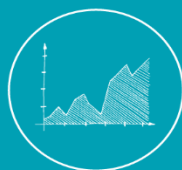
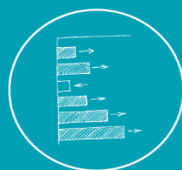




Una aproximación metodológica a la predicción de la etnia: el caso del Ecuador

Septiembre 2024



Autoridades:

Roberto Castillo A.
Director Ejecutivo

Jorge García-Guerrero
Subdirector General

Darío Vélez J.
*Coordinador General Técnico de Innovación en
Métricas y Análisis de la Información*

Carmen Granda E.
Directora de Estudios y Análisis de la Información

Autores:

Diego Del Pozo V.
Instituto Nacional de Estadística y Censos, Ecuador

Andrés Villacís M.
Instituto Nacional de Estadística y Censos, Ecuador

Los Cuadernos de Trabajo Temáticos son documentos que presentan análisis de fenómenos sociales, económicos y ambientales con el objetivo de promover la investigación e incentivar el debate.

Las interpretaciones y opiniones expresadas en este documento pertenecen a los autores y no reflejan el punto de vista oficial del Instituto Nacional de Estadística y Censos (INEC). El INEC ha realizado una revisión del documento, no obstante, no garantiza la exactitud de los datos que figuran en el documento.

Expresamos nuestro agradecimiento a Kathia Pinzón y Cristhian Rosales por su invaluable contribución, la cual fue fundamental para la realización exitosa de este trabajo.

Una aproximación metodológica a la predicción de la etnia: el caso del Ecuador

Diego Del Pozo V.¹; Andrés Villacís M.¹

Resumen

El estudio se centra en la aplicación del método *Bayesian Improved Surname Geocoding* (BISG) para predecir la etnia en el Censo de Población y Vivienda (CPV) de 2010 en Ecuador. Se realizan pruebas exhaustivas del método, explorando diversas combinaciones de apellidos y niveles de desagregación geográfica (provincial, cantonal y parroquial). Se emplearon métricas de evaluación para determinar el desempeño del método, incluyendo el *accuracy* y el ratio de falsos positivos (errores tipo 1) y el ratio de falsos negativos (errores tipo 2). Además, se incorporan covariables como el sexo y la edad con el objetivo de analizar una mejora del rendimiento del método. Finalmente, se llevó a cabo una validación intertemporal del método utilizando datos de la Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU) anual de 2023.

Palabras clave: *Bayesian Improved Surname Geocoding*, BISG, predicción étnica, análisis de apellidos

Introducción

La información étnica tiene alta relevancia en diversas áreas sociales, como la salud, la economía y la política. En el ámbito socioeconómico, su análisis ha permitido identificar y abordar disparidades en equidad entre distintos grupos poblacionales, lo que resulta clave para formular políticas públicas efectivas que reduzcan la desigualdad y promuevan un desarrollo inclusivo (Galarza, 2010). En el campo de la salud, la información étnica es esencial, ya que las diferencias biológicas entre grupos permiten identificar enfermedades específicas y personalizar tratamientos, mejorando así los pronósticos médicos (Adjaye-Gbewonyo et al., 2014; Mays et al., 2003). Asimismo, en el ámbito político y electoral, comprender la información étnica es crucial para entender los patrones de participación y las preferencias de voto, lo que contribuye al desarrollo de estrategias más representativas e inclusivas (Barreto et al., 2022; Imai y Khana, 2016).

Por esta razón, es crucial recopilar información étnica en diversas operaciones estadísticas. Sin embargo, dado que existen limitaciones en la recolección de esta información, se han desarrollado técnicas que permiten predecir la etnia utilizando otras variables. Entre las técnicas evaluadas en la literatura se encuentran los métodos *geocoding only* - GO, *Categorical Surname and Geocoding approach* (CSG) y el enfoque *Bayesian Surname Geocoding* (BSG), los cuales integran apellidos e información geográfica para una estimación más precisa de la etnia. Además, a través de la utilización de registros detallados de apellidos, se mejora al método BSG hacia una nueva metodología denominada *Bayesian Improved Surname Geocoding* (BISG).

En este estudio se aplicó el método BISG sobre el Censo de Población y Vivienda (CPV) de 2010 en Ecuador para predecir la etnia de los individuos. El enfoque del método se basó en una estimación bayesiana la cual permitió predecir la identificación étnica de las personas, utilizando sus apellidos y la ubicación geográfica de su residencia. Sin embargo, con el fin de fortalecer el análisis, se llevaron a cabo diversas pruebas del método, las que incluyeron el uso del primer

¹ Los investigadores pertenecen a la Dirección de Estudios y Análisis de la Información del INEC. Para requerimiento de información del estudio dirijase a: inec@inec.gob.ec.

apellido, el segundo apellido, ambos apellidos agrupados y la combinación de ambos apellidos, en diferentes niveles de desagregación geográfica (provincial, cantonal y parroquial).

Una vez realizada la predicción étnica en el CPV 2010, se la comparó con la etnia autodeclarada real de los individuos y, se construyeron matrices de confusión para estimar métricas de evaluación, tales como la exactitud (*accuracy*)², el ratio de falsos positivos (error tipo 1)³ y el ratio de falsos negativos (error tipo 2)⁴. Estas métricas permitieron evaluar el rendimiento de cada una de las aproximaciones del método BISG. Los resultados mostraron que considerar ambos apellidos agrupados a nivel parroquial fue la técnica de mayor *accuracy* y con menores errores de predicción. Para la identificación del mejor método BISG y siguiendo el enfoque de Imai y Khana (2016), se incorporaron covariables adicionales, tales como el sexo y la edad. Esto con el propósito de proporcionar al método información adicional y, por ende, mejorar aún más el rendimiento del método. Los resultados obtenidos mostraron una ligera mejora en términos de *accuracy*, lo que llevó a definir este método como el preferido para su utilización.

Posteriormente, con la finalidad de llevar a cabo una validación intertemporal del método identificado como el más eficaz, se realizó una prueba de la matriz BISG construida a través del CPV 2010 sobre la Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU) anual de 2023. Los resultados evidenciaron una notable mejora en el *accuracy*. No obstante, debido a que la base de datos de la ENEMDU anual de 2023 tiene menor tamaño en comparación con el CPV 2010, se observó un aumento en los errores tipo 1 y tipo 2. Como complemento, se replicó la construcción de la matriz BISG a partir del CPV 2010, pero esta vez se agruparon las categorías étnicas, para finalmente probar la matriz en la base de datos de ENEMDU anual 2013. Los resultados alcanzados mostraron mejoras en la exactitud de la predicción étnica para ambas fuentes de información en comparación con el análisis original.

La estructura del documento se divide de la siguiente manera: en la primera sección se presenta una revisión de literatura sobre la etnia y su relevancia estadística, así como una revisión de las técnicas estadísticas empleadas en la predicción étnica. La segunda sección describe los datos utilizados y los criterios metodológicos aplicados en el estudio. La tercera sección exhibe los resultados obtenidos en las diversas aproximaciones de la implementación del método BISG. Finalmente, la cuarta sección resume las conclusiones y recomendaciones derivadas del trabajo.

1. Revisión de Literatura

1.1. La etnia y su importancia estadística

La etnia, desde una perspectiva etimológica, proviene del término griego "*ethnos*", que se traduce como pueblo o nación. Este concepto se refiere a aspectos culturales y está usualmente ligado a la construcción de identidad de los individuos dentro de un grupo cultural específico. Este proceso de identidad o identificación es donde las personas se definen a sí mismas, y los elementos distintivos de su cultura se confrontan y entrelazan con los de otros, formando así sus identidades únicas (Galarza, 2010).

La literatura resalta la importancia de la información étnica en diversos sectores, destacando su rol en el ámbito socioeconómico para identificar y abordar las brechas de equidad entre distintos pueblos y nacionalidades frente a grupos que se han beneficiado de los modelos de desarrollo

² Mide el porcentaje de predicciones correctas realizadas por un modelo en relación con el total de predicciones (Hasan et al., 2021).

³ Número de instancias en las que el modelo clasifica incorrectamente un caso como positivo (pertenece a una clase específica) cuando el caso es en realidad negativo (no pertenece a la clase específica) (Hasan et al., 2021).

