenta el cronograma por días y semanas de las principales actividades del levantamiento de información.

1.4.4. Estructura organizacional del operativo.

El proyecto será realizado por el siguiente equipo:

La directora del proyecto será responsable de asesorar todas las actividades relacionadas con el diseño operativo y logístico para la ejecución del proyecto, así como los procesos de diagramación de formularios y la elaboración de manuales para la

recolección de datos, la realización de la prueba piloto, y el sistema de capacitación. Se encargará, de hacer el seguimiento al cronograma de actividades, de ejercer las tareas relacionadas con el aseguramiento de la calidad, de mantener una relación permanente con el comité técnico, y de la coordinación de los informes relacionados con el trabajo de campo. De la dirección de proyecto dependerán el coordinador de trabajo de campo y el responsable de sistemas.

El coordinador de trabajo de campo tendrá a su cargo la elaboración de instructivos para la aplicación, la selección y capacitación de los supervisores y encuestadores, la dirección del proceso de recolección de datos, y el seguimiento del aseguramiento de calidad y cumplimiento del trabajo de campo.

El responsable de sistemas tendrá a su cargo la estructuración de tablas de datos, el procesamiento, consolidación, y análisis de datos y estadísticas, así como el seguimiento y control de calidad de la información. Diseñará y aplicará mallas de validación para la depuración de la información, asesorará la elaboración de los programas de captura y supervisará la captura de la información. Participará en las reuniones del comité técnico, y/o con el cliente, para discutir los requerimientos en cuanto a la presentación de la información y los resultados finales del proyecto.

La estadística será el responsable general del diseño muestral probabilístico, garantizar que la metodología relacionada con su implementación se ejecute con rigurosidad y que esto responda a las expectativas del comité técnico. Adicionalmente la estadística contará con un equipo de apoyo (una estadística y una experta en geomática) para la gestión de insumos de los marcos de muestreo y el análisis estadístico que se requiera, el cual estará bajo la responsabilidad del coordinador de estadística de SEI, cuya labor se orienta a apoyar la labor de la estadística principal y participar en la documentación técnica que se requiera. Específicamente se construirán los códigos computacionales que se requieran a partir de los cuales se seleccionarán las unidades de muestreo.

Previamente, recopilarán la información fuente para la construcción de los marcos de muestreo, elaborarán los protocolos para la implementación del diseño probabilístico en las Unidades Primarias de Muestreo (municipio o área metropolitana) seleccionadas y realizarán el control y seguimiento de las metas de cobertura. Finalmente, construirán y calibrarán los factores de expansión, realizarán el procesamiento de los resultados y elaborarán la documentación técnica de informes según el estándar estadístico requerido.

Los supervisores tendrán como principal función coordinar las visitas de su grupo, en cuanto al recuento de vivienda y hogares, selección de la persona mayor de 18 años dentro del hogar, controlar la cobertura asignada a su grupo y la calidad de la información por medio de las técnicas de supervisión. Es importante destacar que se trata de un operativo donde la coordinación logística juega un papel fundamental.

En el contexto específico de este estudio, los supervisores tendrán las siguientes funciones:

- Organizar el trabajo del grupo a su cargo, planificando día a día las rutas de trabajo de los encuestadores.
- Controlar el desarrollo del operativo y controlar que la cobertura se cumpla en cada uno de los municipios asignados en su ruta.
- Controlar la calidad de la información utilizando para ello las técnicas de supervisión de acompañamiento y verificación.

El cumplimiento de las funciones de supervisión implica, además, el desempeño de algunas actividades de tipo administrativo tales como: manejo de recursos, manejo de personal, contratación de transportes especiales, control de materiales, y reportes de trabajo.

En cuanto al manejo de personal, los supervisores deberán:

- Ayudar a los encuestadores para que puedan dar lo mejor de sí mismos, y asesorarlos permanentemente durante el trabajo de campo.
- Vigilar que las relaciones interpersonales del grupo sean óptimas y dar un tratamiento justo a los delegados.
- No permitir acumulación de trabajo en manos de un encuestador.

Los encuestadores serán los responsables de la localización de las manzanas seleccionadas, el recuento de viviendas y hogares y la selección de la persona mayor de 18 años, de acuerdo con la metodología establecida para ese fin, para encuestarla, haciendo uso del formulario respectivo.

Se tiene prevista la conformación de 33 grupos, cada uno con un supervisor y cuatro encuestadores.

