Distr.: General
21 November 2022

English

Economic Commission for Europe

Conference of European Statisticians

Group of Experts on Migration Statistics

Geneva, Switzerland, 26–28 October 2022
Item A of the provisional agenda
Improvements in use of administrative data for migration statistics

Bases for the conformation of a statistical registry of migrant population

Note by National Administrative Department of Statistics (DANE)

Abstract

With the purpose of advancing in the implementation of the Comprehensive Migration Policy that has as its axis the recognition of Colombians abroad, and the effective enjoyment of the rights of immigrants and returnees, the Sectorial Group of Migration Statistics, in charge of the National Administrative Department of Statistics (DANE), defined the need to implement the Migration Statistics Information System – SIEM (for its acronym in Spanish). The purpose of this system is to organize, consolidate and disseminate information, so that it can be used and processed in an efficient, timely and affordable manner. The SIEM is made up of public and private entities that are producers or users of data, policies, standards, technical processes, and infrastructure involved in the management of information related to migration. Based on the idea of having reliable, first-source, quality information on international immigrants, DANE consulted several information-producing institutions to obtain administrative records of people identified as foreigners to characterize this population. In this way, 11 databases of administrative records were obtained in which foreigners (by country of birth) and Colombians living abroad are found. The objective of this paper is to describe the registry integration methodology and show its potential to characterize international migrants and include them in the Population

*Prepared by Andres Felipe Copete Martínez, Rafael Andrés Urrego Posada, Juan Sebastián Oviedo Mozo, Mariana Francisca Ospina Bohórquez, Steven Cifuentes Rugeles

NOTE: The designations employed in this document do not imply the expression of any opinion whatsoever on the part of the Secretariat of the United Nations concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

Base Registry. The following steps were followed for the process of register integration: I) Data cleaning: process that consists of standardizing the formats and lengths of the variables. Data cleaning included removing special characters that do not allow correct reading in programs for statistical analysis in variables such as residence address or names and surnames. II) Coding of variable categories: unification of numerical codes of variable categories such as gender, countries of birth or nationality (ISO Alpha 3166-3), type of document, department or municipality codes. III) Elimination of deterministic duplicates (by document number) in each administrative record. IV) Elimination of probabilistic duplicates (similarity in names and surnames) in each administrative record. V) Conformation of a single registry with the individual registries. VI) Elimination of probabilistic duplicates of the single record. This methodological procedure makes it possible to demonstrate a significant advance in the creation of a statistical registry of international migrants, which allows the generation of statistics on the location and demographic characteristics of the migrant population, useful for the design of public policies.

I. Introduction

1. Since 2015, a massive immigration phenomenon has been observed from Venezuela, of which Colombia does not have sufficient records, especially because the registry of international entries and exits of people does not contain information on the border with Venezuela, so its documentation, analysis and description become a necessity, not only to characterize the immigrant population, but also to design public policies that allow their integration and effective enjoyment of rights. Despite the efforts of the Colombian Government in humanitarian attention, such as the creation of the Special Permanence Permit (PEP, for its acronym in Spanish), which allowed access to services, such as education and health, or the characterization of the population through the Administrative Registry of Venezuelan Migrants (RAMV, for its acronym in Spanish), the growing number of immigrants has overwhelmed the institutional and administrative capacity to allow the integration of Venezuelan citizens.
2. In this way, the document CONPES 3950 is created, which, among other provisions, guides state institutions towards strengthening information systems on the migrant population. In this sense, DANE has carried out actions to generate a Base Statistical Registry of the Migrant Population (REBPM, for its acronym in Spanish), which serves as a reference for the institutions and from there generate public policies oriented towards the migrant population.
3. This document shows the methodology that was used to generate the REBPM from different administrative records of the migrant population, which some entities of the national order shared. The document is divided into three sections, in the first a description and analysis of completeness and quality of the variables of each of the administrative records is made; in the second, the cleaning and standardization of the variables; and the third describes the processes of elimination of duplicates by deterministic and probabilistic methods and the imputation of the sex variable.
4. It is important to point out that all the refined information on international immigrants is likely to be included in the Population Base Statistical Registry. So that there is a framework where the entire population residing in Colombian territory is included, and thus have a

longitudinal reference to capture the different demographic phenomena, especially those related to the migrant population, which, finally, is the one that most lacks information.

II. Administrative records shared by other entities

5. Due to the growing need for information from the immigrant population, DANE, through the Interinstitutional Migration Board, asked the entities that comprise it to provide administrative records of people who were born in a country other than Colombia. In this way, four institutions - among which is the Colombian Institute of Family Welfare (ICBF), Migration Colombia, the Foreign Ministry, and the Ministry of Labor -, offered 11 administrative records. These administrative records have at least identification variables such as document type and number, names and surnames, country, and date of birth, as well as characterization variables (specific to the topic of each record) and in some cases location variables.

