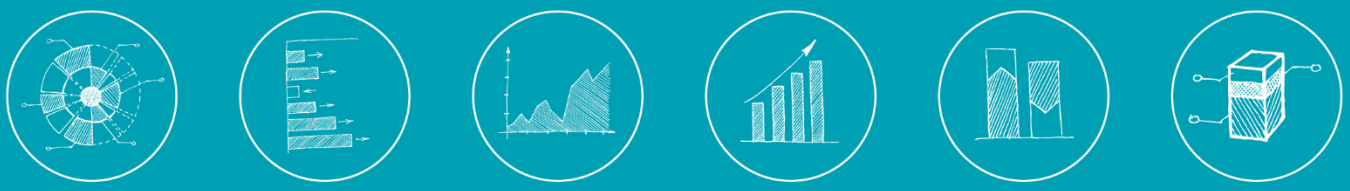
Cuaderno de Trabajo
No. 21 – Junio 2025

Reconstrucción de Redes Familiares: Desafíos en la Identificación de Efectos de Red mediante Proxies Indirectos

2025



Elaboración técnica:

Buenas cifras,
mejores vidas

Autores:

Felipe Brugués.
The Business School at Instituto Tecnológico Autónomo de México, Ciudad de México

Juan Felipe Riaño
Department of Economics, Georgetown University, Washington D.C.

Cristhian Rosales Castillo
Department of Economics, Emory University, Atlanta. GA

Andrés Villacís M
Instituto Nacional de Estadística y Censos, Ecuador

Reconstrucción de Redes Familiares: Desafíos en la Identificación de Efectos de Red mediante Proxies Indirectos

Felipe Brugués[1] Juan Felipe Riaño[2] Cristhian Rosales-Castillo[3]
Andrés Villacís-Miranda[4]

Abstract

Los estudios empíricos sobre redes familiares dependen abrumadoramente de *proxies* indirectos, como los apellidos compartidos, cuya precisión rara vez ha sido contrastada con datos de referencia. Este trabajo presenta una primera validación a gran escala de estos métodos, construyendo la red familiar casi completa de un país entero, Ecuador, a partir de datos administrativos sobre vínculos de parentesco explícitos. Comparamos esta red de referencia con redes contrafactuales construidas utilizando los ampliamente adoptados *proxies* de apellidos, y analizamos las discrepancias en propiedades topológicas generales. Encontramos que, si bien los *proxies* de apellidos pueden replicar parcialmente la distribución del tamaño de las familias, fallan sistemáticamente en capturar la verdadera estructura de la red. Los *proxies* tergiversan el número de unidades familiares distintas (componentes conectados) y subestiman o sobreestiman el tamaño de las mismas. Este hallazgo sugiere que los estudios basados en *proxies* pueden estar afectados por un error de medición no clásico y significativo, lo que podría sesgar las estimaciones de efectos de red en diferentes direcciones. Este documento presenta un primer resultado en la reconstrucción de redes familiares que están siendo usadas en investigaciones más precisas sobre los efectos causales de las redes familiares en los ámbitos económico, social y político.

1. Introducción

Las conexiones familiares influyen en una amplia variedad de resultados en las economías en desarrollo. Estas redes inciden en quién accede a empleos, gana contratos o dirige empresas, a menudo a través de mecanismos informales, difíciles de medir o completamente invisibles para los analistas. La ausencia de datos administrativos sobre vínculos de parentesco ha llevado a los investigadores a recurrir a proxies indirectos—como apellidos compartidos, etnicidad, identificadores fiscales, redes familiares incompletas o lugar de residencia—para reconstruir los lazos familiares (Cruz, Labonne, y Querubín 2017; Gagliarducci y Manacorda 2016; Banerjee et al. 2013). Si bien estos proxies son prácticos, pueden no reflejar con precisión la estructura social subyacente ni los caminos de interdependencia entre unidades, lo que plantea importantes desafíos para la identificación de efectos de red.

[1] The Business School at Instituto Tecnológico Autónomo de México, Río Hondo 1, Altavista, Álvaro Obregón, 01080 Ciudad de México, CDMX, Mexico felipe.brugues@itam.mx.
[2] Department of Economics, Georgetown University, 37th St NW & O St NW, Washington, DC, 20057 juan.felipe.riano@georgetown.edu.
[3] Department of Economics, Emory University, Rich Building, 1602 Fishburne Dr., Atlanta, GA 30322-2240, USA crosal3@emory.edu.
[4] Instituto Nacional de Estadística y Censos del Ecuador, Quito, Ecuador andres_villacis@inec.gob.ec.

En este documento presentamos la reconstrucción de redes familiares a partir del registro único de personas en Ecuador, conocido como el registro de cedulados. Este registro contiene información histórica sobre relaciones familiares—incluyendo vínculos de parentesco con padres, madres y cónyuges—lo que permite reconstruir, con un alto grado de precisión, la estructura familiar de prácticamente toda la población registrada.

Como ejercicio complementario, y con el objetivo de contrastar la validez de este enfoque con métodos indirectos comúnmente utilizados en la literatura, construimos redes familiares alternativas utilizando apellidos paternos y maternos como proxies de parentesco. Esto nos permite realizar una primera comparación sistemática entre la topología de ambas redes, evaluando características como el número de familias identificadas y el tamaño de las mismas, medido en términos del número de individuos asociados. Este estudio constituye un primer esfuerzo en la construcción y validación de bases de datos sobre redes familiares a gran escala, que son el insumo fundamental de futuras investigaciones sobre el papel de las redes familiares en diversos ámbitos económicos y sociales.

Para ilustrar el alcance de este problema, consideremos el uso de proxies en distintos campos. En la investigación política, por ejemplo, Querubín (2016) y Cruz, Labonne y Querubín (2017) utilizan apellidos para rastrear redes familiares de políticos en Filipinas y documentar su influencia sobre los resultados electorales. Querubín (2016) encuentra que los candidatos que apenas ganan su primera elección tienen cinco veces más probabilidades de tener un familiar en el poder en el futuro que aquellos que apenas pierden. De forma similar, Cruz, Labonne y Querubín (2017) muestra que los candidatos políticos tienden a provenir de familias con mayor centralidad en la red familiar del país, y que esta centralidad contribuye a un mayor porcentaje de votos durante las elecciones.