1.4.5. Plan de entrenamiento del personal.

1.4.5.1. Convocatoria y selección de personal de campo.

Para el desarrollo de este operativo se cuenta con colaboradores permanentes a nivel de dirección y coordinación, y contará con personal temporal para la labor de recolección en campo. Es así como se conformará un equipo con encuestadores con experiencia certificada en encuestas poblacionales, quienes contarán con el respaldo de supervisores experimentados que los guiarán en la ejecución del operativo al interior de cada grupo. Los encuestadores serán los directos responsables de recoger la información a través de la aplicación de la encuesta a cada informante idóneo en la

muestra, y los supervisores desempeñarán un papel crucial al proporcionar orientación y apoyo continuo a sus grupos.

Al momento de organizar el proceso de convocatoria de personal temporal se determinarán los perfiles requeridos para cubrir cada cargo, siendo el mínimo solicitado, el que se presenta a continuación:

Perfil del encuestador: Personas con 18 años o más, con estudios aprobados de bachillerato y experiencia en aplicación de encuestas en por lo menos un proyecto de caracterización socioeconómica o de percepción, dirigidos a hogares o personas residentes en zonas urbanas y/o rurales.

Perfil del supervisor: Personas con 18 años o más, con al menos dos semestres de formación técnica, tecnológica o superior y experiencia en actividades de supervisión de encuestas en, por lo menos, dos proyectos de caracterización socioeconómica o de percepción, dirigidos a hogares o personas residentes en zonas urbanas y/o rurales.

Previo a la recolección se adelantará la convocatoria, capacitación y selección del personal temporal. Para iniciar con la convocatoria de personal temporal, se comenzará con la búsqueda en el SÍSEI[63]. Se requiere que el personal por convocar cuente con tiempo completo para el proyecto y disposición, tanto física como anímica, para este tipo de trabajos, excelentes habilidades de comunicación y calidez en el trato. Asimismo, quienes aspiren a ser encuestadores deberán demostrar una vez capacitados, el manejo adecuado del aplicativo de captura. Para contar con una cantidad suficiente de candidatos que permita cubrir eventuales deserciones (durante y posterior a la capacitación) e incumplimiento, se citará a capacitación a un número mayor de personas al requerido para contratar, además de contar con un remanente de eventuales reemplazos para cubrir casos de deserción durante la ejecución del operativo de campo.

A continuación, se hará una capacitación principal del personal que se adelantará en Bogotá, de este equipo base, se seleccionarán los capacitadores que se desplazarán a las principales ciudades (capitales de las áreas metropolitanas) en donde se replicará la inducción (capacitación en cascada). Las ciudades en las que se replicará la inducción y se hará capacitación con personal de la región son: Barranquilla, Cali, Medellín, Bucaramanga y Buenaventura. En general, la capacitación abordará las recomendaciones para aplicar adecuadamente la encuesta, el uso del aplicativo de captura, las técnicas de entrevista sugeridas para evitar sesgos y reducir la fatiga de los encuestados, así como los diferentes controles de calidad que tendrá la información.

[63] SÍSEI es un sistema de información de SEI en el que se almacenan los datos de quienes han trabajado o aspiran a trabajar con la compañía De ser necesario, se publica la invitación en la página y en las redes sociales o en páginas especializadas, para que nuevos candidatos postulen sus hojas de vida. Este proceso se hace a través de la aplicación "Sistema de Hojas de Vida" al que se puede acceder a través del enlace https://www.sei-consultores.com.co/trabaje-con-sei/

Además de los conceptos propios de la encuesta, con el fin de que se familiaricen con la terminología y se sensibilicen respecto a la relevancia de la encuesta, se incorporará un módulo orientado a impartir recomendaciones para el abordaje de la población objetivo, haciendo énfasis en la necesidad de adoptar un trato respetuoso, digno y diferencial con los informantes, basado en el enfoque de género e interseccional. La capacitación incluirá jornadas teórico prácticas, así como momentos para las evaluaciones, que servirán para hacer la selección del personal más idóneo. En lo referente a las jornadas prácticas, la capacitación para encuestadores incluirá el manejo específico del aplicativo de captura, mientras que la de supervisores contendrá además sesiones adicionales con el detalle de sus funciones específicas, los procedimientos de recolección, el control de calidad y el control de la cobertura.

La escogencia de quienes finalmente participarán en el proyecto se llevará a cabo al finalizar la capacitación. Una vez se conozca el listado de seleccionados, se procederá a hacer el trámite de contratación para organizar los temas administrativos respectivos. Los criterios para la selección del personal más idóneo se exponen en detalle en la siguiente sección.

1.4.5.2. Plan de capacitación.

La capacitación será programada con el supervisor de contrato designado por el DNP y estará basará en componentes de orden organizacional y metodológico. El propósito es desarrollar un proceso eficaz de transmisión de información y conocimientos, para brindar las herramientas suficientes que permitan al personal capacitado aplicar, durante el cumplimiento de su función, lo aprendido a lo largo de las sesiones de inducción.