⁴ Número de instancias en las que el modelo clasifica incorrectamente un caso como negativo (no pertenece a una clase específica) cuando el caso es en realidad positivo (pertenece a la clase específica) (Hasan et al., 2021).

y las políticas gubernamentales (Consejo Nacional para la Igualdad de Pueblos y Nacionalidades, 2019; Galarza, 2010).

En el sector de la salud, la identificación étnica resulta clave para obtener estadísticas precisas de morbilidad y mortalidad, facilitando la planificación de políticas públicas efectivas, el monitoreo de necesidades y la identificación de riesgos para la salud. La identificación étnica también permite desarrollar indicadores de riesgo específicos, aprovechando las diferencias biológicas, de comportamiento y de exposición a factores ambientales para detectar susceptibilidades a enfermedades y respuestas a tratamientos, lo que contribuye a disminuir la incertidumbre en los pronósticos de salud (Adjaye-Gbewonyo et al., 2014; Mays et al., 2003).

En el ámbito político y electoral, la literatura señala que la composición étnica de la población es crucial para determinar la participación y la intención de voto de los individuos. Muchos de los analistas políticos emplean esta información para identificar las disparidades en la participación política de los individuos y para detectar polarización étnica (Barreto et al., 2022; Imai y Khana, 2016).

En este contexto, en Ecuador, la recopilación de la información étnica ha cobrado alta importancia como un medio para construir políticas públicas, las cuales se encuentren destinadas a reducir desigualdades y promover el desarrollo integral de los pueblos y nacionalidades, respetando sus principios y cosmovisiones. Esta información se ha convertido en un recurso crucial para la toma de decisiones, la asignación de recursos, especialmente de sus pueblos y nacionalidades (Galarza, 2010).

El proceso de recolección de datos sobre etnicidad en Ecuador comenzó en 2001 con el VI Censo de Población y el V de Vivienda, que incorporó aspectos como la lengua hablada y la autoidentificación étnica. A partir de ese momento, se incluyeron preguntas sobre identificación étnica en otras operaciones estadísticas relevantes, como la Encuesta Demográfica de Salud Materna Infantil (ENDEMAIN) de 2004, la Encuesta de Condiciones de Vida (ECV) de 2006 y la Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU). Estas encuestas han sido esenciales para entender mejor la composición étnica del país y guiar políticas más inclusivas y efectivas.

1.2. Estimación de la etnia en operaciones estadísticas

La información sobre la etnia de los individuos juega un papel crucial en el análisis estadístico, la planificación gubernamental y el estudio de varias dinámicas sociales. Esta importancia motiva la búsqueda de métodos para asegurar la disponibilidad de la información, permitiendo su inferencia a través de otras variables o dimensiones en casos donde no esté directamente disponible.

En este escenario, la identificación étnica de las personas es determinada a partir de una variedad de elementos clave. Por ejemplo, en el caso de pueblos y nacionalidades indígenas sobresalen factores como su linaje, haciendo referencia a las raíces familiares y el territorio en el que habitan, por la importancia de la tierra el desarrollo de su vida. De igual importancia es la autoidentificación o el autorreconocimiento, el cual se refiere a la conciencia de pertenecer a un grupo culturalmente distinto y sirve como una reafirmación de su vinculación a un grupo específico (Galarza, 2010). No obstante, dada la complejidad de capturar la etnia a través de múltiples dimensiones, se ha priorizado el uso de variables comúnmente disponibles en las fuentes de información, que permiten la diferenciación de grupos poblacionales según su etnia. Entre estas se incluyen la ubicación geográfica de la residencia de las personas (geocodificación) y

sus apellidos (Elliot et al., 2009), aspectos que recogen elementos determinantes como son el sentido de autorreconocimiento de los individuos y sus raíces familiares.

Elliot et al. (2009) destacan la viabilidad de estimar la información étnica utilizando variables como la geocodificación y el análisis de apellidos. Métodos como el uso único de la geocodificación (*geocoding only* - GO) han sido aplicados a partir de datos censales y han permitido identificar la etnia de grandes grupos poblacionales basándose en su lugar de residencia. Sin embargo, esta técnica resulta ser limitada para identificar la etnia en grupos específicos. Por lo que, se planteó el uso de la información de apellidos, a través de diccionarios que correlacionan apellidos con etnias, logrando una identificación más precisa, aunque los resultados aún fueron restringidos para algunos grupos étnicos.

Ante las limitaciones de estos dos criterios analizados de manera aislada, se han propuesto métodos híbridos como el enfoque de *Categorical Surname and Geocoding approach* (CSG) (Fiscella y Fremont, 2006) y el enfoque *Bayesian Surname Geocoding* (BSG) (Elliot et al., 2008), que integran apellidos e información censal para una estimación más eficaz de la etnia. Posteriormente, con la construcción mucho más exhaustiva de una tabulación de apellidos, se mejoró el método BSG hacia una nueva metodología denominada *Bayesian Improved Surname Geocoding* (BISG).

2. Datos y metodología

El estudio se enfocó en la aplicación del método BISG en el contexto ecuatoriano utilizando datos censales del Censo de Población y Vivienda (CPV) de 2010, así como datos de ENEMDU anual de 2023⁵. Se detalla la selección y el análisis de las fuentes de información, así como la especificación del método para cada ejercicio llevado a cabo.

2.1. Datos

El estudio inició con el análisis del Censo de Población y Vivienda (CPV) de 2010 de Ecuador, que proporciona una gran cantidad de información sociodemográfica para un total de 14.483.499 personas, incluyendo información detallada como apellidos, sexo, edad, lugar de residencia y etnia autodeclarada. Se complementó esta información con la base de datos del visualizador de nombres y apellidos del Instituto Nacional de Estadística y Censos (INEC) de 2022⁶, que fue usada para validar los registros a través de un contraste de apellidos. La Tabla 1 proporciona una estadística descriptiva del comportamiento de la información étnica en el CPV 2010, así como de su depuración después de validar los apellidos.

Tabla 1. Descripción de la variable de etnia en el CPV 2010

Fuentes de información	Descripción étnica	Número de registros de la fuente	Registros depurados primer apellido*	Registros depurados Segundo apellido**	Registros depurados ambos apellidos agrupados***	Registros depurados ambos apellidos combinados****
CPV 2010	Indígena	1.018.176 (7,03%)	984.191 (6,98)	774.580 (6,52%)	984.191 (6,98)	1.018.030 (7,03%)
	Afroecuatoriana (o)	615.262 (4,25%)	595.814 (4,22%)	489.126 (4,12%)	595.814 (4,22%)	615.110 (4,25%)
	Negra (o)	145.398 (1,00%)	139.626 (0,99%)	112.011 (0,94%)	139.626 (0,99%)	145.359 (1,00)

⁵ La ENEMDU anual es una encuesta que agrega las muestras de las encuestas mensuales dentro del periodo de enero a diciembre de un año definido. La acumulación de 12 meses permite que la encuesta cuente con cobertura nacional, urbano, rural, 5 ciudades autorepresentadas (Quito, Guayaquil, Cuenca, Ambato y Machala) y 24 provincias (INEC, 2021).

⁶ La base de datos proviene de información depurada del registro administrativo de la Dirección General de Registro, Identificación y Cedulación (DIGERCIC).

	Mulata (o)	280.899 (1,94%)	272.037 (1,93%)	223.339 (1,88%)	272.037 (1,93%)	280.812 (1,94%)
	Montubia (o)	1.070.728 (7,39%)	1.054.181 (7,47%)	961.771 (8,10%)	1.054.181 (7,47%)	1.070.675 (7,39%)
	Mestiza (o)	10.417.299 (71,93%)	10.161.669 (72,03%)	8.573.370 (72,20%)	10.161.669 (72,03%)	10.414.677 (71,92%)
	Blanca (o)	882.383 (6,09%)	850.144 (6,03%)	700.605 (5,90%)	850.144 (6,03%)	882.199 (6,09%)
	Otra (o)	53.354 (0,37%)	49.608 (0,35%)	39.068 (0,33%)	49.608 (0,35%)	53.333 (0,37%)

Nota:

*La depuración consistió en tomar únicamente a los primeros apellidos existentes en la tabla de apellidos del INEC 2022.

