README

David Ricardo Gonzales

Marzo 13 del 2023

La base de datos trabajada representa un conjunto de apellidos en Perú (40 mil aproximadamente). Se clasificó el origen del apellido en indígena e (ii) hispano.

Se realizó una base datos adicional, esta contenía aproximadamente 75 mil observaciones (apellidos) de origen hispanos[1] y aproximadamente 2 mil observaciones (apellidos) de origen indígena[2]. El objetivo de esta nueva base de datos era realizar un *merge* con la base de datos original.

El principal problema que encontramos fue que no existían muchas fuentes para apellidos indígenas. Por ende, teníamos tan pocas observaciones y, al realizar el *merge* con la base de datos original, solo 520 apellidos indígenas fueron reconocidos. Sin embargo, como nuestra base de datos de apellidos hispanos era más amplia, se reconocieron 9,801 apellidos.

Para evitar el problema de falta de fuentes, hemos considerado todas las palabras de diccionarios[3]indígenas, precisamente quechua y aymara. Una vez que añadimos diccionarios a nuestras fuentes, teníamos 72167 observaciones indígenas.

Una vez realizada la compilación de los apellido hispanos e indígenas de diversas fuentes, formamos la base de datos presentada en formato excel. Para poder examinar el grueso de apellidos recopilados, el código presentado elabora la base de datos *apellidos_num_fuentes.dta*, donde se detallan todos los apellidos recopilados, cuáles únicamente se encuentran en fuentes hispanas, únicamente en fuentes indígenas, cuáles se encuentran en ambos tipos de fuentes y la cantidad de ocasiones en las que cada apellido se repite entre las fuentes indígenas e hispanas. Un resumen de estos resultados se puede encontrar en el anexo.

Una complicación que encontramos en los apellidos indígenas es que son muy variables, suele cambiar una palabra, agregar -o quitar- un apóstrofe, entre otros problemas. Este tipo de casos también ocurren con los apellidos hispanos, aunque en menor cantidad. Por ende, decidimos buscar un método que nos permita verificar la existencia de apellidos que tienen como origen otros apellidos con los que sí contamos en nuestra base de datos. Por ejemplo, en la base de datos a clasificar tenemos el apellido "HAREBALO" y "QUISP", mientras que en nuestra base de datos españoles contamos con "AREVALO" y en la base de datos indígenas tenemos "QUISPE". Para poder clasificar estos apellidos de la lista planteamos un método de proximidad usando la "Distancia Levenshtein".

La distancia Levenshtein es la mínima cantidad de cambios que se le debe hacer a una palabra para que sea idéntica a otra. Con ella planteamos un puntaje basado en el porcentaje de letras idénticas (por caracter y posición) que comparten los apellidos del listado con cada apellido de nuestras bases de datos, quedándonos con el mayor puntaje de proximidad encontrado. Para poder implementar esta metodología en nuestro análisis se optó por usar la librería "RecordLinkage"para R, así que la base de datos fue procesada usando este lenguaje solo para los cálculos del puntaje de proximidad.

Retomando los ejemplos anteriores, para el caso de HAREBALO y AREVALO, hay una diferencia de 2 caracteres, por lo que el puntaje de proximidad que planteamos sería obtenido por (8-2)/8 = 0.75. En el caso de QUISP y QUISPE la diferencia es de solo un caractér y se obtendría (6-1)/6 = 0.83.

[1]Para hispanos, las fuentes fueron: Apellidos españoles y Apellidos de España. Sin embargo, los gruesos de la base de datos fueron: De Platt (1996) y Nombres y apellidos españoles- INE

[2]Para indígenas, las fuentes fueron: Manuel de Lucca (1983), Valenzuela (2018), Reniec (2012), Lingotario de apellidos Quechuas, Apellidos Quechuas que fueron modificados al español y Apellidos Quechuas

[3]Las fuentes fueron: Diccionario ilustrado de la lengua Aymara y Nuevo Diccionario: Español-Quechua, Quechua-Español

Para realizar la clasificación se debe plantear un puntaje referencial, visto como el mínimo puntaje de similitud que debe haber entre dos palabras para que las podamos considerar relacionadas. Con la intención de poder analizar la cantidad de observaciones que se logran clasificar usando distintos puntajes referenciales, realizamos la clasificación con los puntajes: 0.50, 0.55, 0.60, 0.65, 0.70, 0.75 y 0.80. Los resultados obtenidos con cada puntaje están resumidos en la tabla *Observaciones clasificadas por puntaje de similitud* del anexo.

El algoritmo para la clasificación del origen de los apellidos tomando en cuenta el puntaje referencial mencionado es el siguiente: primero se verifica que al menos uno de los puntajes sea mayor al puntaje referencial, luego se compara el mayor puntaje de proximidad a un apellido español con el mayor puntaje de proximidad a un apellido indígena (los apellidos respectivos se pueden visualizar en los resultados), si uno de los puntajes es mayor al otro, se tomará este como el origen del apellido. En caso de que los puntajes máximos sean iguales, se realizará una comparación entre la cantidad de veces que las palabras relacionadas al apellido analizado se repiten entre las fuentes hispanas e indígenas de nuestra base de datos. Clasificando al apellido con el origen que tenga mayor sustento, en caso la cantidad de fuentes sea igual o no se cumpliese ninguno de los requisitos anteriores, el apellido queda sin clasificar.