Paralelamente a la convocatoria, el personal de apoyo y coordinación del estudio se encargará de alistar los materiales y equipos a utilizar en el transcurso de la capacitación, tanto en contenido como en cantidad. Se hará uso de herramientas pedagógicas, como videos explicativos y presentaciones de Power Point con los contenidos clave de cada tema, cuyo fin es facilitar la asimilación y comprensión de los temas expuestos. Estas ayudas se basarán en presentaciones, manuales, formatos y versiones finales del formulario, elaborados para el proyecto.

SEI recomienda que sean los propios funcionarios del DNP quienes presenten los objetivos del estudio y las generalidades de las diferentes temáticas abordadas en el formulario, para brindar el contexto general del proyecto a todos los asistentes.

Adicionalmente, para controlar la calidad de la capacitación, el personal convocado evaluará cada uno de los temas tratados, de acuerdo con las normas establecidas por la

compañía. Se estima que el tiempo prudencial para dictar la capacitación es de una semana, los cuales incluyen jornadas prácticas de diligenciamiento.

Se diseñarán evaluaciones teóricas y prácticas, para medir el nivel de apropiación de los conocimientos, así como para valorar las habilidades y experiencia del personal convocado. Estas evaluaciones se utilizarán para la selección del personal a contratar y para identificar los temas que necesitan refuerzo.

Una semana antes de la capacitación general se dictará la formación al equipo base en Bogotá, y de allí saldrán los capacitadores que se desplazarán a las principales ciudades para replicar la inducción, en aquellos sitios con mayor concentración de muestra o con mayor número de aspirantes a encuestadores, con el fin de optimizar los recursos disponibles. Para lo anterior, SEI convocará personal que trabaja habitualmente en proyectos de recolección y que reside en las principales ciudades del país, lo cual facilita la organización de estas capacitaciones en cascada. Esta consultoría implementará acciones tendientes a evitar el impacto que trae consigo la atomización de cursos, como son un taller sobre técnicas de capacitación y la elaboración de un instructivo que sirva como herramienta al capacitador y facilite un proceso de estandarización adecuado.

Tanto la agenda de cada capacitación, como el calendario general de las mismas, en las ciudades que se requieran, que en principio son las áreas metropolitanas, serán fruto de la concertación con el DNP para definir la estructura de capacitación que resulte más pertinente para las necesidades del estudio.

A continuación, se presenta el contenido con que se planea desarrollar los cursos en cada una de las instancias de la organización del proyecto (encuestadores y supervisores).

La capacitación del personal incluirá:

- Entrenamiento sobre manejo conceptual y contexto del proyecto.
- Diligenciamiento del formulario.
- Manejo del aplicativo y del smartphone.
- Técnicas de entrevista, encuestas simuladas y ejercicios escritos de diligenciamiento.
- Manejo de la muestra y del recuento.
- Manejo del instrumento de recolección.
- Aspectos administrativos y de contratación.

Adicional a lo anterior los supervisores se capacitarán sobre técnicas de supervisión para el trabajo de campo.

La capacitación se apoyará en dos **técnicas**: la primera corresponde a las exposiciones magistrales de un instructor o facilitador con sesiones de preguntas y aclaraciones e intervenciones del personal. La segunda, a un componente práctico dividido a su vez en ejercicios de entrenamiento en oficina y práctica de campo, donde esta última estará enfocada al ejercicio de hacer recuentos de manzanas para adiestrar al personal en el conteo y selección de unidades de observación.

Una vez finalizada cada una de las sesiones y los ejercicios de campo y oficina, se aplicarán pruebas de evaluación que servirán para seleccionar al personal que finalmente ejecutará el proyecto; se considera que las personas deben acumular por lo menos 60 puntos sobre 100 para ser aceptados. De quienes logren mínimo 60 puntos, se seleccionarán a los que tengan los mejores puntajes.

1.4.6. Estrategias de sensibilización.

Uno de los principales retos de este estudio consiste en convencer a las personas para que participen en la encuesta, y para esto, se requiere habilidad del personal de campo, destinado a contactar a los participantes.

Para ello, el equipo de recolección presentará de manera sencilla y clara los objetivos del estudio, así como la importancia de su participación, a cada persona seleccionada, y estará en capacidad de generar empatía y escuchar atentamente las inquietudes que surjan sobre el proceso. Asimismo, presentará la carta firmada por el DNP que lo respalda como parte del equipo de la firma y suministrará los números de contacto de los responsables del estudio para que se comuniquen cuando deseen ampliar la información; de esta forma se busca generar mayor tranquilidad y confianza.