**La depuración consistió en tomar únicamente los segundos apellidos existentes en la tabla de apellidos del INEC 2022.

*** La depuración consistió en tomar únicamente los primeros apellidos existentes en la tabla de apellidos del INEC 2022. Se mantuvieron los apellidos maternos vacíos.

****La depuración consistió en eliminar casos en el que tanto el primer apellido como el segundo se encuentran vacíos.

Además, con el fin de llevar a cabo una validación intertemporal del desempeño del mejor método BISG implementado sobre el CPV 2010, se aplicó el mismo a la ENEMDU anual de 2023, la cual contiene un total de 304.637 registros. Este proceso buscó confirmar si la eficacia del método mejora al aplicarlo en fuentes de información más recientes.

2.2. Metodología

El método BISG aplicado en este trabajo se inspira en la metodología propuesta por Elliot et al. (2009). Este enfoque utiliza una estimación bayesiana⁷ para pronosticar la identificación étnica de los individuos, basándose en sus apellidos⁸ y la ubicación geográfica de su residencia (geolocalización). Para lo cual, se asume que la ubicación geográfica y el apellido de las personas son condicionalmente independientes de su etnia. Esto significa que, si se conoce la etnia de un individuo, su apellido no provee información de su ubicación geográfica y viceversa, como se expresa en (1).

$$G_i \perp\!\!\!\perp S_i | R_i \quad (1)$$

La expresión bayesiana parte de definir la probabilidad de un individuo i de pertenecer a la etnia r , dado que tiene el apellido s $P(R_i = r | S_i = s)$ ⁹ y la probabilidad de un individuo i de residir en la ubicación geográfica g , dado que pertenece a la etnia r $P(G_i = g | R_i = r)$ ¹⁰. Con las probabilidades previas se establece la probabilidad de un individuo i de tener la etnia r , dado que tiene el apellido s y reside en la ubicación geográfica g , como se expresa en (2).

$$P(R_i = r | S_i = s, G_i = g) = \frac{P(G_i = g | R_i = r) P(R_i = r | S_i = s)}{\sum_{r' \in R} P(G_i = g | R_i = r') P(R_i = r' | S_i = s)} \quad (2)$$

2.2.1. Extensión del método

Siguiendo a Imai y Khana (2016) se prueba una aproximación adicional, en la que se incluyen otras variables sociodemográficas, como son el sexo y la edad de los individuos. El objetivo de la aproximación es mejorar la exactitud de la predicción de la etnia obtenido por el ejercicio original. La incorporación de las covariables (X_i) genera un cambio en la expresión de independencia condicional de (1), obteniendo la expresión (3).

⁷ La especificación del método presentada en este trabajo se inspira en el trabajo realizado por Imai y Khanna (2016).

⁸ En el Anexo 1 se presenta una vista de los 30 primeros lugares del ranking de apellidos por etnia para el CPV 2010 generado en este trabajo de acuerdo con Elliot et al. (2009).

⁹ El Anexo 2 presenta un ejemplo de la matriz de probabilidad de un individuo i de pertenecer a la etnia r , dado que tiene el apellido s .

¹⁰ El Anexo 3 presenta un ejemplo de la matriz de probabilidad de un individuo i de residir en una ubicación g y pertenecer a una etnia r .

$$\{G_i, X_i\} \perp\!\!\!\perp S_i | R_i \quad (3)$$

De este modo, la ecuación (4) presenta una modificación a la especificación original del método, en la que se incluye el parámetro X_i , que corresponde a las dos covariables incluidas.

$$P(R_i = r | S_i = s, G_i = g) = \frac{P(G_i = g, X_i = x | R_i = r) P(R_i = r | S_i = s)}{\sum_{r' \in R} P(G_i = g, X_i = x | R_i = r') P(R_i = r' | S_i = s)} \quad (4)$$

2.2.2. Definición de la etnia predicha y validación de predicciones

Una vez se ha construido la matriz BISG para cada ejercicio y acorde con Imai y Khanna (2016), como criterio de selección se establece elegir como predicción a aquella etnia con la probabilidad más alta dentro de todos los grupos étnicos existentes, tal como se ilustra en el ejemplo presentado en el Anexo 4. Por su parte, la evaluación del método implementado consiste en la comparación de la información de la etnia autodeclarada en las fuentes de información con la etnia inferida por el BISG. El objetivo es analizar el desempeño de la predicción de la etnia a través de la generación de métricas de evaluación desprendidas de la matriz de confusión, tales como *accuracy*, ratio de falsos positivos (error tipo 1) y ratio de falsos negativos (error tipo 2).

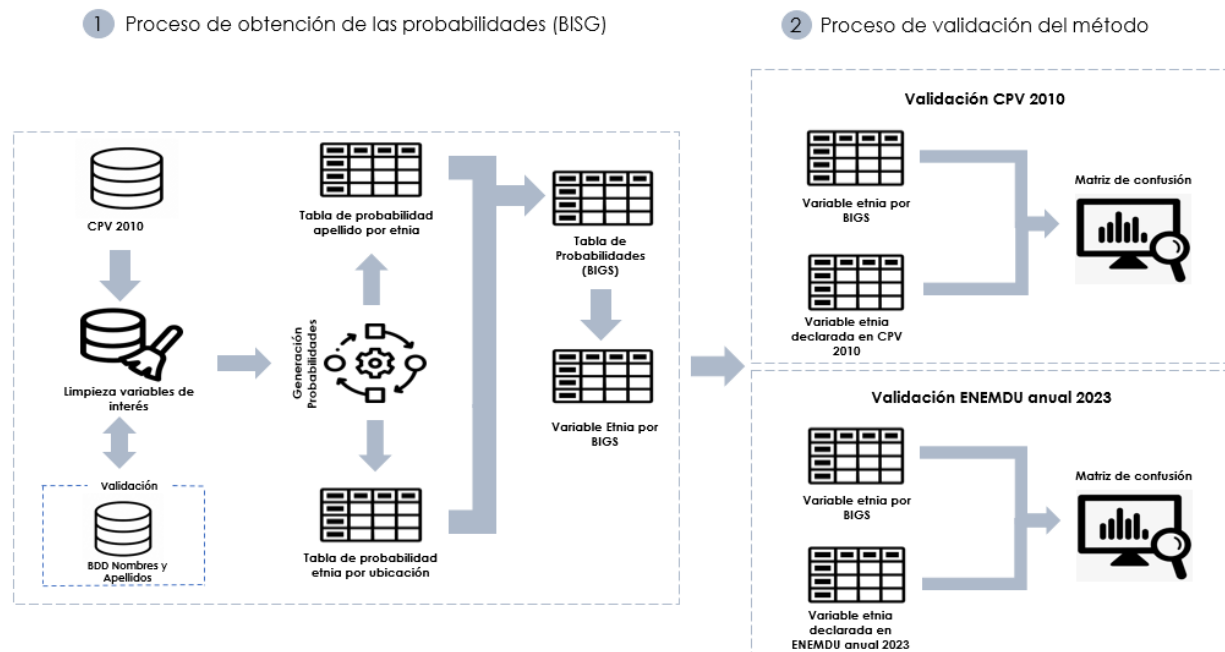
El ejercicio implementado en este trabajo partió de probar el método para los casos del primer apellido, segundo apellido, ambos apellidos agrupados¹¹ (Imai, et al., 2022) y la combinación de ambos apellidos¹² para el CPV 2010. En lo que respecta a la ubicación geográfica se probaron tres dimensiones geográficas de residencia para obtener resultados más precisos de residencia (provincia, cantón y parroquia), tal como ocurre en Clark et al. (2022).

Luego, se comparó el desempeño de la predicción de cada caso en base a métricas de evaluación. Como una aproximación adicional se incluyó en el mejor método de predicción las covariables de sexo y edad con el objetivo de apreciar una mejoría en el desempeño del mismo. Finalmente, el método de mejor desempeño fue probado en la base de datos de ENEMDU anual de 2023. La Figura 1 ilustra el proceso llevado a cabo en la implementación del método BISG.

¹¹ La tabla de probabilidad de apellido por etnia es generada a partir de la agrupación del primer apellido y el segundo apellido.