En economía, Dal Bó, Dal Bó, and Snyder (2009) estudian las dinastías políticas en Estados Unidos mediante el análisis de apellidos a lo largo de generaciones. Los autores muestran que la representación dinástica en el Congreso es más común que en otras profesiones. Utilizando estrategias de variables instrumentales, encuentran que una mayor duración en el cargo incrementa la probabilidad de que familiares del legislador también ocupen cargos en el Congreso. George (2020) , por su parte, examina los efectos del desarrollo asociados a las dinastías políticas en India, utilizando autodeclaraciones de vínculos familiares desde 1947. Encuentra que, si bien ciertas regiones experimentan un crecimiento inicial gracias a los efectos positivos de los fundadores políticos, a largo plazo se empobrecen, pues los efectos negativos de sus descendientes superan las ventajas iniciales.

En el ámbito judicial, Brassiolo et al. (2021) investigan el nepotismo en el sistema judicial mexicano a partir de vínculos familiares inferidos por apellidos compartidos. Documentan que cuando los jueces nombran a familiares en el personal del tribunal, se reduce la productividad, lo que sugiere que estas contrataciones responden a incentivos de extracción de rentas más que a criterios de eficiencia. Además, este tipo de prácticas es más común entre jueces sancionados administrativamente, aquellos que laboran en su estado natal, y quienes ocupan cargos de mayor jerarquía.

En el ámbito académico, Durante, Labartino, and Perotti (2011) documentan la existencia de dinastías académicas en Italia, rastreando la continuidad intergeneracional mediante apellidos. Analizan una reforma de 1998 que descentralizó los procesos de contratación universitaria desde el nivel nacional hacia las universidades locales. Encuentran que la mayor autonomía para las autoridades locales generó un aumento significativo del familismo en regiones con bajo capital cívico, sin observarse efectos similares en áreas con capital cívico más alto.

Este patrón también se extiende al sector privado. Gagliarducci and Manacorda (2016) exploran el nepotismo en empresas italianas midiendo vínculos familiares entre políticos y empleados del sector privado mediante información familiar contenida en identificadores fiscales. Los autores encuentran que los políticos logran transferir rentas significativas a sus familiares en forma de empleos en el sector privado. En Ecuador, Brassiolo, Estrada, and Fajardo (2020) y Brugués, Brugués, and Giambra (2024) estudian los vínculos políticos y su relación con el acceso a empleos públicos y contratos de adquisiciones. Brassiolo, Estrada, and Fajardo (2020) muestra que las personas conectadas con alcaldes apenas electos tienen una mayor probabilidad de obtener empleos municipales en comparación con aquellas vinculadas a candidatos que apenas fueron derrotados. De forma similar, Brugués, Brugués, and Giambra (2024) encuentran que cuando los hermanos de accionistas acceden a altos cargos burocráticos, las empresas vinculadas experimentan un aumento considerable en la probabilidad de ganar contratos públicos, especialmente en aquellos asignados con mayor discrecionalidad. Nuevamente en Filipinas, Fafchamps and Labonne (2017) utilizan un diseño de regresión discontinua para analizar si los familiares de políticos acceden a mejores empleos tras elecciones municipales. Encuentran evidencia de que los parientes de funcionarios en el poder tienden a emplearse en ocupaciones mejor remuneradas.

En todos estos casos, los proxies familiares basados en apellidos y otras convenciones han sido herramientas centrales para medir efectos de red.

A partir de estas metodologías, una nueva generación de estudios ha reemplazado los proxies basados en apellidos por registros administrativos que identifican explícitamente vínculos de parentesco. En el sector público, Riaño (2021) vincula parientes de primer grado en toda la nómina del servicio civil en Colombia, utilizando declaraciones juradas obligatorias. Riaño (2021) encuentra que los funcionarios con conexiones familiares ascienden más rápido y reciben mayores salarios, a pesar de las reglas formales contra el nepotismo. En el sector corporativo, Balán, Dodyk, and Puente (2025) aprovechan las declaraciones de "partes relacionadas" de empresas que cotizan en bolsa en Brasil para identificar vínculos de sangre dentro de los consejos de administración. Demuestran que, incluso después de prohibirse las donaciones corporativas, las empresas controladas por familias continúan canalizando recursos hacia la política mediante contribuciones personales realizadas por familiares. Si bien estos estudios representan avances significativos en la identificación de redes de parentesco, siguen enfrentando limitaciones asociadas a la observabilidad parcial y al muestreo restringido a ciertos nodos de la red, lo que impide reconstruir estructuras familiares completas.

A pesar de estos avances, los enfoques basados en proxies siguen estando limitados por restricciones importantes. Por ejemplo, los apellidos compartidos pueden no reflejar vínculos reales debido a factores omitidos relacionados con convenciones de género, cambios de nombre por matrimonio o patrones culturales en la asignación de apellidos. Incluso los datos administrativos que listan parientes explícitamente—aunque más precisos—suelen capturar solo segmentos parciales de la red familiar, omitiendo relaciones más extendidas o informales. Esta reconstrucción incompleta puede introducir sesgos si los vínculos omitidos están correlacionados tanto con los resultados como con la exposición al tratamiento. Más fundamentalmente, las medidas basadas en proxies pueden generar errores de medición no clásicos que complican la identificación de efectos causales y sesgan las estimaciones de red de forma impredecible (Chandrasekhar y Lewis 2016; Riaño 2021; Banerjee et al. 2013).

Un hallazgo central de este análisis preliminar es que el número de familias —definidas como componentes conectados en la red— es representado con mayor fidelidad al emplear un proxy de dos apellidos. Este resultado se explica, en gran medida, por la capacidad de dicha aproximación para capturar de manera más efectiva a los individuos sin conexiones presentes en la red de referencia (red real).

Al excluir este grupo de individuos aislados, emerge un patrón consistente con las expectativas teóricas: el uso de un solo apellido subestima el número de familias, mientras que el uso de dos apellidos lo sobreestima. En lo que respecta a la distribución del tamaño de las familias, el proxy de dos apellidos se alinea más estrechamente con la distribución empírica para familias de tamaño reducido. No obstante, en términos relativos, las distribuciones generadas por ambos proxies son notablemente similares a la distribución de referencia, como lo corrobora un valor reducido de la Distancia de Movimiento de Tierra (*Earth Mover's Distance*, EMD).