Adicionalmente, SEI y sus colaboradores adoptarán durante todo el operativo, pautas de acción sin daño para el abordaje de las personas, que serán socializados en la capacitación del personal de recolección. Estas contendrán todos los aspectos a considerar para desarrollar la aplicación en un ambiente de respeto, desde el saludo, presentación, tono de voz, etc., enfatizando en la importancia de transmitir calidez y objetividad en todo momento.

Como funciones específicas en el momento de contacto, el personal de recolección debe cumplir a cabalidad las recomendaciones que se enuncian a continuación:

- Tener un comportamiento que haga honor a las instituciones que representa (cliente y SEI).
- Realizar el trabajo con seriedad y profesionalismo para garantizar la calidad de la información.

- Conocer los conceptos del instrumento para generar confianza y habilidad en el manejo de éste, ya que, si no se siente seguro al momento de la aplicación, no podrá desarrollar con éxito su tarea.
- Mantener serenidad en situaciones difíciles o incómodas con gente irritable, mal educada o mal intencionada, y conservar la calma para bajar el nivel de agresividad o incomprensión del interlocutor con el fin de propiciar un diálogo productivo y amable.
- Evitar entrar en discusiones de cualquier índole, tener discreción y ratificar la confidencialidad de la información;
- Ser neutral al comunicarse con los encuestados, evitando polemizar, ya que nunca se debe dar la impresión de aprobar o desaprobar lo que digan.

Adicional a lo anterior, se establecerán de manera puntual las siguientes estrategias:

- Contactar previamente a funcionarios de las alcaldías de municipios de la muestra para informar de la actividad y del envío de un comunicado de presentación del estudio, firmado por el DNP.
- Colgar un banner en la página de SEI, y en sus redes sociales, informando sobre el desarrollo del estudio y las líneas de contacto para ampliar información.
- Colgar un banner también en la página del DNP, y en sus redes sociales, informando sobre el desarrollo del estudio, así como las líneas de contacto para ampliar información.
- Al llegar a un barrio o sector, se deberá identificar y visitar los CAI de policía, juntas de acción comunal y líderes reconocidos por la comunidad para hacer presentación formal del estudio, informar acerca de los sitios a abordar y solicitar colaboración para facilitar la respuesta.
- Entregar cartas personalizadas a los administradores de conjuntos cerrados firmadas por el DNP donde se incluyan los objetivos del estudio, las entidades responsables y la metodología a emplear en el estudio. Esto con el fin de solicitar colaboración para la realización de las encuestas.

En el caso que resulten seleccionadas veredas o centros poblados de comunidades negras y/o indígenas se debe proceder de la siguiente manera:

- Se solicitará autorización a la autoridad indígena o al líder del consejo comunitario para el ingreso al resguardo o a la comunidad afro y para la realización de las encuestas. Se informará qué personas de SEI serán las responsables de la aplicación de las encuestas y con quiénes se pueden comunicar, en caso de tener alguna inquietud con respecto al estudio.

- En caso de obtener una respuesta negativa, se procederá a informar al DNP, para que por intermedio de ellos se realice un acercamiento, con el fin de obtener el respectivo permiso.
- En caso de que la respuesta sea afirmativa, se procederá a establecer con mayor precisión los siguientes detalles: Fechas y días de visita, requisitos que se deben cumplir para el ingreso a las comunidades (actividades y prácticas culturales) y la permanencia en ellas, solicitud de un traductor en el caso de que algunas de las personas a encuestar no hablen español y las demás que establezca las autoridades pertinentes.

Es importante tener en cuenta que la respuesta a la solicitud de ingreso, en algunos casos puede ser un poco demorada, por lo tanto, la gestión se hará de manera anticipada de ser posible.

Acerca del consentimiento informado

En cumplimiento de las condiciones de protección de la información y de conformidad con lo establecido en la Ley Estatutaria 1581 de 2012 y normatividad vigente en lo que a protección de datos personales se refiere, SEI reconoce y protege el derecho que tienen todas las personas a conocer, actualizar y rectificar las informaciones que se hayan recogido sobre ellas en bases de datos o archivos que sean susceptibles de tratamiento por entidades de naturaleza pública o privada. Cuando se habla de datos personales, se está haciendo referencia a toda información asociada a una persona y que permite su identificación, por ejemplo, su documento de identidad, lugar de nacimiento, estado civil, edad, lugar de residencia, trayectoria académica, laboral o profesional. Existe también información sensible como su estado de salud, sus características físicas, ideología política, entre otros aspectos (Superintendencia de Industria y Comercio. Circular externa #001, 2016).