¹² Este método consiste en utilizar las matrices BISG por etnia para el primer apellido y el segundo apellido, para luego comparar las probabilidades y tomar la más alta entre ambos apellidos. El criterio considerado es el siguiente: i) si la probabilidad de la etnia del primer apellido es mayor a la probabilidad de la etnia del segundo apellido, se toma la etnia del primer apellido; ii) si la probabilidad de la etnia del segundo apellido es mayor a la probabilidad de la etnia del primer apellido, se toma la etnia del segundo apellido; y iii) si la probabilidad de la etnia del primer apellido es la misma que la del segundo apellido, se toma la etnia del primer apellido.

Figura 1. Proceso de implementación del método BISG



3. Resultados

Esta sección detalla los resultados obtenidos en términos de desempeño de la aplicación del método BISG sobre el CPV 2010 y consecuentemente como medio de evaluación de consistencia intertemporal sobre la ENEMDU anual 2023.

3.1. BISG para diferentes aproximaciones

A continuación, se compara el rendimiento de las predicciones étnicas en diferentes escenarios, que incluyen el uso del primer apellido, el segundo apellido, ambos apellidos agrupados y la combinación de ambos, considerando los niveles de desagregación geográfica provincial, cantonal y parroquial de residencia de los individuos en el CPV 2010. La Tabla 2 muestra una comparativa detallada de estos ejercicios en términos de métricas de evaluación como el *accuracy*, así como los errores de predicción, tales como el ratio de falsos positivos (error tipo 1) y el ratio de falsos negativos (error tipo 2).

Tabla 2. Comparación de ejercicios de implementación del método BISG

Provincia										
BISG	Indicador	General	Indígena	Afroecuatoriana (o)	Negra (o)	Mulata (o)	Montubia (o)	Mestiza (o)	Blanca (o)	Otra (o)
Primer_apellido	Accuracy	74,19%	79,16%	55,43%	57,02%	50,51%	57,43%	60,60%	51,18%	54,24%
Segundo_apellido		74,01%	79,24%	54,91%	56,29%	50,72%	57,91%	60,12%	50,71%	52,12%
Ambos_apellidos		80,01%	90,07%	67,33%	68,77%	63,21%	71,46%	72,83%	62,11%	64,98%
Combinacion_pre_d		74,96%	80,48%	54,13%	57,55%	50,50%	57,50%	60,83%	51,24%	54,44%
Primer_apellido		7,21%	2,94%	3,78%	0,85%	1,91%	6,41%	35,61%	5,89%	0,32%

Segundo_apellido	Ratio de Falsos positivos (error tipo 1)	7,42%	2,73%	3,73%	0,83%	1,85%	6,88%	37,22%	5,82%	0,32%
Ambos_apellidos		5,17%	1,38%	2,79%	0,62%	1,42%	4,29%	25,97%	4,63%	0,25%
Combinacion_pre d		6,71%	2,77%	3,89%	0,84%	1,91%	6,36%	31,70%	5,90%	0,32%
Primer_apellido	Ratio de Falsos negativos (error tipo 2)	47,29%	33,34%	62,46%	72,58%	65,14%	60,68%	23,13%	31,05%	29,95%
Segundo_apellido		48,10%	34,33%	62,26%	72,88%	68,80%	60,38%	23,20%	33,49%	29,47%
Ambos_apellidos		32,94%	19,84%	37,95%	51,70%	34,32%	48,91%	16,48%	22,46%	31,88%
Combinacion_pre d		42,02%	28,29%	55,33%	70,99%	56,62%	58,86%	23,09%	19,68%	23,30%
Cantón										
BISG	Indicador	Genera l	Indígen a	Afroecua toriana (o)	Negra (o)	Mulata (o)	Montubia (o)	Mestiza (o)	Blanca (o)	Otra (o)
Primer_apellido	Accuracy	75,65%	82,92%	56,95%	57,46%	50,76%	68,16%	64,65%	51,26%	54,36%
Segundo_apellido		75,61%	83,20%	56,33%	57,38%	50,92%	68,85%	64,49%	50,78%	52,14%
Ambos_apellidos		81,02%	91,56%	67,71%	69,38%	64,01%	77,61%	74,88%	62,15%	65,34%
Combinacion_pre d		76,29%	83,92%	55,81%	58,29%	50,73%	68,59%	64,91%	51,33%	54,46%
Primer_apellido	Ratio de Falsos positivos (error tipo 1)	6,64%	2,40%	3,65%	0,84%	1,90%	4,80%	33,35%	5,88%	0,32%
Segundo_apellido		6,74%	2,20%	3,61%	0,80%	1,85%	5,09%	34,27%	5,81%	0,32%
Ambos_apellidos		4,92%	1,17%	2,75%	0,61%	1,39%	3,32%	25,22%	4,62%	0,24%
Combinacion_pre d		6,32%	2,28%	3,75%	0,83%	1,91%	4,71%	30,87%	5,89%	0,32%
Primer_apellido	Ratio de Falsos negativos (error tipo 2)	46,43%	28,20%	61,51%	68,80%	73,41%	51,65%	21,00%	33,64%	33,24%
Segundo_apellido		46,18%	29,15%	61,78%	68,80%	72,34%	50,76%	20,92%	35,83%	29,85%
Ambos_apellidos		32,20%	17,70%	37,69%	49,48%	37,29%	45,27%	15,21%	21,97%	32,95%
Combinacion_pre d		41,67%	24,72%	58,28%	66,83%	67,08%	50,19%	20,95%	22,31%	23,01%
Parroquia										
BISG	Indicador	Genera l	Indígen a	Afroecua toriana (o)	Negra (o)	Mulata (o)	Montubia (o)	Mestiza (o)	Blanca (o)	Otra (o)
Primer_apellido	Accuracy	76,62%	85,43%	57,64%	57,66%	51,06%	69,53%	65,89%	51,34%	54,37%
Segundo_apellido		76,57%	85,57%	57,08%	57,40%	51,12%	70,21%	65,69%	50,83%	52,23%
Ambos_apellidos		81,53%	92,45%	68,13%	69,51%	64,30%	78,16%	75,47%	62,64%	65,81%
Combinacion_pre d		77,13%	85,95%	56,69%	58,03%	51,06%	69,90%	65,99%	51,39%	54,54%
Primer_apellido	Ratio de Falsos positivos (error tipo 1)	6,22%	2,06%	3,59%	0,84%	1,89%	4,59%	30,57%	5,87%	0,32%
Segundo_apellido		6,32%	1,90%	3,55%	0,80%	1,84%	4,87%	31,48%	5,81%	0,31%
Ambos_apellidos		4,75%	1,05%	2,72%	0,60%	1,38%	3,24%	24,17%	4,56%	0,24%
Combinacion_pre d		5,95%	2,00%	3,68%	0,83%	1,89%	4,51%	28,50%	5,89%	0,32%
Primer_apellido	Ratio de Falsos negativos (error tipo 2)	44,82%	22,52%	59,70%	67,85%	71,51%	49,45%	20,40%	35,23%	31,91%
Segundo_apellido		46,52%	23,41%	59,79%	67,66%	70,96%	48,57%	20,35%	37,20%	44,26%
Ambos_apellidos		32,09%	14,95%	36,97%	48,83%	37,78%	43,95%	14,90%	23,96%	35,36%
Combinacion_pre d		40,48%	20,17%	55,89%	65,84%	66,36%	48,10%	20,42%	23,90%	23,13%

Los resultados obtenidos al implementar el método en diferentes dimensiones de análisis revelaron que el mejor rendimiento se encontró en el BISG utilizando ambos apellidos agrupados a nivel parroquial. Este hallazgo se replicó tanto de manera global como para todas las categorías étnicas.

En términos generales, el método mostró un accuracy del 81,53%, aproximadamente 1,5 puntos porcentuales por encima de los otros métodos probados. Además, en cuanto a los errores tipo 1 y tipo 2, se registraron cifras alrededor del 4,75% y 32,09%, respectivamente, considerablemente

más bajos que en los otros ejercicios realizados. Aunque la predicción general y por categoría mostraron los mejores resultados entre los ejercicios probados, el análisis detallado reveló que las categorías étnicas con los mayores niveles de precisión son indígena, montubia (o) y mestiza (o), con cifras de exactitud del 92,45%, 78,47% y 75,47%, respectivamente. Esto sugiere que el método clasifica de manera más precisa a estos grupos étnicos.