A. ICBF administrative registers

6. The Colombian Institute of Family Welfare (ICBF) has two administrative records. 1) "Cuéntame", which is an information system aimed at supporting the management and collection of information on the services offered by the ICBF Early Childhood Directorate in the national territory. And 2) Administrative Processes for the Restoration of Rights (PARD), which is an instrument to guarantee the effective exercise of the rights of children and adolescents in the face of their non-observance, threat or violation. These records are divided into foreign population and Venezuelan population. The number of records each of the databases contains is shown below.

Administrative Register	Year	Number of registers
Cuéntame foreign	2018	70.299
	2019	116.359
	2020	113.140
	2021	95.263
Cuéntame Venezuelans	2018	66.270
	2019	112.220
	2020	109.182
	2021	87.146
PARD foreign	2018	1.310
	2019	2.421
	2020	2.957
	2021	3.133
PARD Venezuelans	2018	1.310
	2019	2.421
	2020	2.957
	2021	3.133

Table 1. ICBF administrative registers

B. Administrative registers of Migration Colombia

7. Within the administrative records of Migration Colombia there are three types. 1) Immigration Certificates: documents delivered to foreigners who have carried out an administrative procedure and intend to have permanent residence in Colombia. 2) Special Permanence Permit (PEP): Permit granted to people from Venezuela as a migration facilitation mechanism for Venezuelan nationals, which would allow the preservation of internal and social order, avoid the labor exploitation of these foreigners, and ensure their permanence in decent conditions in the country (Resolution 5797 of 2016[1]). And 3) Unique Registry of Venezuelan Migrants (RUMV): "aims to collect and update information as input for the formulation and design of public policies, and to identify migrants of Venezuelan nationality[2] (...)".

8. These records contain identification information, such as document type and number, names and surnames, country of birth, nationality, and date of birth. As evidenced, the gender variable was not provided, and it was necessary to impute it. The method used for the imputation and the results obtained will be presented later.

Administrative Register	Number of registers
Foreigner IDs	100.390
Special Stay Permit (PEP)	187.870
Unique Registry of Venezuelan Migrants (RUMV)	1.048.001

Table 2. Administrative registers of Migration Colombia

C. Administrative registers of the Ministry of Foreign Affairs

9. One of the functions of the Foreign Ministry is to grant passports and manage the registry of Colombians abroad (Consular Registry), these two registries are oriented only to people with Colombian nationality, so they do not provide information about foreigners in Colombia. However, this information makes it possible to characterize those people who apply for a passport or who are residing abroad. On the other hand, the Foreign Ministry shared the administrative records of visas granted to citizens of other countries so that they are allowed to enter and stay in Colombia. The number of records in each database is presented below.

Administrative Register	Number of registers
Consular registration	423.682
Passaports	3.832.930
Visas	338.629

Table 3. Administrative registers Ministry of Foreign Affairs

D. Administrative register of the Ministry of Labor

10. The Ministry of Labor has an administrative registry, the Single Registry of Foreign Workers in Colombia (RUTEC), which has information on foreign workers linked to or

[1] https://dapre.presidencia.gov.co/normativa/normativa/DECRETO%20216%20DEL%201%20DE%20MARZO%20DE%202021.pdf

[2] Decree 2016 of 2021

formally hired in Colombia. This record is filled out by the employer, it provides information about work activity and identification variables, demographic and socioeconomic variables. Next, the number of records that make up the database is presented.

Administrative Register	Number of registers
RUTEC	51.465

Table 4. Administrative register of the Ministry of Labor

III. Cleaning and standardization of information

11. Given that the administrative records shared by the entities do not have statistical purposes, it is necessary to purge them in such a way as to achieve an integration of all the registers and thus perform queries optimally. This section will present the procedures to standardize the variables of interest such as names, surnames, document types and gender.

12. The Cuéntame and PARD registries of the ICBF, in the year 2018, have the coding M for “Male” and F for “Female”. However, for the following years the coding changes, M for "Woman" and H for "Man". The standardization of this variable is established as 1 for “Men” and 2 for “Women”.

13. On the other hand, the categories of the Document Type, Ethnicity, and Class variables are standardized. The coding is shown in the following tables.