Es importante mencionar que prevalecerá el criterio de participación voluntaria e informada, respeto a la dignidad, protección de derechos y bienestar de las personas participantes. El propósito y las características del estudio serán explicados claramente a los encuestados, solicitando la aceptación del consentimiento en voz alta, para que quede registrado en el audio asociado a cada encuesta, como muestra de su aceptación voluntaria y del entendimiento de autorizar el uso de sus datos con fines estadísticos, garantizando en todo momento su anonimato, es decir la protección de su identidad. También deberá quedar claro que la persona podrá rehusarse a contestar la encuesta o parte de ella, sin perjuicio alguno por esa decisión. El guion de ese consentimiento será redactado de manera convincente para incentivar que la persona seleccionada acepte participar en el estudio. Una vez se cuente con la aceptación del informante seleccionado, se podrá iniciar con la encuesta.

1.4.7. Seguimiento y control.

Para garantizar las tres características que hacen que la información sea confiable (comparabilidad, integridad y veracidad), SEI organizará procesos de supervisión y depuración de los datos para entregar información de calidad. La función de control del trabajo es el objetivo básico de la supervisión y se concreta en la coordinación del personal a su cargo para cumplir las metas de calidad y tiempo de ejecución del proyecto, y en la supervisión directa de su trabajo.

Para conseguir la comparabilidad, integridad y veracidad de la información, los supervisores desarrollarán diferentes técnicas de supervisión, las cuales se describen en los siguientes párrafos. Cabe anotar que las técnicas de supervisión son registradas directamente en un aplicativo diseñado para tal fin, y sus resultados se monitorean a través del dashboard interno para el seguimiento de la coordinación.

1. **Verificación de la información**. Como no es práctico verificar en cada unidad de observación la información de las encuestas, se recurrirá a una muestra en la cual se verificarán presencialmente algunas variables clave de los hogares seleccionados. Esta verificación se hará sobre la información obtenida por el encuestador. Si existen grandes diferencias, el supervisor debe aclarar con el informante la causa de esto y tomar una decisión acerca de la permanencia del encuestador en el proyecto. Esta técnica de supervisión se aplicará para algunas unidades de observación y su selección se hará de manera discrecional por parte del supervisor, basado en criterios como rendimiento del encuestador, oportunidad y optimización de recursos. El supervisor puede, en todo caso, optar por verificar algunas encuestas donde encuentre inconsistencias o que le dejen dudas sobre el trabajo del encuestador. Asimismo, puede decidirse por verificar una muestra mayor de algunos encuestadores, por ejemplo, de aquellos que presentan mayores deficiencias, o son inexpertos o tienen un rendimiento muy elevado o presentan mayores diferencias.

   En general, es recomendable cambiar las preguntas que se verifican en distintas unidades de observación visitadas por un mismo encuestador, y no debe permitirse a los encuestadores el acceso a la información verificada o que conozcan las preguntas que usualmente verifica el supervisor. Deben elegirse preguntas que permitan verificar la consistencia de varias preguntas simultáneamente o que correspondan a información vital para el estudio y cuya veracidad deba garantizarse.

   Los resultados de la verificación requieren un análisis juicioso, específicamente cuando las respuestas no coinciden, ya que el dilema es aceptar la encuesta o rechazarla. La primera situación dependerá de las explicaciones del encuestador y de su veracidad, y el supervisor debe comprobarlas con el informante o con testigos presenciales. Si se decide rechazar el formulario, habrá que verificar las otras

encuestas realizadas por el encuestador y, además, prescindir de sus servicios si la situación se presenta en otros casos.

2. **Acompañamiento.** Otra de las maneras de hacer supervisión y controlar la calidad, consiste en acompañar al encuestador durante algunas aplicaciones, sin interferir con las normas sobre confidencialidad y anonimato de la encuesta. El acompañamiento sustituye a la verificación en cuanto se hace directamente, lo cual le permite asegurarse de que la información es verídica. La función del supervisor durante el acompañamiento es la participación como "observador activo" de la aplicación, lo cual le da ocasión para observar la técnica del encuestador y su habilidad; y por lo tanto permite corregir sus deficiencias.

3. **Asesoría al personal operativo.** Parte de la función de supervisión consiste en la asesoría permanente que se debe dar al personal en todo lo relativo a su trabajo. Esta actividad debe orientarse siempre a mejorar la calidad y el rendimiento en el trabajo y apoyándose en el instructivo de recolección y en las circulares que se emitan con posterioridad. Durante todo el tiempo, el supervisor apoyará a los encuestadores en este aspecto y completará el entrenamiento a través de instrucciones adicionales. Tal actividad es especialmente importante al comienzo del trabajo.