Sin embargo, se observaron altos errores tipo 1 para la categoría mestizo en comparación con otras categorías, alcanzando el 24,17%, lo que evidenció la posibilidad de clasificar erróneamente a individuos de otras categorías en esta. Además, en cuanto a los errores tipo 2, se detectaron cifras superiores al 40% para las etnias negra (o) y montubia (o), lo que sugiere una subestimación de estas categorías, con la posibilidad de que el método clasifique erróneamente a individuos pertenecientes a estas etnias en otra categoría.

Como medida adicional para fortalecer el análisis, se incorporaron covariables de sexo y edad al mejor método (ambos apellidos agrupados a nivel parroquial) con el fin de mejorar el accuracy y reducir los errores en las predicciones. No obstante, como se muestra en la Tabla 3, el cambio en las métricas obtenidas es marginal. A pesar de ello, esta aproximación se posicionó como la mejor en comparación con los otros ejercicios realizados.

Tabla 3. Desempeño del método por dos apellidos agregados y por parroquia, sexo y edad

BISG	Indicador	General	Indígena	Afroecuatoriana (o)	Negra (o)	Mulata (o)	Montubia (o)	Mestiza (o)	Blanca (o)	Otra (o)
Ambos_a pellidos (Sexo - Edad)	Accuracy	81,54%	92,54%	68,19%	69,68%	64,61%	78,43%	75,61%	62,81%	66,31%
	Ratio de Falsos positivos (error tipo 1)	4,76%	1,03%	2,71%	0,60%	1,37%	3,19%	24,37%	4,54%	0,24%
	Ratio de Falsos negativos (error tipo 2)	32,31%	14,92%	36,78%	48,69%	38,57%	44,14%	14,79%	24,63%	35,95%

3.2. Validación del método

Con el objetivo de verificar la consistencia temporal del método final, se aplicó el método por dos apellidos agregados y por parroquia, edad y sexo sobre base de datos de la ENEMDU del 2023. La Tabla 4 muestra la distribución de la población a través de las diferentes etnias de la encuesta antes y después de la validación de los apellidos de los individuos.

Tabla 4. Descripción de la variable de etnia en ENEMDU 2023

Fuentes de información	Descripción étnica	Número de registros	Registros depurados ambos apellidos agrupados por sexo, edad y parroquia *
	Indígena	22.217 (7,29%)	10.509 (8,18%)
	Afroecuatoriana (o)	4.428 (1,45%)	1.608 (1,25%)

ENEMDU anual 2023	Negra (o)	5.131 (1,68%)	1.840 (1,43%)
	Mulata (o)	2.941 (0,97%)	1.041 (0,81%)
	Montubia (o)	4.551 (1,49%)	2.365 (1,84%)
	Mestiza (o)	262.238 (86,08%)	109.958 (85,55%)
	Blanca (o)	3.035 (1,00%)	1.172 (0,91%)
	Otra (o)	96 (0,03%)	37 (0,03%)

Nota:

*Los registros considerados en la predicción son aquellos que se emparejan de acuerdo al primer apellido, segundo apellido, sexo, edad y ubicación de residencia (parroquia) con el CPV 2010.

Los resultados alcanzados se presentan en la Tabla 5, donde se puede ver que de forma general el *accuracy* crece respecto al ejercicio implementado, bordeando de forma general el 87,91%. En tanto que, los errores tipo 1 y tipo 2 aumentan, lo cual se atribuye al tamaño de los grupos poblacionales que conforman la ENEMDU 2023, que en su mayoría son mestizos e indígenas.

Al analizar las métricas de evaluación por categorías, se observa una ligera disminución en el *accuracy* para todas las categorías, excepto para indígena y afroecuatoriana (o), que aumentan en aproximadamente 2 puntos porcentuales. Los errores tipo 1 se mantienen estables, excepto para la categoría mulata (o), que presenta cifras más altas. Por otro lado, el error tipo 2 aumenta en todas las categorías, excepto en la categoría mulata (o), donde se reduce.

Tabla 5. Desempeño del método BISG por dos apellidos agregados para el nivel geográfico de parroquia por sexo y edad para la ENEMDU anual 2023

BISG	Indicador	General	Indígena	Afroecuatoriana (o)	Negra (o)	Mulata (o)	Montubia (o)	Mestiza (o)	Blanca (o)	Otra (o)
Ambos, a apellidos (Sexo - Edad)	Accuracy	87,91%	93,38%	73,55%	67,13%	56,45%	69,79%	83,16%	53,20%	64,75%
	Ratio de Falsos positivos (error tipo 1)	5,57%	0,65%	0,85%	1,03%	39,23%	1,10%	0,71%	0,94%	0,02%
	Ratio de Falsos negativos (error tipo 2)	62,98%	74,89%	94,43%	17,75%	4,49%	75,30%	85,22%	55,42%	96,35%

3.3. Agregación de categorías étnicas

Con el objetivo de abordar el limitado tamaño de ciertos grupos poblacionales, se exploró una aproximación alternativa. Esta consistió en aplicar el mejor método identificado (que utiliza ambos apellidos agrupados por parroquia, edad y sexo), pero con la diferencia de que esta vez se agruparon en una sola categoría las etnias mulata (o), negra (o) y Afroecuatoriana (o). Al

reducir las categorías, el método mostró mejoras tanto en términos generales como por categorías, como se puede observar en la Tabla 6.

El desempeño del *accuracy* general pasó del 81,54% al 82,17%. En el caso de los errores tipo 1 el comportamiento fue diferente ya que aumentaron, pasando de 4,76% a 6,30%. Mientras que, los errores tipo 2 se redujeron de 32,31% a 26,84%. Al analizar las categorías, se replicó el patrón general de mejoras en el *accuracy* de la predicción. Especialmente notable fue la mejora en la categoría que agrupa las etnias mulata (o), negra (o) y afroecuatoriana (o), cuyo *accuracy* aumentó del 67,47% en promedio de las tres categorías al 72,97% al fusionar las categorías. Sin embargo, los errores tipo 1 y tipo 2 se mantuvieron relativamente estables para todas las categorías.

Tabla 6. Desempeño del método para dos apellidos agregados por parroquia, sexo y edad para etnias agregadas para el CPV 2010

BISG	Indicador	General	Indígena	Afro/ Negra(o)/ Mulata(o)	Montubia(o)	Mestiza(o)	Blanca(o)	Otra(o)
Ambos apellidos	Accuracy	82,17%	92,54%	72,97%	78,42%	75,87%	62,82%	66,32%
	Ratio de Falsos positivos (error tipo 1)	6,30%	1,03%	3,95%	3,20%	24,84%	4,54%	0,24%
	Ratio de Falsos negativos (error tipo 2)	26,84%	14,92%	26,90%	44,13%	14,59%	24,64%	35,85%

Al igual que en el ejercicio anterior, se realizó la predicción de la etnia para todas las categorías étnicas utilizando la base de datos de la ENEMDU anual 2023. Los resultados, al igual que con la información del CPV 2010, mostraron mejoras significativas, como se detalla en la Tabla 7. En términos generales, el *accuracy* de la predicción aumentó del 82,17% en el CPV 2010 al 89,09% en la ENEMDU anual 2023. A pesar de estas mejoras sustanciales en el *accuracy*, los errores tipo 1 y tipo 2 también aumentaron, pasando del 6,30% al 7,28% y del 26,84% al 55,27%, respectivamente. Por categorías, se observa un aumento en el *accuracy* de todas las categorías étnicas, excepto para las etnias blanca (o) y otra (o), que disminuyeron. En cuanto a los errores tipo 1, se observa una mejora para todas las categorías, excepto para mestiza (o), donde el error se incrementó. Por otro lado, los errores tipo 2 disminuyeron para las categorías mestiza (o) y montubia (o).