Un resultado de particular importancia es la identificación, en la red de referencia, de un componente conectado de tamaño masivo, que aglutina a aproximadamente 16 millones de individuos. La caracterización y análisis de este macro-componente es el objeto de una investigación en marcha. Dicho trabajo se enfoca en tres objetivos principales: primero, estimar los sesgos que surgen en la estimación de efectos de red debido al uso de proxies; segundo, proponer metodologías para la corrección de dichos sesgos; y tercero, usar la red familiar real en el análisis de efectos de red en diferentes contextos económicos, políticos y sociales.

La estructura del presente documento se articula como sigue: la Sección 2 escribe las fuentes de datos empleadas, detalla el proceso de depuración de estos y expone la metodología para la construcción de las redes analizadas. Subsiguientemente, la Sección 3 presenta los resultados descriptivos fundamentales relativos a la topología de dichas redes, con énfasis en el número y tamaño de las 'familias' identificadas. La Sección 4 ofrece una discusión sobre la aplicación de estas bases de datos en la investigación empírica de efectos de red. Finalmente, la Sección 5 resume las principales conclusiones del estudio y delinea posibles avenidas para futuras investigaciones.

2. Datos y Metodología

2.1. Datos

La presente investigación se fundamenta en el análisis de registros administrativos. Estos se definen como datos generados sistemáticamente por entidades públicas en el ejercicio de sus funciones, susceptibles de ser reutilizados con fines estadísticos. La literatura especializada ha reconocido extensamente el valor de este enfoque para la producción de estadísticas caracterizadas por su oportunidad, granularidad y eficiencia en costos (Wallgren y Wallgren 2007; Statistics-Denmark 1995).

El Instituto Nacional de Estadística y Censos (INEC) de Ecuador ha establecido directrices metodológicas para la adaptación de registros administrativos a su uso estadístico. Dichos procedimientos comprenden la evaluación estructural del registro, la identificación y subsanación de duplicidades, la armonización de variables críticas y la generación de identificadores únicos y consistentes (INEC 2022) . Es pertinente destacar que los registros específicos empleados en este estudio han satisfecho los criterios estipulados por el INEC, asegurando así su idoneidad para la realización de análisis cuantitativos rigurosos.

Para garantizar la confidencialidad de los individuos, los registros son sometidos a un proceso de seudonimización, mediante el cual los identificadores personales directos se sustituyen por códigos alfanuméricos anónimos. Si bien este mecanismo constituye una salvaguarda esencial para la protección de la identidad, introduce limitaciones analíticas, como la posibilidad de errores en la vinculación de datos (*linkage errors*) o una menor trazabilidad entre distintas bases de datos[5].

Los datos de esta investigación provienen en su totalidad de una única fuente administrativa: la Dirección General de Registro Civil, Identificación y Cedulación (DIGERCIC). Dicha institución es la responsable en Ecuador del registro del número único de identidad (cédula) de cada residente[6], así como de sus vínculos de parentesco explícitos (padre, madre, cónyuge) y variables sociodemográficas clave[7].

El análisis emplea dos bases de datos distintas derivadas de esta fuente: «cedulados» y «apellidos». La primera contiene los registros seudonimizados de la población residente, incluyendo los identificadores de sus padres y cónyuge, junto con fechas de nacimiento y defunción. La segunda base consiste únicamente en una cadena de texto con los nombres y apellidos completos de cada individuo[8]. Un aspecto crucial del protocolo de datos es que estas dos bases no pueden ser vinculadas a nivel individual, una medida diseñada para garantizar una estricta confidencialidad.

[5] Para mayor detalle revisar INEC (2022).
[6] Dado el carácter histórico del registro de cedulados, este puede incluir personas que, al momento del análisis, hayan fallecido o emigrado del país.
[7] Para una descripción detallada de la fuente de datos, véase INEC (2024).
[8] El acceso a los registros administrativos «cedulados» y «apellidos» está restringido y solo es posible a través del Laboratorio de Procesamiento de Información del Instituto Nacional de Estadística y Censos (INEC), previa solicitud formal a la institución.

2.2. Metodología

2.2.1. Depuración del registro administrativo de apellidos

El registro administrativo utilizado corresponde a un padrón histórico con corte al año 2022. Este contiene una única variable de texto en la que se registran los nombres y apellidos completos de los individuos en una sola cadena de caracteres. Cada fila representa un individuo único, sumando aproximadamente 21 millones de registros. El propósito de usar esta base de datos es identificar posibles vínculos entre personas a partir de sus apellidos, los cuales pueden reflejar relaciones familiares directas o afinidades nominales que sugieren agrupamientos. Para ello, es necesario descomponer la cadena de texto original y extraer dos variables independientes: apellido1 y apellido2.

En el Ecuador, conforme a la Ley Orgánica de Gestión de la Identidad y Datos Civiles (artículos 36 y 37), se establece que a un individuo se le puede asignar hasta dos nombres simples o uno compuesto, y que los apellidos corresponderán, en general, al primero de cada progenitor. A partir de esta disposición, se espera que la estructura del nombre completo siga el patrón «apellido1 apellido2 nombre1 nombre2». Esta lógica es consistente con los patrones observados en la base de datos analizada en este estudio.

Como se muestra en la Tabla 1, la mayoría de los registros presenta tres espacios entre caracteres, lo que concuerda con una estructura de cuatro componentes sin espacios iniciales ni finales: uno entre apellido1 y apellido2, otro entre apellido2 y nombre1, y el tercero entre nombre1 y nombre2. No obstante, también se identifican registros con menos o más separaciones, lo cual puede responder a nombres compuestos, falta de información, personas con un solo nombre o apellido, nombres extranjeros u otras inconsistencias.

Tabla 1. Distribución de la estructura de nombres y apellidos

Número de espacios por cadena	Número de elementos por cadena	Registros
3	4	19.355.002
2	3	1.076.032
4	5	607.986
5	6	189.787
1	2	71.970
6	7	12.752
7	8	2.670
0	1	510
8	9	122
9	10	15
12	13	1
Total registros		**21.316.847**

El procedimiento para la extracción de apellidos a partir de las cadenas de texto brutas comprende tres pasos de pre-procesamiento. Primero, cada cadena de texto se segmentó en unidades discretas (*tokens*) utilizando el espacio como delimitador. Segundo, se ejecutó un proceso de normalización que incluyó la eliminación de espacios múltiples y la conversión de todo el texto a mayúsculas para asegurar la consistencia. Finalmente, se depuraron los registros que carecían de información válida, descartando específicamente aquellas entradas con marcadores de texto como «NO HAY» o «NO HAY TARJETA».