Adicionalmente a los controles de calidad descritos, desde el nivel central se contará con el apoyo de un equipo de escucha de audios, quienes, como su nombre lo indica, serán responsables de escuchar algunos audios de encuesta para verificar técnicas de entrevista, lectura del consentimiento informado, sondeo, y consistencia de preguntas, entre otras. Con base en las revisiones, cada dos días estarán enviando reportes al coordinador de trabajo de campo para informar del avance de su gestión.

Con todos estos controles se generarán indicadores por grupo y por encuestador, y cada supervisor será responsable de retroalimentar a su grupo con el fin de mejorar la aplicación y garantizar mejores estándares de desempeño. Se tiene previsto realizar técnicas de supervisión entre acompañamientos, verificación y escucha de audios, a un 20% de los registros en la muestra, es decir que, alrededor de 2.680 encuestas contarán con algún tipo de verificación de las ya mencionadas.

A medida que el trabajo de campo avance y con la información disponible, se generarán tablas de frecuencia de las respuestas de la encuesta para identificar posibles errores que no hayan sido detectados por la malla de validación. Para tener el reporte completo no solo del avance en campo y de las eventuales situaciones que se puedan presentar, es fundamental que el comité técnico se reúna periódicamente para tener conocimiento del avance y poder así dar soporte y soluciones oportunas.

De otra parte, para controlar el avance del trabajo de campo, la cobertura y la calidad de la información en cuanto a variables críticas, se generarán una serie de reportes periódicos sistematizados que serán usados por los supervisores de campo y la coordinación general a nivel central que servirán como insumos para presentar resultados al CTS.

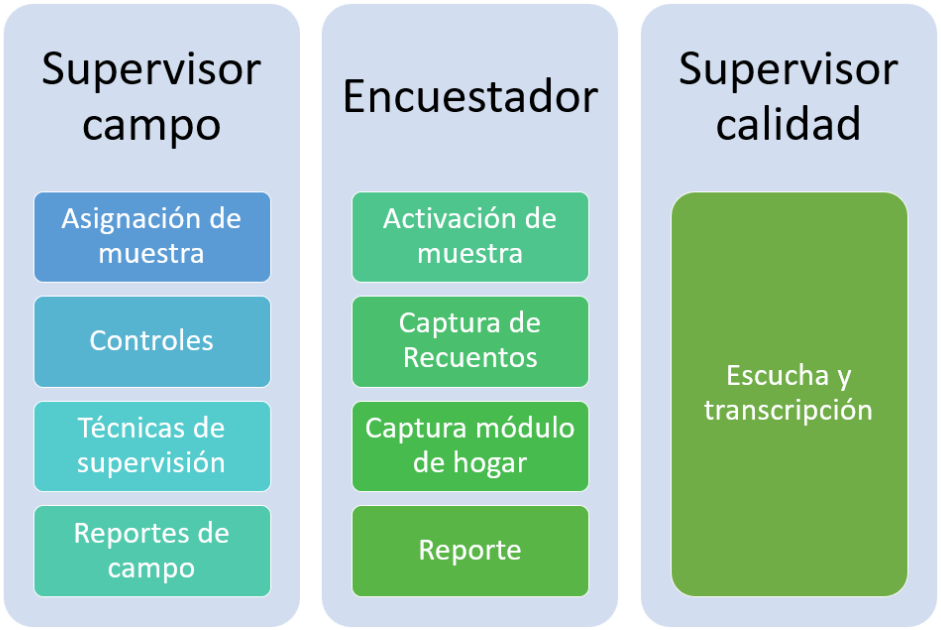
1.4.8. Sistemas de captura y georreferenciación de los datos recolectados.

La información será obtenida utilizando el programa CSPro, la cual es una poderosa herramienta de software ampliamente utilizada en el ámbito de la estadística y la investigación social para la captura y gestión de datos. SEI diseña y configura un aplicativo de forma personalizada para este levantamiento permitiendo configurar controles y validaciones para una mejor calidad de los datos. En la Figura 10 se presenta un esquema general del proceso de captura de la información, el cual tiene entre otras, las siguientes funcionalidades:

- Personalización de formularios: Según sus necesidades específicas, lo que les permite recopilar información relevante de manera eficiente además permite definir reglas de validación y lógica para garantizar la integridad de los datos capturados.
- Validación en tiempo real: El aplicativo realiza validaciones en tiempo real para garantizar la precisión de los datos. Los errores se pueden identificar y corregir de inmediato, lo que reduce la probabilidad de errores en los datos finales.
- Exportación de datos: Una vez que los datos se han capturado y validado, se permite la exportación de datos en varios formatos, como CSV, Excel o bases de datos SQL, para su análisis posterior.
- Seguridad de los datos: se proporciona funciones de seguridad para proteger los datos sensibles. Los formularios y datos están encriptados y protegidos con contraseñas.
- Soporte para dispositivos móviles: el aplicativo es compatible con dispositivos móviles, lo que facilita la captura de datos en smartphones, el uso de smartphones representa múltiples ventajas frente a otros tipos de dispositivos, ya que se trata de elementos pequeños, lo cual facilita su porte de manera segura, y cuentan con óptimo rendimiento y procesamiento. Además, los smartphones permiten almacenar grandes cantidades de información y procesarla de forma eficaz, garantizando conectividad con redes que permiten acceder a internet por varios medios (WIFI-Ethernet-Móvil), lo que facilita la transmisión de datos sin importar su ubicación geográfica.

En resumen, el sistema de captura en CSPro es una herramienta que permite a los investigadores y encuestadores recopilar, validar y gestionar datos de manera eficiente y precisa, contribuyendo así a la calidad y la confiabilidad de los datos recopilados.

Figura 10. Estructura general del sistema de captura de la información.



Fuente: SEI

Georreferenciación

En el proceso de recopilación de datos, los encuestadores emplean una técnica conocida como georreferenciación, la cual les permite obtener una ubicación precisa en el campo, garantizando la calidad y precisión de los datos recopilados. Para llevar a cabo este proceso, utilizan mapas tanto impresos como digitales como referencia.

La georreferenciación se apoya en combinación con sistemas de captura, que juegan un papel esencial en la obtención de la ubicación exacta. Los encuestadores recopilan puntos de GPS durante sus actividades de campo, marcando con precisión las coordenadas.

Un aspecto crucial de este proceso es la comparación de estos puntos de GPS con los puntos de referencia ubicados en el centroide de la manzana o la vereda correspondiente. Este enfoque garantiza que la información geográfica recopilada esté relacionada de manera precisa con las áreas geográficas específicas que se están estudiando.

Es importante destacar que los datos relacionados con la georreferenciación se manejan con un alto nivel de seguridad y privacidad. Estos datos se almacenan de manera segura y están encriptados para proteger su confidencialidad y evitar cualquier acceso no autorizado.

1.4.9. Transmisión de datos.

La transmisión de datos se lleva a cabo de manera constante, las 24 horas al día y los 7 días a la semana, siempre que haya disponibilidad de conexión a Internet por parte de los encuestadores en las diversas zonas donde se encuentren desplegados. Este proceso es esencial para mantener la información actualizada y disponible en tiempo real. Para garantizar la seguridad y confiabilidad de esta transmisión de datos, se implementan rigurosas medidas de protección. La información se encuentra resguardada mediante un sólido sistema de encriptación, lo que significa que los datos se convierten en un formato ilegible para cualquier persona no autorizada. Además, se establecen conexiones seguras y directas con servidores de confianza, lo que minimiza los riesgos de intercepción o manipulación de datos durante el proceso de transmisión.

Esta infraestructura de transmisión de datos proporciona una base sólida para la gestión y el resguardo de información sensible, lo que es esencial en proyectos de encuestas y recopilación de datos.

Al mantener la información encriptada y utilizar conexiones seguras y confiables, se garantiza la integridad y la confidencialidad de los datos recopilados, lo que a su vez contribuye a la credibilidad y la precisión de los resultados finales.

1.5. Diseño del procesamiento de los datos

1.5.1. Consolidación de archivos de datos.

Esta consolidación se realiza de forma automática en distintos momentos del día para a medida que se recopilan los datos en campo, las encuestas completas se sincronizan en la nube, lo que permite una gestión eficaz y oportuna de la información recopilada. Esta sincronización agiliza significativamente el flujo de datos y proporciona a los responsables de la consolidación acceso inmediato a la información, facilitando así la toma de decisiones basada en datos actualizados y precisos.

La depuración de datos se desarrollará en varios momentos: el primero es paralelo con la recolección de información, gracias a la incorporación de la malla de validación en el aplicativo de captura, de forma que el programa sólo acepte valores válidos y haga cruces entre preguntas; el segundo momento comienza una vez se consolida la información, luego de lo cual se podrán revisar las frecuencias simples e incorporar nuevas validaciones que hayan surgido durante la recolección, o validaciones que sirvan como control y verificación de la consistencia de la información.

