



Nuevas tecnologías en encuestas de hogares



Fernando Olguín Ibarra
Director de Encuestas Regulares en Hogares

20 noviembre 2024

Innovaciones en la Recolección de Datos



Captación de información CAPI

La captación de encuestas en hogares que realiza el INEGI se centra en la recolección de información a través de **entrevistas tipo CAPI (Dispositivos móviles)**. Esto agiliza los procesos, **reduce errores** y permite georreferenciar las actividades que realizan tanto entrevistadores como supervisores.

Sin embargo, hay encuestas que por sus propias dinámicas y complejidades se siguen captando a través de papel, o bien, algún otro método de captación.

Por ejemplo, la Encuesta Nacional de Ocupación y Empleo (ENOE), que se levanta de manera continua a lo largo del año, actualmente se encuentra en una **transición para ser captada en dispositivos electrónicos**, su complejidad radica en la cantidad de validaciones y verificaciones que tiene el proyecto y que tienen que ser reflejadas en el dispositivo. La transición concluye este mes.

Captación de información en CATI

Las entrevistas tipo CATI (Entrevistas telefónicas), fueron un parteaguas para la generación de estadísticas en periodos adversos, como lo fue el periodo de pandemia. Si bien el INEGI no estaba preparado para llevarlas a cabo en ese momento, se construyó una estrategia para su implementación regular.

La **ENOE** tiene incorporada como parte de su metodología la **captación de información a través de entrevistas telefónicas**, lo cual permite contar con **información en periodos adversos** en los que no es posible estar en campo, por ejemplo, por pandemias, por zonas de inseguridad o por informantes que son complicados de encontrar en sus viviendas.

La complejidad para su implementación es que, si bien, no existen diferencias entre la información captada de manera telefónica vs lo cara a cara, sí existe una menor participación de las personas en el proyecto, lo que **disminuye la calidad estadística de la información**.

Captación de información en CAWI

En lo que el INEGI ha centrado esfuerzos para la captación de encuestas en Web (CAWI), ha sido en **temas emergentes a grupos específicos**, en los que se busca indagar por características que no se preguntan en cuestionarios tradicionales.

De igual manera, son encuestas que, si bien van enfocadas a una población en muchos casos reducida, se busca que puedan **participar la mayor cantidad de personas**, en términos de comunicación **favorece a la confianza** en la encuesta.

Los resultados de las entrevistas realizadas a partir de este tipo de instrumentos **de ninguna manera pueden ser generalizables a la población objetivo**. Se trata de un estudio de tipo cualitativo que permite complementar la información que se obtiene de una encuesta probabilística en la temática.

El INEGI llevó a cabo la Encuesta Nacional sobre Diversidad Sexual y de Género (ENDISEG) tanto en una muestra probabilística como en CAWI.

Captación de información con otros métodos

Otros métodos de captación que ha implementado el INEGI en encuestas en hogares:

Encuestas en dispositivos electrónicos con audífonos; se han implementado en encuestas con temáticas delicadas, donde el informante pudiera no sentirse cómodo con la lectura de las preguntas y la declaración de las respuestas.

Invitación por cartas; en una muestra aleatoria enviamos cartas por correo postal, invitando a la población a participar en una encuesta, contactándonos de manera telefónica o por mail, donde se le aplicaría la encuesta de empleo.

Entrevistas de población de estratos altos; por medio de entrevistas cara a cara por Zoom, Teams

Desafíos en la recolección de datos

Acceso a Zonas Rurales y Marginales: La infraestructura limitada en algunas zonas dificulta la recolección de datos.

Desconfianza y Seguridad: La desconfianza en las instituciones y la inseguridad ciudadana afectan la participación.

Brecha Digital: El acceso desigual a internet limita el uso de encuestas web en algunos sectores de la población.

Participación ciudadana: Es importante que la población esté informada sobre los esquemas de captación que se implementan, generando confianza y seguridad para los informantes.

Encuestas telefónicas



Evaluación de entrevistas telefónicas

La ENOE es una encuesta que se levanta de manera trimestral, cada vivienda participa en cinco ocasiones a lo largo de un año tres meses. Actualmente la **estrategia de captación** para el 90% de las viviendas es con **captación en papel/CAPI**, mientras que el 10% restante se **levanta por CATI**, con la excepción que la primera visita debe de ser cara a cara (para generar confianza), mientras que el resto de las visitas son con seguimiento telefónico.