Casos considerados en la limpieza de datos:

1. **Apellidos con prefijo:** Existen algunos casos de apellidos con algún prefijo como -viuda-. Ejemplo el apellido "vda. de garcía". En ese tipo el apellido buscado fue "García". Se creó una dummy caracterizando este caso

2. **Apellidos no factibles:**"A.", "G", "L", etc. En ese caso, se realizó una limpieza adicional. Se creó una dummy caracterizando este caso

3. **Apellidos de 2 letras:** LY, FU, JO, YI, LI, DE etc. En este caso, se intuye que algunos representan un error (como el caso de "DE"), pero otros podrían ser apellidos reales (aunque parecen de origen asiático).

   Con respecto a este último, se creo una dummy caracterizando ese caso. Además, realizamos un *merge* con una base de datos de apellidos japoneses, sin embargo, no hubo ningún *match*.

4. **Apellidos conjuntos:** Carpio Avedanho, Huaman de los Heros, Pacheco Talavera, Concha Fernandez, etc. Lo complicado fue el caso de clasificación en aquellos donde el primer apellido es hispano y el segundo indígena, como en el caso de Huaman de los Heros.

   En este caso, se colocó el origen del primer apellido y adicionalmente se creó otra variable con el origen del segundo apellido.

5. **Apellidos ingleses:** Smith, Hakanson, Williamson, etc. Se creó una dummy caracterizando este caso.

6. **Apellidos croatas:** Miovich, Jovich, Ivancovich, etc. Se creó una dummy caracterizando este caso.

Explicación de las variables en las bases de resultados:

Además de las variables presentadas en la siguiente lista, en la base de datos se encontrarán variables con terminación "_1", "_2" y "_3", estas variables solo toman valores en los casos de apellidos conjuntos explicados anteriormente, y son el equivalente de la variable sin la terminación, pero para cada miembro del apellido conjunto.

1. **Apellido:** Lista con los apellidos entregados en la base de datos original.

2. **APELLIDO:** Resultado de la limpieza de los apellidos realizada con STATA.

3. **origen:** Variable categórica que clasifica los orígenes requeridos tomando en cuenta el método mencionado anteriormente, 0 indica que su origen no es ni Hispano, ni indígena, toma valor 1 cuando el origen es indígena y por último, toma valor 2 cuando el origen es hispano.

4. **hisp_word:** Es el apellido perteneciente a nuestra base de datos de apellidos españoles más cercana al apellido a clasificar.

5. **ind_word:** Es el apellido perteneciente a nuestra base de datos de apellidos españoles más cercana al apellido a clasificar.
6. **index_hisp:** Puntaje de similitud obtenido al aplicar el método de la distancia Levenshtein entre la palabra española de mayor similitud con el apellido de la base de datos original.
7. **index_ind:** Puntaje de similitud obtenido al aplicar el método de la distancia Levenshtein entre la palabra indígena de mayor similitud con el apellido de la base de datos original.
8. **caso_2:** Variable dummy: valor 1 cuando el apellido original es no factible.
9. **nom_dos_letras:** Variable dummy: valor 1 cuando el apellido original es de 2 letras.
10. **viuda:** Variable dummy: valor 1 cuando el apellido original tienen prefijo vda. o contiene la palabra viuda.
11. **apellidos_conjuntos:** Variable dummy:valor 1 cuando el apellidos es conjuntos
12. **origen_eng:** Variable dummy: valor 1 cuando el apellido es de origen inglés, visto como aquellos que contienen la conjunción "TH", propia del lenguaje inglés.
13. **origen_croata:** Variable dummy: valor 1 cuando el apellido es de origen croata, visto como aquellos apellidos con terminación "VICH", propio del lenguaje croata.
14. **fuentes_hisp:** Indica la cantidad de veces que *hisp_word* se repite entre las fuentes hispanas.
15. **fuentes_ind:** Indica la cantidad de veces que *ind_word* se repite entre las fuentes indígenas.

Anexos

Base de datos de apellidos y sus fuentes por origen

solo_hisp	74, 298
solo_ind	31, 742
ambas_bases	902

Ejemplos de casos de empate

Apellido	**Significado Indígena**	**Significado Hispano**
ROCA	Proveniente de Ruq'a : El que obra y tiene prudencia. Ej. Inca Roca	Piedra
RATA	Enredadera de la Yunga Sirwana: Yunka rata rata sirwana	Roedor
LUNA	frutero, con el frutero, tapa bien la fruta.	Satélite Natural
OLAYA	Variedad de cuarzo compacto, lo que produce chispas	Eulaios, Elocuente
MANCO	Proveniente del Manqu: Valiente. Ej. Manco Capac	Que ha perdido un brazo

Observaciones clasificadas por puntaje de similitud

	indígenas	hispanos	origen_eng	origen_croata	mismo_puntaje	empate_no_clasi	clasificados
bd_index_50	8, 114	24, 470	187	40	9, 326	9, 324	32, 475
bd_index_55	8, 069	24, 406	187	40	9, 326	9, 324	32, 372
bd_index_60	7, 947	24, 300	187	40	9, 326	9, 324	32, 154
bd_index_65	7, 384	23, 642	187	40	9, 326	9, 324	30, 957
bd_index_70	6, 735	22, 490	187	40	9, 326	9, 324	29, 172
bd_index_75	5, 556	20, 567	187	40	9, 326	9, 324	26, 093
bd_index_80	4, 250	17, 650	187	40	9, 326	9, 324	21, 901

Referencias

De Platt, L. (1996). *Hispanic surnames and family history.* Genealogical Publishing Com.

Manuel de Lucca, D. (1983). *Diccionario aymara-castellano, castellano-aymara.* Comisión de Alfabetización y Literatura en Aymara.

Reniec (2012). Introducción a un tesoro de nombres quechuas en Apurimac.

Valenzuela, P. (2018). Tesoro de nombres shipibo-konibo.