Un desafío clave en el procesamiento de nombres hispánicos es el tratamiento de apellidos compuestos. Para abordar esta cuestión, previo a la segmentación, se implementó una función basada en expresiones regulares. El propósito de esta función es identificar y tratar aglomeraciones de apellidos (e.g., «DE LA TORRE») como una única unidad léxica, evitando su separación incorrecta.

Una vez pre-procesadas estas unidades, la extracción de los apellidos paterno y materno se rige por un algoritmo condicional que opera sobre el número de *tokens* en cada registro. Para el caso estándar, definido como registros con cuatro *tokens*, se asume la estructura [apellido1, apellido2, nombre1, nombre2] y se asignan los dos primeros *tokens* como los apellidos. Para registros con un número de *tokens* distinto, el algoritmo aplica un conjunto de reglas heurísticas. Estas reglas se fundamentan en la convención nominal hispánica, priorizando sistemáticamente los primeros *tokens* como apellidos bajo el supuesto de que estos preceden a los nombres. Si bien este procedimiento está diseñado para maximizar la precisión, se reconoce un margen de error inherente en casos con estructuras nominales atípicas o ambiguas. El resultado final de este protocolo es una base de datos estructurada con los apellidos paterno y materno de cada individuo, insumo fundamental para la construcción de las redes familiares proxy.

2.2.2. Depuración del registro administrativo de cedulados

El registro de «cedulados», que al igual que la base de apellidos comprende a más de 21 millones de individuos con datos históricos a 2022 (con registros que se remontan a las primeras décadas del siglo XX), constituye la fuente principal para la construcción de la red familiar de referencia. Cada registro individual contiene identificadores seudonimizados para el propio individuo, sus progenitores y su cónyuge, además de variables sociodemográficas clave como las fechas de nacimiento y defunción. A pesar de que esta base de datos fue sometida a una estandarización inicial por parte de la agencia estadística (véase Sección 2.1), nuestro protocolo de investigación implementó una fase de validación adicional. Este proceso tuvo como objetivo central verificar la coherencia lógica y cronológica de los vínculos de parentesco, con un enfoque particular en las relaciones paterno-filiales.

Dada la imposibilidad de verificar externamente la validez de los códigos seudonimizados, se implementó un protocolo de validación de consistencia interna. Este protocolo se basó en la eliminación de registros que violaban un conjunto de criterios lógicos y cronológicos, articulado en tres etapas:

1. **Consistencia de Identificadores:** Se excluyeron los registros con identificadores duplicados en roles familiares lógicamente incompatibles. Esto comprendió casos en los que un individuo compartía su identificador con el de uno de sus progenitores o su cónyuge.

2. **Consistencia Cronológica:** Se verificó la plausibilidad temporal de los vínculos. Utilizando las fechas de nacimiento y defunción, se identificaron y descartaron relaciones biológicamente inviables, tales como progenitores más jóvenes que sus hijos, nacimientos registrados *post-mortem* del progenitor, o intervalos inter-nacimiento para una misma madre inferiores al mínimo biológico (e.g., <196 días).

3. **Consistencia Estructural de Roles:** Se depuraron registros con asignaciones de roles familiares inconsistentes. Esto incluyó relaciones recíprocas (e.g., dos individuos listados mutuamente como padre e hijo) y casos donde un mismo individuo figuraba con roles de género inconsistentes a través de la base de datos (e.g., como padre en un registro y como madre en otro).

El volumen de registros eliminados en cada etapa de este proceso de depuración se detalla en la Tabla 2.

Tabla 2. Total de registros depurados por etapa

Etapas depuración	Total
1. Consistencia de Identificadores	607
2. Consistencia Cronológica	97.649
3. Consistencia Estructural de Roles	18.827
Total depuraciones	**117.083**

Por último, se llevó a cabo un proceso de normalización textual, con el fin de eliminar caracteres especiales, saltos de línea y espacios redundantes.

2.2.3. Construcción de redes familiares

Se construyeron dos tipos de redes familiares proxy basadas en apellidos. La primera, denominada *Proxy Paterno*, establece un vínculo no dirigido entre dos individuos si comparten su primer apellido (almacenado en la variable apellido1). La segunda, el *Proxy Compuesto*, define el vínculo si los individuos comparten la combinación exacta de su primer y segundo apellido (la concatenación de apellido1 y apellido2). En este marco, una 'familia' se define como el conjunto de individuos que comparten el mismo identificador de apellido (paterno o compuesto), y su tamaño corresponde al número de miembros en dicho conjunto.

La red de referencia, o red 'real', se construyó como un grafo no dirigido a partir de los vínculos de parentesco explícitos (padre, madre y cónyuge) disponibles en los registros administrativos. Esta estructura permite la identificación precisa de un amplio espectro de relaciones familiares, incluyendo hermanos (individuos que comparten al menos un progenitor), abuelos (padres de un progenitor), tíos, suegros y primos, entre otros. Es importante destacar una limitación inherente a los datos: el

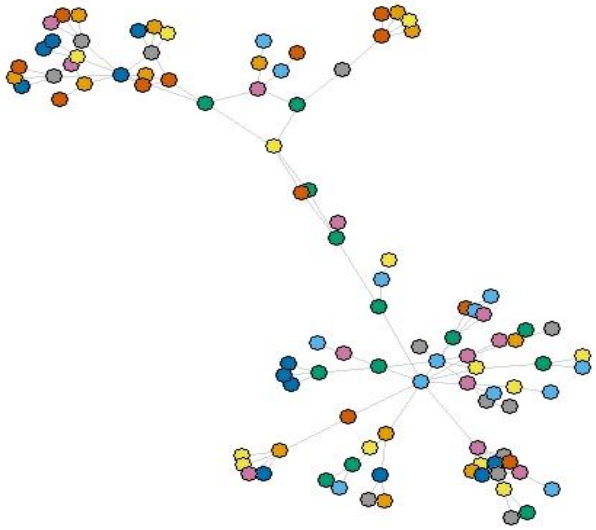
registro solo contiene la relación conyugal más reciente, lo que impide la observación de uniones pasadas. En esta red, una 'familia' se define formalmente como un componente conectado, y su tamaño es el número de nodos (individuos) dentro de dicho componente.