Antes de su implementación se diseñaron algunas estrategias para evaluar cuál sería la mejor en su implementación:

- Entrevistas telefónicas con visita previa
- Entrevistas telefónicas sin visita previa
- Entrevistas a número aleatorios

Evaluación de entrevistas telefónicas

Las **pruebas estadísticas** indicaron que **no existen diferencias estadísticamente significativas** entre los tres métodos de captación en el caso de los indicadores del mercado laboral, sin embargo, **sí se encontró un efecto en la tasa de respuesta de los informantes**, lo cual impacta en la calidad estadística de la información.

El **marco provee los números telefónicos** de las dos primeras estrategias donde se observan las siguientes tasas de respuesta

- Entrevistas telefónicas con visita previa (60% de respuesta)
- Entrevistas telefónicas sin visita previa (40% de respuesta)

En el caso de los **números aleatorios** se toma como base el padrón público a seis dígitos de los números telefónicos.

- Entrevistas a número aleatorios con padrón de 6 dígitos (5% de respuesta).

Innovaciones en el Análisis de Datos



Anonimización de Microdatos



Anonimización de Microdatos

Los microdatos que se publican tradicionalmente en las encuestas en hogares deben de guardar **principios de confidencialidad**, un principio importante es el de **no publicar las clasificaciones geográficas a un nivel subestatal** (municipio, localidad, etc).

Sin embargo, también es importante conocer si **a partir de la información publicada** es posible hacer una **identificación directa** de los informantes **entre proyectos** estadísticos.

Por lo que se evaluaron tres programas; el Censo de población y Vivienda 2020, la Encuesta Nacional de Ocupación y Empleo y la Encuesta Nacional de Ingresos y Gastos de los hogares, con la finalidad de **analizar el riesgo de que se materialice una re identificación de los Informantes**.

Anonimización de Microdatos

A partir de la **información vinculante** entre los proyectos, como las características sociodemográficas, socioeconómicas, las características de las viviendas, entre otras, **se realizó un modelo para evaluar su identificación.**

Método Generative Adversarial Networks (GAN): La idea es crear **dos modelos de redes neuronales**. Uno llamado generador el cual recibe **valores aleatorios y los transforma en un registro**. El otro modelo es llamado el discriminador, **recibe un registro y trata de discernir si es real o sintético.**

Estos dos modelos **son entrenados de manera iterativa** con el objetivo de vencer a su contraparte intentando lograr un equilibrio entre ambos. El generador es entrenado para poder crear registros que parezcan reales, mientras el discriminador es entrenado para encontrar las sutiles diferencias entre los datos reales y generados.

Se determinó que **no existen probabilidades significativas** que permitan concluir que la información es vinculante entre proyectos estadísticos.

Vinculación de cifras entre Encuestas y Registros Administrativos



Vinculación

Es importante que las **encuestas en hogares** puedan **confirmar que las tendencias y las magnitudes** que se presentan en los resultados reflejan una realidad de país, **más allá de la calidad estadística que pudieran tener.**

La validación de información con el Censo de Población y Vivienda es una actividad habitual, sin embargo, el CPV tiene limitantes que no permiten contar con muchas características de la población.

Utilizar registros administrativos permite contar con valores especializados de la población y es una labor del instituto para que cada vez más unidades de estados participen en la generación y producción de información estadística.

Vinculación

Actualmente en el país los **registros administrativos no son lo suficientemente especializados** para generar una gran cantidad de indicadores clave. Los más importantes son:

- El número de trabajadores registrados en el Instituto Mexicano de Seguridad Social (IMSS).
- El ingreso promedio de los trabajadores registrados en el IMSS.
- Remesas
- Estadísticas vitales (nacimientos, defunciones, etc.)

Por ejemplo, los últimos cambios que ha sufrido el país en cuanto al **incrementos al salario mínimo**, así como una disminución en la pobreza, son aspectos que han detectado tanto las encuestas como los registros administrativos en magnitudes similares, lo que abona a la calidad de la información.