Tabla 7. Desempeño del método para dos apellidos agregados por parroquia, sexo y edad para etnias agregadas para la ENEMDU 2023

BISG	Indicador	General	Indígena	Afro/ Negra(o)/ Mulata(o)	Montubia(o)	Mestiza(o)	Blanca(o)	Otra(o)
Ambos apellidos	Accuracy	89,09%	93,38%	86,95%	69,80%	83,47%	53,20%	64,75%
	Ratio de Falsos positivos (error tipo 1)	7,28%	0,88%	0,85%	1,03%	39,81%	1,10%	0,02%

Ratio de Falsos negativos (error tipo 2)	55,27%	43,55%	94,42%	17,74%	4,35%	75,25%	96,35%
--	--------	--------	--------	--------	-------	--------	--------

4. Conclusiones

La información de la etnia de los individuos juega un rol importante en diferentes aristas socioeconómicas como una fuente de información necesaria para la reducción de la desigualdad y la elaboración de políticas públicas eficientes que garanticen la equidad de todos los grupos sociales nacionales (Galarza, 2010). Así también, la información étnica se ha convertido en un medio necesario en el campo de la salud, donde a través de la información se pueden obtener diagnósticos médicos y tratamientos específicos para grupos poblacionales específicos, logrando una atención eficiente de los pacientes (Adjaye-Gbewonyo et al., 2014; Mays et al., 2003). En el ámbito político y electoral, la composición étnica es determinante en el análisis de la participación electoral de los individuos y de su intención de voto (Barreto et al., 2022; Imai y Khana, 2016).

Dada la importancia de la información étnica, varios son los estudios desarrollados para poder subsanar su ausencia en fuentes de información, dentro de los métodos implementados se puede destacar aquellos que predicen la etnia a través del aprovechamiento de la información de la ubicación geográfica de residencia y los apellidos de los individuos como son: *Categorical Surname and Geocoding approach* (CSG) (Fiscella y Fremont, 2006), el enfoque *Bayesian Surname Geocoding* (BSG) (Elliot et al., 2008) y *Bayesian Improved Surname Geocoding* (BISG) (Elliott, et al., 2009).

En ese sentido, este estudio se propuso como finalidad probar el método *Bayesian Improved Surname Geocoding* (BISG) desarrollado por Elliot et al. (2009) para la predicción de la identificación étnica de los individuos en el Ecuador. El método se basa en la aplicación de estimaciones bayesianas, en las que considerando el apellido y la ubicación geográfica de residencia de los individuos se predice su etnia.

Para este estudio, se optó por inicialmente probar el método BISG establecido por Elliot et al. (2009) utilizando los datos del CPV 2010. En este proceso, se llevó a cabo la predicción de la etnia de los individuos basada en el primer apellido y la ubicación geográfica de su residencia. Sin embargo, con el propósito de fortalecer el análisis, se realizaron pruebas adicionales que incluyeron el uso del segundo apellido, la agrupación de ambos apellidos y la combinación de estos. Además, se evaluaron los resultados a través de tres niveles geográficos de la residencia: provincias, cantones y parroquias. Cada análisis se sometió a un pre-procesamiento específico de los apellidos del CPV 2010, utilizando la base de datos del visualizador de nombres y apellidos del INEC de 2022 para definir los casos a analizar.

Una vez realizada la predicción de la etnia, se compararon los resultados con la información autodeclarada por los individuos en el CPV 2010. A través de métricas derivadas de una matriz de confusión y se evaluó el rendimiento del método. Los resultados mostraron que el mejor rendimiento del método, en términos de exactitud (*accuracy*) y menor cantidad de errores de predicción, tanto de tipo 1 como de tipo 2, se obtuvo al agrupar ambos apellidos y considerar el nivel parroquial de desagregación geográfica de la residencia.

Para complementar el análisis, y con el objetivo de mejorar el rendimiento del método identificado, siguiendo la metodología de Imai y Khana (2016), se incluyeron covariables como la edad y el sexo. Los resultados experimentaron fluctuaciones leves, con un ligero aumento en el *accuracy* del método. Por lo tanto, se estableció como modelo definitivo de predicción aquel

que utiliza la agrupación de los dos apellidos, considera la dimensión geográfica parroquial, y además incorpora la información de sexo y edad.

Una vez identificado el mejor método de predicción de la etnia en el CPV 2010, se procedió a probarlo en la ENEMDU anual de 2023. Los resultados mostraron un mejor desempeño del método en términos de *accuracy*, lo que sugirió que el método tiende a ofrecer resultados consistentes a lo largo del tiempo. Sin embargo, al analizar los resultados por categorías étnicas, se observó que los grupos poblacionales más representados presentaban un *accuracy* más alto. Específicamente, se notó un alto error tipo 1 en la etnia mestiza (o), atribuible a la alta concentración poblacional de este grupo, lo que puede llevar a la clasificación errónea de individuos de otras categorías. Además, se identificó que, para todas las categorías excepto mestiza (o) y montubia (o), el error tipo 2 fue notablemente alto, indicando la posibilidad de clasificar erróneamente a individuos de estas etnias en otras categorías debido a su tamaño poblacional limitado.

En conclusión, el método BISG para dos apellidos agregados, por parroquia, sexo y edad demostró ser una herramienta útil para predecir la etnia en casos donde la información es escasa. No obstante, presentó limitaciones, como altos valores de error tipo 1 y tipo 2, especialmente en grupos poblacionales pequeños.

Con miras a mejorar el desempeño del método, el ejercicio demostró que la calidad de la información de los apellidos y el tamaño de la muestra pueden influir en la precisión de las predicciones. Por lo tanto, es recomendable realizar pruebas del método en otras fuentes de información para validar la consistencia del método y su funcionalidad. Adicionalmente, considerar la inclusión de otras variables como la lengua hablada o de los padres, e incluso explorar ubicaciones geográficas más detalladas como zonas o sectores, podría mejorar la predicción étnica. Asimismo, comparar los resultados del método BISG con otras técnicas de predicción, como modelos de clasificación basados en machine learning, permitiría identificar posibles mejoras y limitaciones del enfoque utilizado, tal como se destaca en el estudio de Decter-Frain (2022).

Bibliografía

- Adjaye-Gbewonyo, D., Bednarczyk, R., Davis, L., & Omer, S. (2014). Using the Bayesian Improved Surname Geocoding Method (BISG) to Create a Working Classification of Race and Ethnicity in a Diverse Managed Care Population: A Validation Study. *Health Services Research*, 268 - 283.
- Barreto, M., Cohen, M., Collingwood, L., Dunn, C., & Waknin, S. (2022). A NOVEL METHOD FOR SHOWING RACIALLY POLARIZED VOTING: BAYESIAN IMPROVED SURNAME GEOCODING. *N.Y.U. REVIEW OF LAW & SOCIAL CHANGE*, 1 - 42.
- Clark, J., Curiel, J., & Steelman, T. (2022). Minmaxing of Bayesian Improved Surname Geocoding and Geography Level Ups in Predicting Race. *Politycal Analysis*, 456 - 462.
- Consejo Nacional para la Igualdad de Pueblos y Nacionalidades. (2021). Agenda para la igualdad de derechos de las nacionalidades y pueblos indígenas , pueblo afroecuatoriano y pueblo montubio 2019 - 2021.

- Decter - Frain, A. (2022). HOW SHOULD WE PROXY FOR RACE/ETHNICITY? COMPARIN BAYESIAN IMPROVED SURNAME GEOCODING TO MACHINE LEARNING METHODS. *arXiv*, 1-15.
- Elliot, M., Morrison, P., Fremont, A., McCaffrey, A., Pantoja, P., & Lurie, N. (2009). Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Serv Outcomes Res Method*, 69:83.
- Elliott, Fremont, Morrison, Pantoja, & Lurie. (2008). A new method for estimating race/ethnicity and associated disparities where administrative records lack self-reported race/ethnicity. *Health Serv. Res.* 43(5p1), 722–1736.
- Fiscella, K., & Fremont, A. (2006). Use of geocoding and surname analysis to estimate race and ethnicity. *Health Serv. Res.* 41(4 Pt 1), 1482–1500.
- Galarza, P. (2010). *Inclusión de la variable etnia en las fuentes de información sociodemográfica del Ecuador*. Santiago de Chile: Naciones Unidas.
- Hasan, Ahraful, Shidhartho, Aishwariya, Tasnim, & Sunanda. (2021). Missing value imputation affects the performance of machine learning: A review and analysis of the literature (2010–2021). *Informatics in Medicine Unlocked* 27, 1-23.
- Imai, K., & Khanna, K. (2016). Improving Ecological Inference by Predicting Individual Ethnicity from Voter Registration Records. *Political Analysis*, 263-272.
- Imai, K., Olivella, S., & Rosenman, E. (2022). Addressing Census data problems in race imputation via fully Bayesian Improved Surname Geocoding and name supplements. *arXiv*, 1-20.
- Instituto Nacional de Estadística y Censos (INEC). (2021). *Diseño muestral de la Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU)*.
- Mays, V., Ponce, N., Washington, D., & Cochran, S. (2003). CLASSIFICATION OF RACE AND ETHNICITY: Implications for Public Health. *Annu. Rev. Public Health* , 83–110.
- Mosquera, M., & Ayala, L. (2020). La interculturalidad y la plurinacionalidad del Ecuador en el marco de los derechos del Buen Vivir. *Revista digital de Ciencia, Tecnología e Innovación*, 988-998.