En ambos tipos de redes (proxy y de referencia) pueden existir nodos aislados. En la red de referencia, estos corresponden a individuos sin vínculos familiares registrados. En las redes proxy, un nodo aislado representa a un individuo cuyo apellido o combinación de apellidos es único en la base de datos. Para el análisis comparativo, nos centramos en dos métricas topológicas fundamentales: el número de familias (componentes) y la distribución de sus tamaños. Dada la extrema densidad de las redes proxy —lo que hace computacionalmente inviable su renderización explícita como grafos—, estas métricas se calcularon directamente a partir de las frecuencias de los apellidos. Este enfoque es suficiente para nuestros objetivos de investigación, que no requieren el análisis de la topología completa de las redes proxy. Por el contrario, la red de referencia, por su naturaleza más dispersa, fue construida y analizada computacionalmente utilizando los paquetes *networkx* en *Python* e *igraph* en *R*[9].

3. Resultados

La Figura 1 ofrece una visualización ilustrativa de un subgrafo extraído de la red de referencia. En esta representación, los individuos son denotados por nodos (esferas) y los vínculos de parentesco explícitos que los conectan son representados por enlaces (líneas).

Figura 1. Representación red familiar (subgrafo)

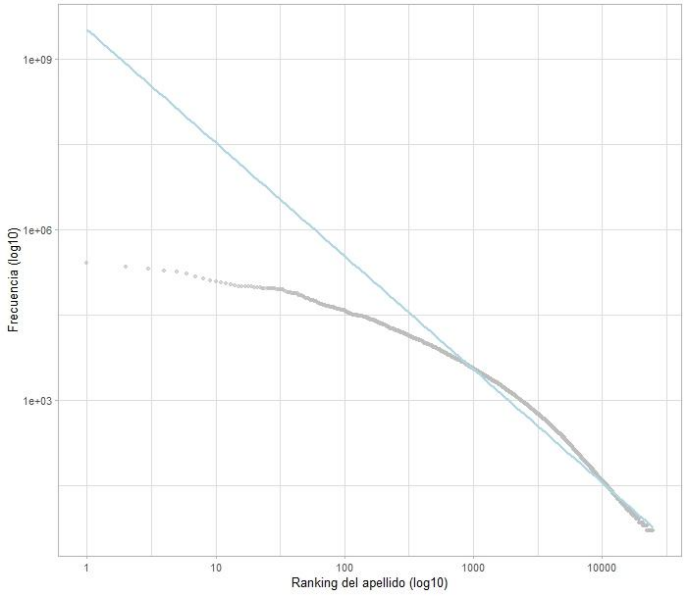


[9] Como decisión metodológica, en esta etapa inicial no se excluyen los registros de residentes sin nacionalidad ecuatoriana. Este enfoque inclusivo permite capturar la totalidad de los vínculos familiares presentes en el registro, incluyendo aquellos que trascienden la nacionalidad (e.g., entre extranjeros, o entre extranjeros y sus descendientes ecuatorianos).

Como se ha establecido, la construcción explícita de las redes proxy es computacionalmente inviable. Este desafío emana de la distribución de frecuencias de los apellidos, la cual, de manera consistente con la literatura (Panaretos 1989; Rossi 2013; Tucker 2013) presenta una cola pesada (*heavy-tailed*). Este patrón, a menudo compatible con una ley de potencia, implica que un número reducido de apellidos de alta frecuencia origina nodos con un grado de conectividad desproporcionadamente elevado.

Para verificar este comportamiento en nuestros datos, la Figura 2 presenta un gráfico de rango-frecuencia en escala logarítmica. Dicha visualización, donde una relación lineal es indicativa de una ley de potencia, muestra que la distribución de los apellidos más frecuentes se alinea con este patrón. Para formalizar esta observación, se realizó una regresión lineal sobre la transformación log-log de la frecuencia y el rango de los apellidos. El análisis del ajuste confirma que la linealidad es particularmente pronunciada para el segmento superior de la distribución (aproximadamente, los primeros 1.000 apellidos en el ranking), corroborando la presencia de un comportamiento tipo ley de potencia para los apellidos más comunes.

Figura 2. Distribución Log – Log de los apellidos



3.1. Comparación Topológica de las Redes

Con el fin de comparar las estructuras topológicas resultantes de cada método de construcción, se generaron subgrafos de 500 nodos, inducidos aleatoriamente para cada una de las tres redes: la de referencia, la del proxy paterno y la del proxy compuesto. Este procedimiento de muestreo a tamaño fijo permite aislar las diferencias en los patrones de conectividad que son directamente atribuibles a la definición del vínculo, controlando por los efectos de escala que de otro modo confundirían la comparación.

Figura 3. Subgrafos de redes familiares construidas según aproximación

A pesar de las diferencias en sus mecanismos de generación y en la naturaleza de los vínculos que representan, la comparación de redes a través de sus propiedades estructurales es una práctica metodológica consolidada. La literatura sobre análisis de redes ha demostrado que grafos de distinta naturaleza —que representan desde amistad hasta participación política— tienden a exhibir regularidades topológicas en métricas como el grado o la centralidad (Hashmi et al. 2012). De hecho, análisis sobre amplias bases de datos de redes empíricas confirman que ciertas configuraciones estructurales son recurrentes a pesar de la diversidad de sus orígenes, lo que sugiere la existencia de patrones sociales subyacentes (Kantarci y Labatut 2013).

Este principio encuentra un correlato directo en la literatura econométrica. De particular relevancia para nuestro estudio, Goldsmith-Pinkham e Imbens (2013) argumentan que, incluso cuando los vínculos se generan mediante *proxies* y están sujetos a error de medición, el análisis estructural puede revelar patrones sociales válidos bajo supuestos de identificación adecuados. Más recientemente, Jochmans (2023) refuerza esta idea proponiendo estrategias de inferencia basadas en subestructuras observables, como componentes aislados o subconjuntos excluyentes, para identificar patrones válidos incluso en redes parcialmente observadas.

En línea con la literatura citada, nuestro análisis comparativo se centra en dos propiedades topológicas fundamentales: el número de componentes conectados ('familias') y la distribución de sus tamaños. Esta última métrica es particularmente informativa sobre la capacidad de una red para formar clústeres sociales. A diferencia de métricas de conectividad global o de grado nodal, el análisis de componentes captura la estructura de grupos cohesivos donde todos los miembros están conectados, ya sea directa o indirectamente, lo que lo hace ideal para el estudio de agrupaciones familiares.