1.5.2. Codificación.

El proceso de codificación es un elemento fundamental en la gestión de datos, y su implementación se produce en diferentes momentos. En primer lugar, comienza durante el diseño mismo de la encuesta, donde se establece un sistema de codificación uniforme para todas las preguntas del cuestionario. Esto implica asignar códigos numéricos o alfanuméricos a cada posible respuesta o categoría de respuesta. Además, el uso del sistema de captura de datos ayuda a garantizar que esta codificación se aplique de manera consistente, evitando omisiones o la falta de información, así como asegurando que las respuestas se alineen con las categorías previamente definidas.

En segundo lugar, el proceso de codificación se aplica a las preguntas abiertas, como aquellas que requieren respuestas en formato de texto libre. En este caso, una persona especializada revisa y recategoriza las respuestas, convirtiéndolas en opciones estandarizadas o agrupándolas en categorías relevantes.

1.5.3. Diccionario de datos.

Los diccionarios de datos proporcionan una descripción detallada de los elementos de datos utilizados en el instrumento del levantamiento de la información.

- Recuento de viviendas
- Recuento de hogares
- Modulo del hogar y persona seleccionada

A continuación se detallan algunos de los componentes de los diccionarios de datos:

Nombre del elemento de datos (Variables): Este es el nombre único que se le da a un elemento de datos específico.

Descripción o Label: Proporciona una explicación clara y concisa del elemento de datos. Esto ayuda a los usuarios a comprender su significado y su uso.

Tipo de datos: Indica qué tipo de datos se almacenan en el elemento, como texto, número, fecha, etc.

Longitud o tamaño: En el caso de datos de texto o cadenas, se especifica la cantidad máxima de caracteres que puede contener el elemento. En el caso de números, puede indicar la cantidad de dígitos o decimales permitidos.

Restricciones: Pueden incluir reglas o restricciones sobre cómo se deben ingresar o validar los datos en el elemento. Por ejemplo, un formato de fecha específico o un rango de valores permitidos.

Clave primaria o foránea: Indica si el elemento es parte de una clave primaria (utilizada para identificar de forma única un registro) o una clave foránea (que hace referencia a otro registro en una tabla relacionada).

Un diccionario de datos es una herramienta esencial para garantizar la consistencia y la comprensión de los datos, lo que facilita la comunicación entre el equipo del proyecto.

1.5.4. Revisión y validación de los archivos de datos.

Este proceso implica una evaluación exhaustiva de los datos recopilados, comenzando por la verificación de la consistencia y la integridad de la información. Se debe comprobar que no haya datos faltantes, duplicados o erróneos que puedan sesgar los resultados. Además, de realizar una revisión de los códigos y etiquetas utilizados para asegurar que sean coherentes y representen adecuadamente las variables. Nuestra validación de datos comienza desde la misma captura en campo de estos al tener una malla de validación que permite validar el dato en sitio.

En última instancia, la validación de los archivos de datos asegura que los investigadores puedan confiar en la solidez de sus hallazgos y tomar decisiones basadas en datos confiables.

1.5.5. Instrumentos de edición (validación y consistencia) e imputación de datos.

Remitirse a la sección 1.3.17. numeral iii. análisis de la no respuesta e imputación de datos, donde se describe en detalle el manejo de la no respuesta y el procedimiento para la imputación de datos.

1.5.6. Cuadros de salida.

Remitirse a sección 1.3.13. metodología de la estimación, en la que se describen características de los cuadros de salida y otras formas propuestas para la visualización de los datos recolectados y análisis.

1.5.7. Definición de variables y construcción de indicadores.

Las variables contenidas en el instrumento de encuesta fueron construidas por los expertos del DNP, quienes determinaron su relevancia para la medición de la percepción de la ciudadanía, acerca del contenido del Plan Nacional de Desarrollo. En el Anexo 1, se presenta la forma en que se construyeron los principales indicadores del instrumento, a modo ilustrativo.

2. PRIMERA ENTREGA DOCUMENTACIÓN DE BASES DE DATOS DE ENCUESTA

En el Anexo 7 se presenta la primera entrega del formato Nesstar con la estructura para metadatos y microdatos.

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.

Lavallee, 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.

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 Poststratication: State-Level Estimates from National Polls.

4. ANEXOS

ANEXO 1. RELACIÓN DE INDICADORES PRINCIPALES.pdf

ANEXO 2. FORMULARIO.pdf

ANEXO 3. ANEXO ESTADÍSTICO.zip

ANEXO 4. CRONOGRAMA DE RUTAS.pdf

ANEXO 5. CRONOGRAMA DEL PROYECTO.pdf

ANEXO 6. MANUAL ENCUESTADOR_ARD649.pdf

ANEXO 7. PND_2022-2026.Nesstar

ANEXO 8. GLOSARIO TERMINOLOGÍA EPC.pdf