Anexos

Anexo 1: Ranking de primeros apellidos por etnia a partir del CPV 2010

Nombres	Ranking*	Cuenta**	prop100k	acum_prop100k	Prc_Indígena	Prc_Afroecuatoriana(o)	Prc_Negra(o)	Prc_Montubia(o)	Prc_Mestiza(o)	Prc_Blanca(o)	Prc_Otra(o)	Prc_Mulata(o)
ZAMBRANO	1	176.597	1.251,82	1.251,82	0,30%	4,58%	0,63%	16,18%	68,60%	7,56%	0,25%	1,90%
SANCHEZ	2	134.504	953,44	2.205,25	1,80%	3,48%	0,61%	10,63%	75,06%	6,41%	0,27%	1,75%
VERA	3	125.662	890,76	3.096,01	0,52%	4,77%	0,62%	17,14%	67,30%	6,85%	0,39%	2,43%
GARCIA	4	124.791	884,59	3.980,60	1,14%	4,89%	1,14%	11,31%	70,84%	7,76%	0,33%	2,59%
RODRIGUEZ	5	120.637	855,14	4.835,74	0,59%	5,49%	1,43%	9,45%	73,38%	6,78%	0,59%	2,30%
LOPEZ	6	108.434	768,64	5.604,38	1,89%	3,13%	0,35%	6,87%	78,55%	7,35%	0,32%	1,54%
CEDENO	7	100.373	711,50	6.315,88	0,11%	4,50%	0,60%	17,68%	67,26%	7,77%	0,24%	1,84%
TORRES	8	85.592	606,72	6.922,60	1,23%	4,20%	1,05%	6,70%	77,67%	6,70%	0,31%	2,13%
CASTRO	9	82.732	586,45	7.509,05	2,05%	4,45%	0,80%	14,94%	68,88%	6,45%	0,30%	2,12%
MENDOZA	10	76.989	545,74	8.054,79	1,03%	4,05%	0,49%	15,38%	68,99%	7,89%	0,25%	1,92%
MACIAS	11	73.839	523,41	8.578,20	0,10%	4,24%	0,56%	19,94%	65,32%	7,82%	0,27%	1,76%
ESPINOZA	12	67.951	481,67	9.059,87	1,51%	6,36%	1,60%	11,62%	68,72%	7,46%	0,24%	2,49%
FLORES	13	67.298	477,04	9.536,92	2,46%	3,14%	0,40%	3,60%	82,04%	6,47%	0,33%	1,57%
GUAMAN	14	65.933	467,37	10.004,29	26,38%	1,50%	0,08%	0,91%	68,04%	2,08%	0,17%	0,84%
DELGADO	15	65.792	466,37	10.470,66	1,50%	6,61%	1,33%	9,10%	72,20%	6,72%	0,40%	2,14%
ALVARADO	16	64.354	456,18	10.926,83	14,23%	3,11%	0,41%	15,30%	58,57%	6,36%	0,30%	1,71%
LOOR	17	64.166	454,84	11.381,68	0,12%	4,55%	0,48%	16,49%	68,02%	8,22%	0,27%	1,85%
ROMERO	18	64.066	454,13	11.835,81	1,77%	3,06%	0,35%	8,31%	77,28%	7,52%	0,25%	1,46%
LEON	19	62.432	442,55	12.278,36	2,60%	3,87%	0,59%	11,74%	71,95%	6,83%	0,41%	2,02%
BRAVO	20	62.080	440,06	12.718,42	1,69%	4,47%	0,79%	12,60%	71,39%	6,95%	0,22%	1,88%
MOREIRA	21	61.226	434,00	13.152,42	0,09%	4,24%	0,43%	16,45%	68,70%	8,05%	0,23%	1,81%
REYES	22	61.150	433,46	13.585,89	0,28%	5,24%	0,71%	8,79%	76,45%	5,72%	0,66%	2,14%
MUNOZ	23	61.091	433,05	14.018,93	1,91%	3,58%	0,43%	13,05%	71,16%	7,31%	0,49%	2,06%
MORALES	24	60.689	430,20	14.449,13	6,68%	3,97%	0,97%	5,65%	74,92%	5,57%	0,28%	1,96%
JIMENEZ	25	59.930	424,82	14.873,95	0,87%	2,92%	0,45%	11,46%	76,73%	5,83%	0,22%	1,54%

PEREZ	26	59.503	421,79	15.295,74	2,29%	2,72%	0,39%	6,84%	78,96%	6,82%	0,35%	1,63%
CASTILLO	27	59.190	419,57	15.715,31	1,30%	9,41%	3,82%	3,39%	72,88%	5,83%	0,19%	3,19%
ORTIZ	28	57.913	410,52	16.125,83	3,50%	7,62%	2,86%	4,84%	71,59%	5,52%	0,28%	3,78%
VARGAS	29	57.435	407,13	16.532,96	11,87%	2,86%	0,33%	9,92%	67,71%	5,58%	0,22%	1,52%
SALAZAR	30	56.595	401,18	16.934,13	4,89%	3,88%	0,82%	8,18%	73,40%	6,70%	0,24%	1,89%

Nota:

* Existe un total de 467.557 individuos con apellidos con una frecuencia menor a 100 catalogados con ranking 0.

** El total de apellidos analizados corresponde a 14.107.270

Anexo 2: Matriz de probabilidad de pertenencia de un individuo a una etnia por apellido

Apellido paterno	Indígena	Afroecuatoriana(o)	Negra(o)	Mulata(o)	Montubia(o)	Mestiza(o)	Blanca(o)	Otra(o)
AAGUILAR	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%	0,00%	0,00%
AAMODT	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%	0,00%	0,00%
AANDSTAD	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%	0,00%
AAPIK	92,31%	0,00%	0,00%	0,00%	0,00%	7,69%	0,00%	0,00%
AARIAS	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%	0,00%	0,00%
AARON	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%	0,00%	0,00%
ABA	3,57%	10,71%	0,00%	3,57%	3,57%	75,00%	3,57%	0,00%
ABAAD	0,00%	11,36%	0,00%	0,00%	4,55%	81,82%	2,27%	0,00%
ABAB	0,00%	5,26%	0,00%	7,89%	7,89%	78,95%	0,00%	0,00%
ABACA	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%	0,00%	0,00%
ABACO	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%	0,00%	0,00%
ABAD	0,18%	2,82%	0,48%	1,37%	4,98%	82,25%	7,62%	0,30%
ABADA	0,00%	0,00%	0,00%	0,00%	0,00%	90,91%	9,09%	0,00%
ABADEANO	0,00%	7,69%	0,00%	0,00%	0,00%	92,31%	0,00%	0,00%
ABADIA	0,00%	0,00%	0,00%	0,00%	8,33%	66,67%	25,00%	0,00%
ABADIANO	2,20%	0,00%	0,00%	0,00%	0,00%	96,70%	1,10%	0,00%
ABADIAS	0,00%	0,00%	100,00%	0,00%	0,00%	0,00%	0,00%	0,00%
ABADIE	0,00%	1,56%	0,00%	0,00%	0,00%	82,81%	15,63%	0,00%