Los resultados de esta comparación, detallados en la Tabla 3, revelan diferencias estructurales significativas entre las redes. La red de referencia identifica aproximadamente 3,5 millones de familias; sin embargo, este número está dominado por nodos aislados (componentes de tamaño uno). Al restringir el análisis

a familias con más de un miembro —un enfoque más relevante para el estudio de interacciones—, el número de componentes en la red de referencia desciende a aproximadamente 390.000. Frente a este punto de referencia, los *proxies* exhiben sesgos opuestos: el *Proxy Paterno* subestima marcadamente el número de familias (≈ 89.000), mientras que el *Proxy Compuesto* lo sobreestima de manera considerable (≈ 1,7 millones). Este patrón es consistente con los mecanismos de construcción de los *proxies*, pues mientras el apellido paterno es un criterio demasiado inclusivo, requerir la coincidencia de ambos apellidos resulta excesivamente restrictivo. Estas divergencias se reflejan también en el tamaño promedio de las familias, donde se observa un patrón de sesgo análogo.

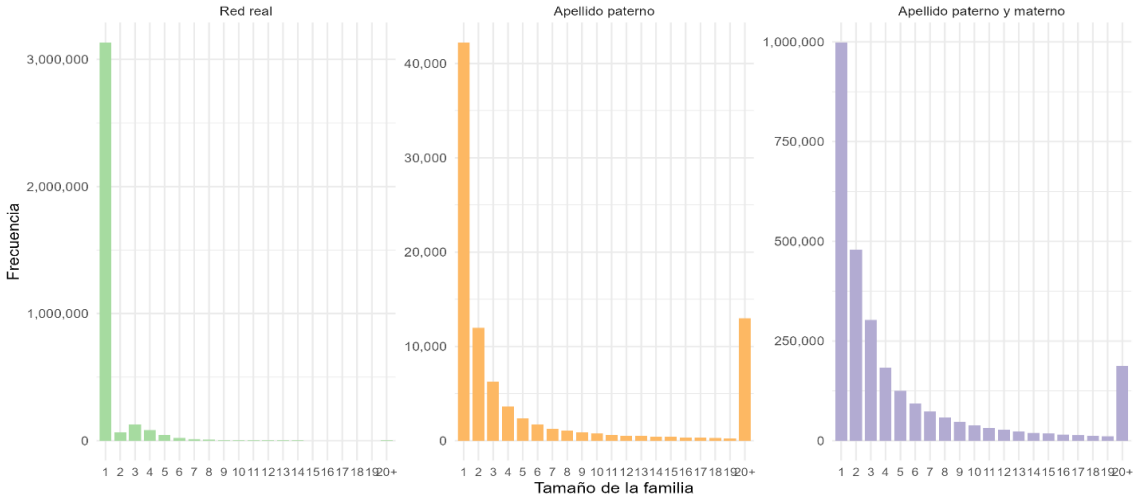
Tabla 3. Características generales de las redes familiares

Red		Número de familias (componentes conectados)	Número de personas (nodos)	Tamaño promedio de familia
Real (referencia)	Con nodos aislados	3.522.810	21.159.334	6
	Sin nodos aislados	389.730	18.026.254	46
Proxy Paterno	Con nodos aislados	88.791	21.300.777	240
	Sin nodos aislados	46.576	21.258.562	456
Proxy compuesto	Con nodos aislados	2.757.579	21.300.777	8
	Sin nodos aislados	1.759.041	20.302.239	12

Quizás el hallazgo más sorprendente de la red de referencia es la existencia de un componente gigante' que aglutina a aproximadamente 16,3 millones de individuos. La presencia de este macro-componente es el principal factor que explica el elevado tamaño promedio de las familias en la Tabla 3 (al excluir nodos aislados). La emergencia de una estructura de esta magnitud puede representar una característica fundamental y hasta ahora no documentada de la red social del país, o bien, podría ser parcialmente un artefacto de relaciones espurias generadas durante el proceso de seudonimización o por errores en los registros originales. Precisamente por esta dualidad, el análisis de la formación, estructura y, crucialmente, la robustez de este componente gigante frente a potenciales errores de registro, constituye el foco de una investigación complementaria en curso.

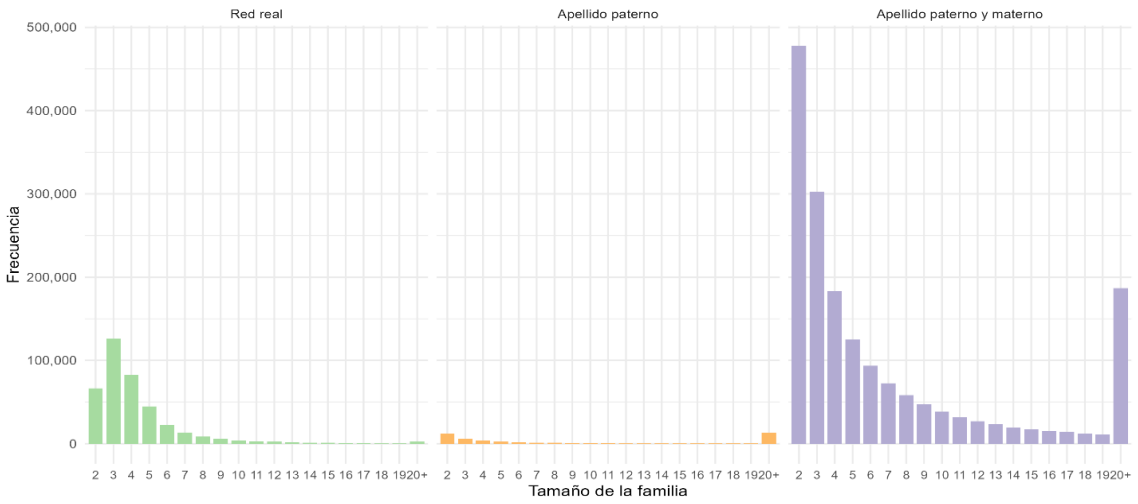
El análisis se extiende a la distribución del tamaño de las familias, presentada en la Figura 4. A primera vista, las distribuciones de las redes proxy parecen replicar la forma general de la red de referencia, caracterizada por una alta concentración en familias de tamaño reducido. Sin embargo, esta aparente similitud es un artefacto de dos factores de confusión: la elevada prevalencia de nodos aislados (familias de tamaño uno) y las drásticas diferencias en la escala de frecuencia (eje Y) entre las tres redes, que impiden una comparación directa.