NAME_TYPE_IDENTIFICATION	ID_TYPE_IDENTIFICATION
CIVIL REGISTRATION	1
IDENTITY CARD (Children)	2
CITIZENSHIP CARD	3
FOREIGNER ID	4
PASSPORT	5
DIPLOMATIC CARD	6
TAX IDENTIFICATION NUMBER	7
DANE´S CERTIFICATE OF LIVE BIRTH	8
SAFE-CONDUCT OF PERMANENCE	9
SPECIAL RESIDENCE PERMIT	10
UNIDENTIFIED ADULT	11
UNIDENTIFIED MINOR	12
VISA	13
BORDER MOBILITY CARD	14
TEMPORARY PROTECTION PERMIT	15
FOREIGN IDENTIFICATION	98
NO DOCUMENT TYPE	99

Table 5. Coding of the Document Type variable

14. Names and surnames may contain special characters such as accents, tildes, umlauts, etc., which complicates the reading of data in different information processing programs, and eventually does not allow the matching of individuals because they are written differently. different ways. To solve this inconvenience, a standardization of "regular expressions" is made. In addition, it is necessary to transform all the character strings into uppercase letters, in this way different administrative records can be linked using the names of the people as a key (in case of not having the type and number of the document).

15. The variables of departments, municipalities and countries are standardized through the Colombian political-administrative division codes (DIVIPOLA), and country of birth and nationality codes according to the ISO 3166-3 coding.

IV. Identification of duplicates

A. Deterministic Duplicates

16. The identification of deterministic duplicates was carried out based on the information of the identification variables (type and number of document). An order must be made, first by the type of document and then by the number. In this way, which ones are repeated were identified and the first of these was preserved. This procedure was carried out in each of the records and the following results were obtained:

Font	Administrative Register	Cases	Deterministic Duplicates
Migración Colombia	PEP	187.870	3.602
	RUMV	1.048.001	1
	Foreigner IDs	100.390	218
ICBF	Cuéntame foreign	395.061	0
	Cuéntame Venezuelans	374.818	0
	PARD foreign	9.821	34
	PARD Venezuelans	8.742	25
Ministry of Foreign Affairs	Passports	3.832.930	100.966
	Visas	338.629	116.436
	Consular Registry	423.682	0
Ministry of Labor	RUTEC	51.465	35

Table 6. Deterministic duplicates identified in each Administrative Register

B. Deterministic Duplicates

17. Despite the use of a basic matching of the records through variables such as names, surnames and date of birth, always looking for attributes that make the instance as unique as possible, in this way it is not possible to identify levels of similarity, since this method it only identifies those that were exactly the same.

18. The method described by Fellegi and Sunter [1969] is used to perform the record linkage task using statistical and probabilistic principles. The first step is to establish the linking rules, which a priori form sets of records with high similarity, moderately similar and those that are different from each other, the method requires knowledge of these parameters for classification, that is, knowing the distribution of the three sets formed by the binding rule.

19. In the process of identifying the fairly similar, there is a calculation process, and it is the Jaro Winkler method, which consists of calculating the number of changes that must be incurred so that two-character strings are equal, in the calculation process it uses the Levenstein measure and Q-grams.

20. The ECM algorithm (conditional maximum likelihood estimation) was used to estimate the parameters established in the linkage rule, making the comparison between the attributes independent given the unknown state of the link or linkage, obtaining optimal convergence properties as a result.

21. The algorithm is mainly based on the choice of attributes that provide a high cardinality to the set of records, consequently, for the exercise the surname of the people is chosen. This is because previously, in a frequency graph, a greater number of groups could be discriminated. In contrast to the names, the groups in the surnames variable occurred more frequently and consequently the power of distinction of the linking rule was reduced.

22. Using a Cartesian product, the probability of similarity between the records classified by the linkage rule was obtained. Finally, those that showed a similarity probability greater than 85% were evaluated to determine their quality as duplicates (Enamorado, 2019).

23. The linking of records from different information sources, which lack an identifier variable (ID), leads to generating complex procedures that allow identifying the similarity between a pair of records by comparing attributes such as name, surname and date of birth. Thus, for the identification of probabilistic duplicates, the record linkage algorithm was deployed to identify the number of matches within the data set that were classified as duplicates.

C. Record Linkage

24. Record Linkage is the term used to indicate the process of joining records from two or more information sources, which are believed to belong to the same entity, or to find duplicates in a single information source. It is usually used to connect records for which a unique identifier is not known, in this case, the data is linked using attributes such as name, surname, gender, date of birth or municipality of residence.

25. In the Record Linkage process, comparing all pairs of records (in all attributes) can be computationally expensive, therefore, several techniques have been developed to make a strategic choice of the records to be compared. These methods are based on the fact that many records do not belong to the same entity and are usually called indexing processes.