ABADIES	0,00%	0,00%	0,00%	0,00%	0,00%	50,00%	50,00%	0,00%
ABAHONZA	0,00%	0,00%	0,00%	0,00%	0,00%	83,33%	16,67%	0,00%
ABAL	0,00%	11,11%	0,00%	5,56%	5,56%	72,22%	5,56%	0,00%
ABALCO	15,15%	0,61%	0,00%	0,00%	0,30%	81,82%	2,12%	0,00%
ABALDE	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%	0,00%	0,00%
ABALLES	0,00%	100,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
ABALO	0,00%	0,00%	0,00%	0,00%	0,00%	95,24%	4,76%	0,00%
ABALOS	8,61%	1,32%	0,66%	0,00%	0,00%	82,12%	7,28%	0,00%
ABAMBARA	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%	0,00%	0,00%
ABAMBARI	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%	0,00%	0,00%
ABAMBARY	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%	0,00%	0,00%

Anexo 3: Matriz de probabilidad de pertenencia de un individuo a una ubicación por etnia

Parroquia	Indígena	Afroecuatoriana(o)	Negra(o)	Mulata(o)	Montubia(o)	Mestiza(o)	Blanca(o)	Otra(o)
010150	0,37%	1,01%	0,34%	0,81%	0,14%	2,82%	2,76%	1,49%
010151	0,03%	0,02%	0,01%	0,01%	0,00%	0,15%	0,07%	0,04%
010152	0,01%	0,02%	0,00%	0,00%	0,00%	0,05%	0,01%	0,01%
010153	0,00%	0,00%	0,00%	0,00%	0,00%	0,01%	0,00%	0,00%
010154	0,01%	0,01%	0,00%	0,00%	0,00%	0,02%	0,02%	0,01%
010155	0,00%	0,01%	0,00%	0,01%	0,00%	0,04%	0,01%	0,00%
010156	0,00%	0,00%	0,00%	0,01%	0,00%	0,05%	0,01%	0,00%
010157	0,00%	0,02%	0,00%	0,03%	0,01%	0,06%	0,01%	0,00%
010158	0,00%	0,00%	0,00%	0,01%	0,00%	0,04%	0,01%	0,02%
010159	0,01%	0,00%	0,00%	0,00%	0,00%	0,02%	0,00%	0,00%
010160	0,00%	0,01%	0,00%	0,00%	0,00%	0,06%	0,02%	0,00%
010161	0,14%	0,00%	0,00%	0,00%	0,00%	0,06%	0,02%	0,09%
010162	0,02%	0,05%	0,01%	0,03%	0,01%	0,18%	0,06%	0,04%

010163	0,01%	0,01%	0,00%	0,01%	0,00%	0,07%	0,03%	0,02%
010164	0,04%	0,01%	0,00%	0,01%	0,00%	0,05%	0,02%	0,01%
010165	0,01%	0,03%	0,01%	0,01%	0,00%	0,08%	0,03%	0,02%
010166	0,00%	0,00%	0,00%	0,00%	0,00%	0,04%	0,01%	0,00%
010167	0,01%	0,01%	0,00%	0,02%	0,01%	0,15%	0,03%	0,03%
010168	0,12%	0,01%	0,00%	0,00%	0,00%	0,09%	0,02%	0,03%
010169	0,07%	0,02%	0,00%	0,02%	0,00%	0,08%	0,03%	0,02%
010170	0,03%	0,03%	0,00%	0,02%	0,01%	0,23%	0,06%	0,02%
010171	0,03%	0,00%	0,00%	0,00%	0,00%	0,05%	0,01%	0,00%
010250	0,01%	0,02%	0,00%	0,01%	0,00%	0,08%	0,03%	0,03%
010251	0,00%	0,00%	0,00%	0,01%	0,00%	0,03%	0,02%	0,00%
010252	0,00%	0,00%	0,00%	0,00%	0,00%	0,01%	0,00%	0,00%
010350	0,04%	0,09%	0,01%	0,05%	0,01%	0,18%	0,13%	0,10%
010352	0,00%	0,00%	0,00%	0,00%	0,00%	0,02%	0,00%	0,00%
010353	0,04%	0,00%	0,00%	0,00%	0,00%	0,04%	0,01%	0,01%
010354	0,00%	0,00%	0,00%	0,00%	0,00%	0,02%	0,01%	0,00%
010356	0,00%	0,00%	0,00%	0,00%	0,00%	0,01%	0,01%	0,00%

Anexo 4: Matriz BISG

Apellido paterno	Parroquia	Indígena	Afroecuatoriana(o)	Negra(o)	Mulata(o)	Montubia(o)	Mestiza(o)	Blanca(o)	Otra(o)
CHALCO	010150	0,79%	0,39%	0,01%	0,19%	0,03%	96,36%	2,10%	0,14%
FLORES	010150	0,36%	1,24%	0,05%	0,50%	0,20%	90,47%	6,99%	0,19%
FAICAN	010150	0,06%	0,54%	0,00%	0,05%	0,02%	96,88%	2,44%	0,00%
GUTAMA	010150	0,07%	0,97%	0,02%	0,45%	0,07%	96,72%	1,68%	0,02%
BALLADARES	010150	0,12%	1,13%	0,03%	0,36%	0,22%	88,97%	8,96%	0,21%
VALLADARES	010150	0,11%	0,56%	0,01%	0,49%	0,06%	91,19%	7,18%	0,39%

GALARZA	010150	0,72%	1,17%	0,04%	0,49%	0,40%	89,17%	7,85%	0,16%
JARAMILLO	010150	0,05%	1,21%	0,11%	0,46%	0,17%	88,29%	9,51%	0,21%
CARPIO	010150	0,20%	1,35%	0,05%	0,65%	1,04%	88,25%	8,30%	0,16%
MENDOZA	010150	0,17%	1,82%	0,07%	0,69%	0,97%	86,43%	9,68%	0,17%
VEINTIMILLA	010150	0,25%	0,90%	0,05%	0,52%	0,06%	89,70%	8,40%	0,11%
OCHOA	010150	0,26%	0,95%	0,03%	0,45%	0,40%	90,05%	7,72%	0,15%
ENCALADA	010150	0,19%	0,72%	0,02%	0,22%	0,05%	91,16%	7,52%	0,13%
JARA	010150	0,10%	0,81%	0,03%	0,37%	0,18%	90,39%	8,00%	0,12%
CONTRERAS	010150	0,03%	1,75%	0,13%	0,82%	1,35%	88,50%	7,27%	0,15%
DOMINGUES	010150	0,29%	2,43%	0,07%	0,72%	0,26%	88,48%	7,42%	0,33%
MUNOZ	010150	0,31%	1,58%	0,06%	0,73%	0,81%	87,40%	8,79%	0,32%
SENTENA	010150	0,00%	0,00%	0,79%	0,00%	0,99%	78,89%	19,33%	0,00%
BARBECHO	010150	0,08%	0,33%	0,00%	0,18%	0,04%	97,70%	1,68%	0,00%
LUCERO	010150	0,54%	0,58%	0,02%	0,40%	0,04%	94,29%	4,04%	0,09%
DOMINGUEZ	010150	0,22%	1,92%	0,06%	0,59%	0,43%	89,74%	6,39%	0,64%
GUEBARA	010150	0,04%	1,09%	0,29%	0,26%	0,12%	88,20%	9,84%	0,16%
ANDRADE	010150	0,33%	1,10%	0,06%	0,54%	0,31%	87,81%	9,72%	0,13%
PAQUI	010150	22,80%	0,50%	0,00%	0,00%	0,01%	74,69%	1,38%	0,62%
CARVAJAL	010150	0,23%	1,46%	0,11%	0,98%	0,21%	89,28%	7,48%	0,23%
MEDINA	010150	1,07%	3,67%	0,39%	0,72%	0,45%	86,96%	6,54%	0,21%
QUINDE	010150	0,42%	1,74%	0,04%	0,65%	0,32%	90,40%	5,94%	0,50%
URGILES	010150	0,09%	0,56%	0,01%	0,14%	0,07%	93,00%	6,01%	0,11%
SARI	010150	0,20%	0,46%	0,00%	0,07%	0,02%	97,38%	1,82%	0,06%
ROMERO	010150	0,27%	1,26%	0,05%	0,48%	0,48%	88,84%	8,47%	0,15%

Nota:

La etnia a la que corresponde cada apellido paterno en determinada parroquia es la probabilidad más alta (**texto en negrita**).



@ecuadorencifras



@ecuadorencifras



@InecEcuador



t.me/equadorencifras



INEC/Ecuador



INECEcuador

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

www.ecuadorencifras.gob.ec