Figura 4. Distribución del tamaño de familias



Para facilitar una comparación rigurosa, la Figura 5 presenta las mismas distribuciones con dos ajustes metodológicos: excluye los componentes de tamaño uno y normaliza la escala del eje de frecuencia. Bajo estas condiciones controladas, emergen sesgos sistemáticos y opuestos. La distribución del *Proxy Paterno* se sitúa consistentemente por debajo de la red de referencia, indicando una subestimación del número de familias en cada tramo de tamaño. Inversamente, la distribución del *Proxy Compuesto* se ubica por encima, lo que evidencia una sobreestimación sistemática. Este patrón de subestimación y sobreestimación es plenamente consistente con los hallazgos del análisis de conteo de componentes (Tabla 3), reforzando la conclusión sobre la naturaleza de los sesgos introducidos por cada tipo de proxy.

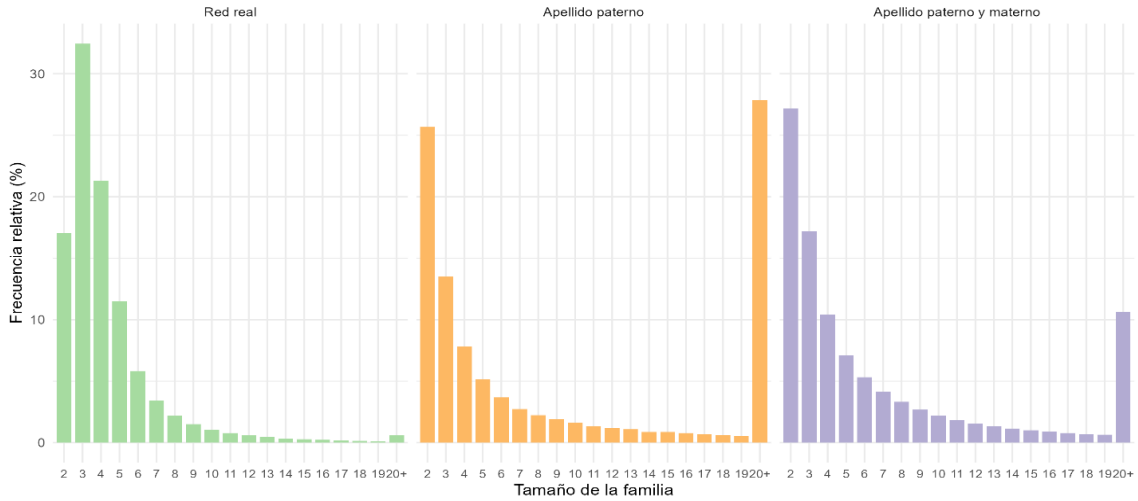
Figura 5. Distribución del tamaño de familias con más de un miembro (misma escala)



La Figura 6 transita del análisis de conteos absolutos a la comparación de las distribuciones relativas del tamaño de las familias (i.e., la proporción de familias de cada tamaño). A nivel general, las distribuciones de ambos proxies se asemejan a

la forma de la red de referencia, capturando la alta prevalencia de familias de tamaño pequeño.

Figura 6. Distribución relativa del tamaño de familias con más de un miembro



Sin embargo, se observa una divergencia sistemática en la cola derecha de la distribución. Específicamente, ambos proxies sobreestiman la proporción de familias grandes (definidas aquí como aquellas con 20 o más miembros). La causa de este sesgo difiere según el proxy:

1. En el caso del **Proxy Paterno**, la sobreestimación es un resultado directo de la aglomeración de individuos no emparentados bajo apellidos comunes, lo que crea un exceso de componentes artificialmente grandes.

2. En el del **Proxy Compuesto**, el sesgo surge de un mecanismo diferente. Aunque este método fractura el 'componente gigante' de la red de referencia, lo hace generando un número de componentes de tamaño intermedio-grande que es desproporcionadamente mayor al observado en la red de referencia (donde la mayoría de los componentes no-gigantes son muy pequeños).

Para cuantificar la similitud entre las distribuciones de tamaño, complementamos el análisis visual con el cálculo de la Distancia de Movimiento de Tierra (*Earth Mover's Distance*, EMD). Siguiendo a Kantarci y Labatut (2013), empleamos esta métrica no paramétrica que estima el 'costo' agregado de transformar una distribución en otra, proveyendo una medida robusta de la proximidad entre sus formas generales. Los resultados de este cálculo se presentan en la Tabla 4.

Tabla 4. EMD entre distribuciones de tamaño de familias

	Red real	Proxy Paterno	Proxy Compuesto
Red real	0,000	0,022	0,021
Proxy Paterno	0,022	0,000	0,012
Proxy Compuesto	0,021	0,012	0,000

Los valores de la EMD obtenidos para ambos proxies son relativamente bajos, lo que indica que, en un sentido agregado, sus distribuciones no divergen drásticamente de la red de referencia. Este resultado cuantitativo no contradice el análisis visual anterior; más bien, lo matiza. Sugiere que si bien los proxies introducen sesgos locales y sistemáticos —particularmente en la sobreestimación de la cola de la distribución—, la forma global de sus distribuciones de tamaño mantiene una similitud estructural considerable con la de la red de referencia. Por lo tanto, aunque imperfectos, los proxies logran replicar la estructura general de agrupamiento familiar.

4. Discusión

El presente estudio constituye un primer paso fundamental en la construcción y validación de una red familiar a escala nacional para Ecuador. Al cuantificar las discrepancias entre la red de referencia y los proxies de apellidos, este trabajo sienta un precedente para el desarrollo de nuevas líneas de investigación empírica basadas en registros administrativos. Las cuales actualmente se encuentran siendo desarrolladas en las temáticas de análisis estructural de la red y sus componentes, estimación de efectos sociales mediados por vínculos familiares, evaluación de sesgos derivados del uso de información proxy, y diseño de estrategias metodológicas para su corrección. Asimismo, la disponibilidad de una red validada abre la posibilidad de explorar aplicaciones como el análisis de estructuras sociales, su evolución en el tiempo y su incidencia en distintos ámbitos económicos y sociales, tanto en el sector público como privado.

Estas líneas futuras no solo permitirán avanzar en el conocimiento sobre la función de las redes familiares en diversos contextos, sino también aportar evidencia útil para el diseño de políticas públicas fundamentadas en la estructura relacional de la población.