26. The indexing technique consists of finding pairs of records that possibly belong to the same entity. After their identification, the comparison between records is made only on the possible candidates and the other pairs of records are not considered in the process. Given its importance, indexing must be done carefully because if a pair of records are not candidates, they can never be linked. Also, if many records are candidates, the computational cost remains high.

27. The most widely used indexing method is known as standard indexing or blocking. In this method the record pairs are compared by a single record attribute called the blocking key. All the records that match the lock key are assigned to only one block, therefore, the obtained blocks are mutually exclusive. The highly recommended attributes to be a lock key are those with very few errors, with little missing information, and stable over time.

28. After the creation of each block, the comparison between records is done using similarity measures. If two records are identical, the measure between them is 1, on the contrary, if the attributes of two records are completely different, the similarity will be zero. The similarity measures and the Record Linkage process used by Fellegi and Sunter are described below.

D. Theoretical framework of Fellegi and Sunter

29. In 1969 Fellegi and Sunter assume two populations A and B, including the special case where A=B, for the identification of duplicates in a single source of information. Each element in the populations has a specific number of attributes, for example, age, gender, and date of birth.

30. A record a belonging to population A can represent the same entity (individual, company, etc.) as record b belonging to population B. The idea is to match record a and record b and decide if they represent the same entity. Therefore, the set A x B (Cartesian product between A and B) is created, where all the possible combinations between the records of population A and population B are found.

31. The set A x B is divided into two disjoint subsets, a subset where all the pairs of records (a,b) represent the same entity (or the same person), called the link set, and another subset where all the pairs of records (a,b) do not represent the same entity, called a non-link set. Each pair of records (a,b) has a true link status M, which is considered a random variable and takes the value of 1 in all the pairs that belong to the link set and the value of 0 if the pairs belong to the non-link set. Since the true matching state between the records is unknown, the goal is to estimate M for each pair of records.

32. Each pair of records (a,b) is compared by means of a similarity measure (or comparison function) that compares k attributes between the records. The differences between each of the attributes are stored in a k-dimensional vector denoted by y, called the comparison vector. The comparison vector is further assumed to be a realization of the k-dimensional random vector Y, which represents the true (unknown) differences between each pair of records.

33. For each pair of records we have the random vector (Y,M). The realization of Y can be observed through y; however, the true matching state M is an unobserved latent variable that is related to the differences between attributes y.

34. Fellegi and Sunter formulate their theory based on linkage rules. These rules are used to classify each of the pairs of records (a,b) into link set and non-link set subsets. The rules they propose are closely related to the rules established in Decision Theory and are based on the following conditional probabilities:

    - $m(y) = P(Y = y \,/\, M = 1)$ Probability that the vector y is a realization of Y given that it belongs to the link set (it is a true link).
    - $u(y) = P(Y = y \,/\, M = 0)$ Probability that the vector y is a realization of Y given that it belongs to the non-link set.

35. The main objective of the authors is to find an optimal and powerful decision rule to distinguish between the distributions of true links (record pairs that belong to the link set) and true non-links (record pairs that belong to the non-link set). set). The binding rule is based on the following likelihood ratio:

$$(y) = \frac{P(Y = y \,/\, M = 1)}{P(Y = y \,/\, M = 0)} = \frac{m(y)}{u(y)}$$

36. For more details on the definition of the binding rule and all the proofs of why it is the most powerful rule, (see Fellegi and Sunter, 1969, p. 1201-1207).

37. In the model proposed by Fellegi and Sunter, the parameters of interest are the probabilities m, u and π, where π=P(M=1) is known as the linkage prevalence. The authors use Bayes'

Theorem to find expressions involving the three parameters of interest and facilitate the application of iterative estimation algorithms. The expressions found are:

- The probability that it is a link (successful match) given the vector of differences between the attributes Y:

$$P(M = 1 \,/\, Y = y) \;=\; \frac{m(y)\,\pi}{m(y)\,\pi \;+\; u(y)\,(1-\pi)}$$

- The probability that it is not a link, given the vector of differences between the attributes Y:

$$P(M = 0 \,/\, Y = \; y\,) \;=\; \frac{u(y)\,(1-\pi)}{u(y)\,(1-\pi) \;+\; m(y)\,\pi}$$

38. Jonathan de Bruin proposes that the algorithm used for the estimation of the model parameters be the EM (Expectation-Maximization) Algorithm, since after comparing different estimation methods such as loglinear models, Bayesian networks and the original estimation method proposed by Fellegi and Sunter, the EM algorithm was the best performer.