Uso de Machine Learning en encuestas en hogares



Codificación de la información

La Encuesta Nacional de Ocupación y Empleo (ENOE) recaba información sobre las características laborales de la población, en particular se **pide describir las actividades que realizan en su trabajo, así como las actividades que realiza el negocio o empresa en la que trabajan.**

Dichas **variables se codifican en una variable numérica** en dos etapas:

- La primera un **sistema algorítmico** busca las palabras claves y les asocia un valor en el clasificador.
- En el caso de que no se logran clasificar a partir del algoritmo anterior, **una persona hace la clasificación manual** de su información.

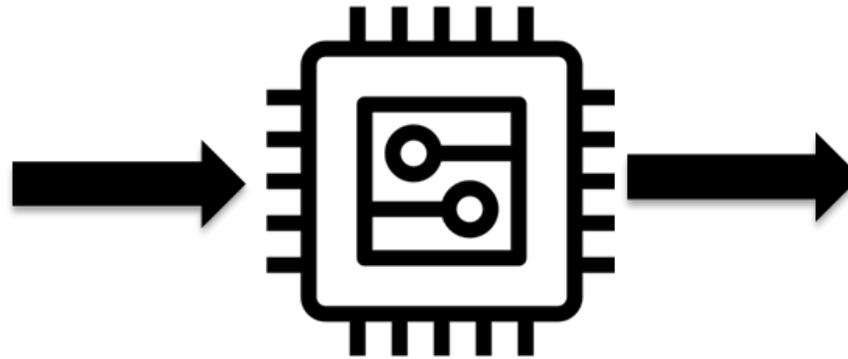
Sin embargo, es importante tener en cuenta que **la clasificación inicial puede fallar para ciertas palabras**, además de que, en la segunda etapa, **puede haber errores humanos** que interfieran en la correcta clasificación.

Codificación de la información

**Descripción de la
unidad económica
donde labora**

**Codificación
automática o asistida
con base en la
clasificación oficial**

“PREPARACIÓN DE TACOS
EN LA VÍA PÚBLICA”



CÓDIGO: 7221
Servicios de preparación
de alimentos y bebidas

Clasificación por Machine Learning

El proyecto de codificación involucra incorporar **algoritmos de Inteligencia Artificial y Procesamiento de Lenguaje Natural** para complementar los procesos de codificación para las variables de Ocupación y Actividad económica de la ENOE.

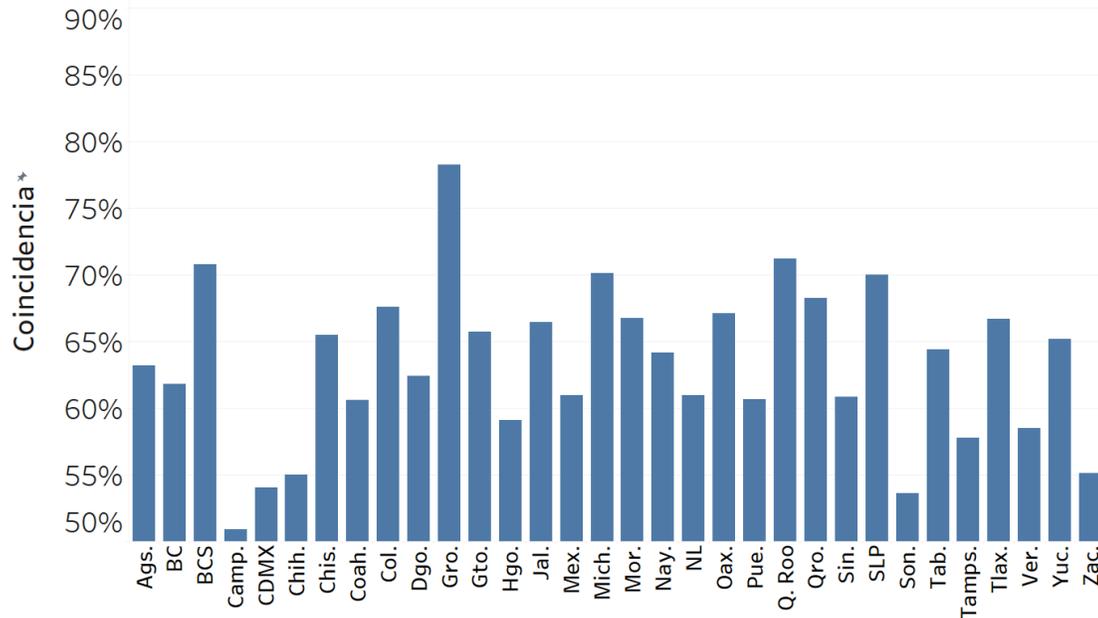
Las variables que se toman en el modelo son **las descripciones historias** con la que cuenta el proyecto, así como **sus respectivas clasificaciones**.