5. Conclusiones

La investigación empírica sobre los efectos de las redes familiares ha dependido de manera abrumadora de proxies indirectos, como son los apellidos compartidos (Brassiolo et al. 2021; Mendoza & Bro, 2021), cuya validez rara vez ha sido contrastada con datos de referencia a gran escala. Este trabajo aborda esta brecha fundamental mediante la construcción de la red familiar casi completa para Ecuador y una comparación sistemática de su topología con la de las redes generadas por los proxies de apellidos más comunes.

Nuestros hallazgos revelan que, si bien los proxies pueden replicar ciertos aspectos generales de la estructura familiar, introducen sesgos significativos y sistemáticos. Demostramos que los proxies tergiversan el número total de familias —subestimándolo en un caso y sobreestimándolo en otro— y distorsionan la distribución de sus tamaños.

La implicación principal de nuestro estudio es que las estimaciones de efectos de red basadas en proxies de apellidos pueden estar sujetas a un error de medición no clásico y de dirección impredecible, lo que exige una reevaluación cautelosa de sus resultados. Al documentar la naturaleza de estos sesgos y proveer una red de

referencia validada, este documento ofrece un insumo crucial y sienta las bases para una nueva generación de investigaciones más precisas sobre el impacto causal de las redes familiares en los ámbitos económico, político y social.

6. Bibliografía

Balan, Pablo, Dodyk, Juan, & Puente, Ignacio. 2025. "Family Ties as Corporate Power". *The Journal of Politics*.

Banerjee, Abhijit, Chandrasekhar, Arun G., Duflo , Esther, & Jackson , Matthew O. 2013. "The Diffusion of Microfinance". *Science* 341 (6144): 1236498.

Brassiolo, Pablo, Estrada, Ricardo, Fajardo, Gustavo & Martinez-Correa Julian. 2021. "Family Rules: Nepotism in the Mexican Judiciary". *Mimeo*.

Brassiolo, Pablo, Estrada, Ricardo & Fajardo, Gustavo. 2020. "My (Running) Mate, the Mayor: Political Ties and Access to Public Sector Jobs in Ecuador". *Journal of Public Economics* 191: 104286.

Brugués, Felipe, Javier Brugués, & Samuele Giambra. 2024. "Political Connections and Misallocation of Procurement Contracts: Evidence from Ecuador". *Journal of Development Economics* 170: 103296.

Chandrasekhar, & Randall Lewis. 2016. "Econometrics of Sampled Networks".

Cruz, Cesi, Labonne, Julien, & Querubín , Pablo. 2017. "Politician Family Networks and Electoral Outcomes: Evidence from the Philippines". *American Economic Review* 107 (10): 3006–37.

Dal-Bó, Ernesto, Dal-Bó, Pedro, & Jason Snyder. 2009. "Political Dynasties". *The Review of Economic Studies* 76 (1): 115–42.

Durante, Ruben, Labartino, Giovanna, & Perotti, Roberto. 2011. "Academic Dynasties: Decentralization and Familism in the Italian Academia".

Fafchamps, Marcel & Labonne, Julien. 2017. "Do Politicians' Relatives Get Better Jobs? Evidence from Municipal Elections." *The Journal of Law, Economics, and Organization*, 268–300.

Gagliarducci, Stefano & Manacorda, Marco. 2016. "Politics in the Family: Nepotism and the Hiring Decisions of Italian Firms". *SSRN Electronic Journal*.

Goldsmith-Pinkham, Paul, & Imbens, Guido W. 2013. "Social Networks and the Identification of Peer Effects". *Journal of Business & Economic Statistics* 31 (3): 253–64.

Hashmi, A., F. Zaidi, Sallaberry & T. Mehmood. 2012. "Are All Social Networks Structurally Similar?" *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 310–14.

INEC. 2022. "Metodología Para Transformar Registros Administrativos En Registros Estadísticos". https://www.ecuadorencifras.gob.ec/documentos/web-inec/Bibliotecas/Libros/Metod_para_transformar_registros_admin_en_registros_estad.pdf.

INEC. 2024. "Documento Metodológico Del Registro Estadístico Base de Población Del Ecuador (REBPE)". https://www.ecuadorencifras.gob.ec/documentos/web-inec/Registros_Administrativos/rebpe/2022/202212_Metodologia_REBPE.pdf.

Jochmans, Koen. 2023. "Peer Effects and Endogenous Social Interactions". *Journal of Econometrics* 235 (2): 1203–14.

Kantarci, Burcu, & Labatut, Vincent. 2013. "Classification of Complex Networks Based on Topological Properties". *2013 International Conference on Cloud and Green Computing*, 297–304.

Mendoza, M., & Bro, N. (2021). Predicting affinity ties in a surname network. *PLoS One*.

Panaretos, John. 1989. "On the Evolution of Surnames". *International Statistical Review / Revue Internationale de Statistique* 57 (2): 161.

Querubin, Pablo. 2016. "Family and Politics: Dynastic Persistence in the Philippines". *Quarterly Journal of Political Science* 11 (2): 151–81.

Riaño, Juan Felipe. 2021. "Bureaucratic Nepotism". *SSRN Electronic Journal*.

Rossi, Paolo. 2013. "Surname Distribution in Population Genetics and in Statistical Physics". *Physics of Life Reviews* 10 (4): 395–415.

Statistics-Denmark. 1995. "Statistics on Persons in Denmark–Aregister-Based Statistical System". Luxembourg: Eurostat.

Siddharth, George. 2020. "Like Father, like Son? The Effect of Political Dynasties on Economic Development". *Mimeo*.

Tucker, D. K. 2013. "Distribution of Forenames, Surnames, and Forename–Surname Pairs in the United States". *Names* 49 (2): 69–96.

Wallgren, Anders & Wallgren, Britt. 2007. *Register-Based Statistics: Administrative Data for Statistical Purposes*. John Wiley & Sons.

INEC | Buenas cifras, **mejores vidas**

@ecuadorencifras

@ecuadorencifras

@InecEcuador

t.me/ecuadorencifras

INEC/Ecuador

INECEcuador

Administración Central (Quito)
Juan Larrea N15-36 y José Riofrío,
Teléfonos: (02) 2544 326 - 2544 561 Fax: (02) 2509 836
Código postal: 170410
correo-e: inec@inec.gob.ec

www.ecuadorencifras.gob.ec