La finalidad es **disminuir** la proporción de registros codificados de **forma asistida**, con un **ahorro significativo en el tiempo** que tomaría esta actividad, así como la **disminución de errores** y disminuir el factor humano de la codificación manual.

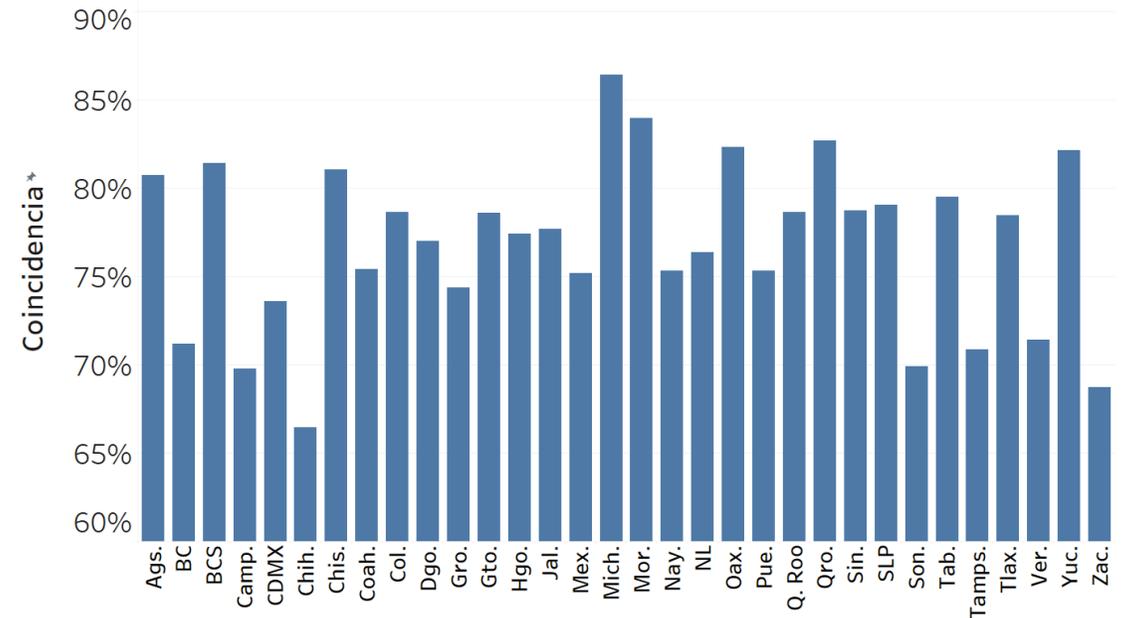
Clasificación por Machine Learning

Sin embargo, la codificación manual por entidad federativa presenta áreas de oportunidad, ya que la codificación manual en algunas entidades cobra una relevancia importante por la dinámica laboral que se pueda presentar en cada zona y que es distinta a las demás.

Ocupación



Actividad económica



Clasificación por Machine Learning

Los métodos de Machine Learning harán más **eficientes los procesos de codificación** de información en las encuestas en hogares, sin embargo, es necesario que los resultados **no reflejen un cambio de tendencia o de magnitud en sus resultados anteriores**.

Se continuarán **afinando los algoritmos de lenguaje**, así como de inteligencia artificial, de tal manera que puedan **reflejar una mejor coincidencia para cada entidad federativa**.

Se llevarán a cabo **pruebas de hipótesis** para determinar los cambios que pueden existir entre la clasificación anterior y la actual.

Otros análisis que se realizan en el INEGI



Análisis recientes

Estimaciones del ingreso y de condición de ocupación en áreas pequeñas, tomando como base las encuestas en hogares.

- Estimaciones a nivel municipal a través de métodos de áreas pequeñas

Nowcasting del ingreso corriente corriente en la Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH).

- Técnica que utiliza datos indicadores económicos clave para estimar el ingreso de los hogares.
- El modelo logra estimar el ingreso corriente promedio por decil dentro de los intervalos de confianza para la ENIGH 2022. Generando estimaciones oportunas confiables
- En marzo de 2025 se esperan contar con los resultados oportunos para 2024, donde se podrá evaluar de nueva cuenta el proyecto.

Análisis recientes

Nowcasting de la distribución del ingreso. Microsimulaciones

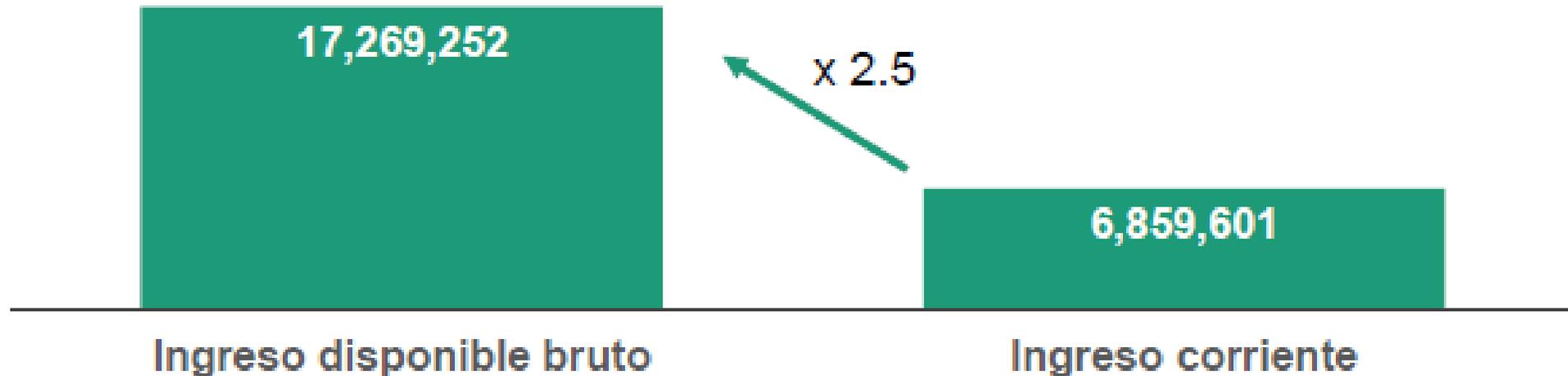
- Las microsimulaciones se basan en actualizar o ajustar datos que reflejen los efectos de cambios socioeconómicos, por ejemplo, aspectos demográficos, o bien, registros administrativos como el ingreso, remesas, impuestos, transferencias o beneficios sobre unidades económicas a nivel micro, como hogares e individuos.
- Los modelos de microsimulación pueden generar estimaciones detalladas de la distribución del ingreso a nivel individual y familiar.
- Se ajustan los factores de expansión y el ingreso conforme las variables demográficas y económicas que se toman en cuenta.
- Los modelos de microsimulación son una muy útil herramienta para obtener indicadores para medir la distribución del ingreso no más oportunos, pero si más frecuentes, por lo menos anualmente

Análisis recientes

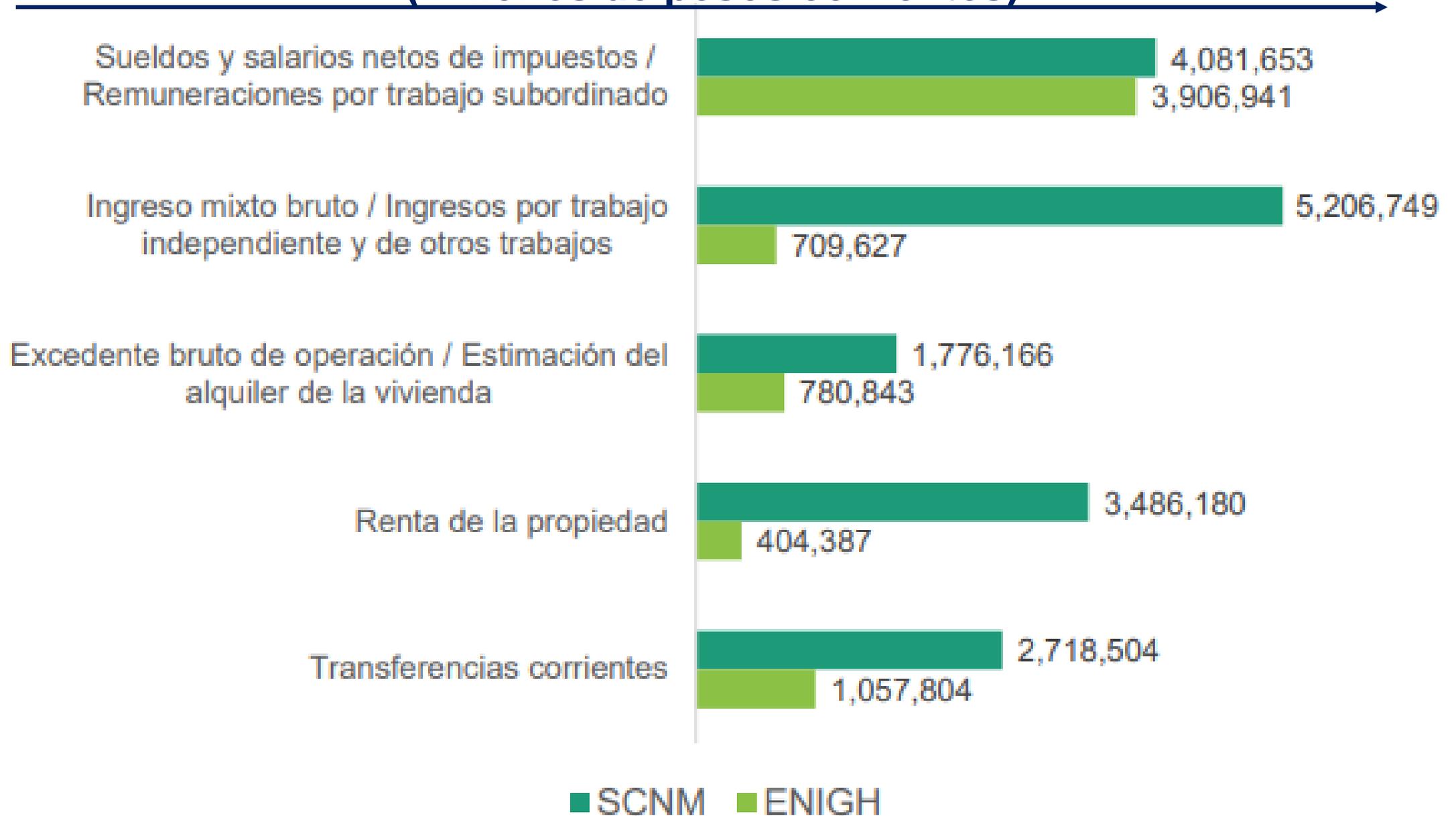
Ingresos ajustados a cuentas nacionales.

- Los datos de las encuestas no son directamente compatibles con los agregados macroeconómicos que componen el Sistema de Cuentas Nacionales de México (SCNM).

Ingreso anual de los hogares en el 2018
Millones de pesos



Ingresos anuales de los hogares en la ENIGH y el SCNM, 2018 (Millones de pesos corrientes)



Análisis recientes

- Las encuestas subestiman el ingreso de los hogares, especialmente en los deciles más altos.
- Se propone una técnica de ajuste que considera elimina las diferencias micro y macro, tomando en cuenta los ingresos y los gastos de forma simultánea, así como el método de microsimulaciones.
- La técnica empleada logró útil para generar resultados consistentes con el SCNM.
- Los resultados de este proyecto permitirán generar Matrices de Contabilidad Social con gran flexibilidad en la clasificación de los hogares.

Desafíos en el análisis de datos



Desafíos

Capacitación: Se requiere capacitación en técnicas avanzadas de análisis para el personal estadístico.

Confidencialidad: La protección de la confidencialidad de los datos es crucial al usar nuevas fuentes y métodos.

Calidad de los Datos: La integración de distintas fuentes requiere rigurosos controles de calidad.

Estandarización: Se requieren metodologías y documentación técnica que pueda ser analizada y replicable por instituciones nacionales o internacionales.

Más información:
<https://www.inegi.org.mx/>

GRACIAS



Conociendo
México

800 111 46 34
www.inegi.org.mx
atencion.usuarios@inegi.org.mx

    **INEGI** Informa