39. The EM Algorithm is an iterative algorithm used to calculate maximum likelihood estimates, mainly in incomplete data problems. In the context of Record Linkage, the true link status M is a latent variable and is seen in the algorithm as the incomplete data. The algorithm allows estimating the probabilities m(y) and u(y) from the iterative expressions found by Bayes' Theorem.

E. Identification of probabilistic duplicates in the migration RRAA

40. The algorithm developed by Fellegi and Sunter and optimized by Brain, was used to identify probabilistic duplicates among the migration RRAA, listed below:

Tables	*Registres*	*Duplicates*	*%Duplicates*
Foreigner IDs	100.172	78	0,08%
Cuéntame foreign	395.061	134.131	33,95%
Cuéntame Venezuelans	374.818	127.384	33,99%
PARD foreign	9.787	207	2,12%
PARD Venezuelans	8.717	189	2,17%
Passaports	3.731.964	1.192	0,03%
PEP	184.268	434	0,24%
Consular registration	423.682	3.015	0,71%
RUMV	1.048.000	10.885	1,04%
RUTEC	51.465	1.193	2,32%
Visas	222.193	13.188	5,94%

Table 7. Migration RRAA list with number of duplicates

41. Subsequently, all the tables were joined vertically to finally search for duplicates by deterministic and probabilistic methods; Once the database is purified by identifying and extracting duplicates by type of document and document number, the duplicates are identified using the probabilistic method. The results are summarized in Table 8. Annex 1 presents the algorithm that was carried out in the collaborative environment of Google.

Migration table processing	*Registers*
Join table	6.258.191
Deterministic duplicates	-256.584
Probabilistic Duplicates	-469.813
Total migration records	**5.531.794**

Table 8. Union of the refined RRAA, Migration table

F. The processing stages are defined in 5 steps:

42. Variable cleaning: The most common data inconsistencies involve variations in name spellings, such as nicknames (for example, Rob and Robert), date of birth formats, data encoding, and missing information. Techniques used to resolve these inconsistencies to convert data into standard forms include: editing, standardization, deduplication, and matching.

43. Standardization is the formatting of data elements so that they are represented consistently across all data sets. For example, dates can be represented in different formats, such as DD/MM/YYYY or MM/DD/YYYY. One format must be selected and used on all data sets for the project.

44. data linking; involves a lot of record comparisons. Ideally, each record in dataset A is compared to each record in dataset B, to find which pairs of records are most likely to be links. However, if each record is compared between two data sets containing 100,000 records each, this would require 10 billion comparisons. Even using advanced computing power, this would take considerable time.

45. As a way to save time, "locking" is used to reduce the number of record comparisons required to find potential record pairs. For example, when using gender for blocking, only records of the same gender (male or female) are compared against each other, typically halving the number of required comparisons. However, gender isn't too useful for blocking, as it just separates the dataset into two big blocks, so a lot of comparisons are still required.

46. Ideally, a locking strategy should generate small blocks of the same size in each data set. For example, using the month of birth would result in 12 blocks (one for each month) and would be expected to have an even number of records in each block. A common strategy is to keep block sizes as small as possible and thus compare as few records as possible.

47. In our case, the variables used are: Names, Surnames and Date of birth. Just by looking at a bar chart of the variables involved, it was possible to establish that the surnames formed smaller comparison groups, thus, the titanic task of making a comparison using the Cartesian product of a record versus n – 1 registers remaining in the table, became to compare only those records that have the same last name.

48. The Splink configuration; it was made under a Json schema. Under this configuration, the default parameters in the Splink library can be modified. In this scheme, the "last name" variable is established as a condition, which generates the blocks identified in the previous point (the blocking rule). In the type of linking, the "dedupe_only" parameter was chosen, which works only to identify duplicates within the same RRAA, in this way the comparison columns are configured, which in this case are "Full names" and "Date of birth". Additionally, it is essential for the operation of the algorithm, an identifier variable, unique for each record, in this way the concordance by pair of rows can be identified.

49. In the configuration of the comparative in the names, 3 levels of comparison were established, where:

    - Level 2: The compared strings are the same
    - Level 1: The compared text string are similar
    - Level 0: There is no similarity in the text strings.

50. Date of Birth is not configured, but by default there are 2 levels:

- Level 1: The compared text strings are the same.
- Level 0: the compared text string is different.

51. The comparison can be performed in many different methods to calculate similarity values on a string of text, numeric values, or dates. In our scenario, where we are calculating the similarity score for the values of the text string, the Jaro Winkler algorithm has been used, which is a hybrid between Q-grams and the Levenstein distance.

52. Set the classification threshold; After running the EM algorithm to estimate the probabilities m(y) and u(y) of each pair of records in each block, match probabilities by name, match probabilities by date of birth, and overall match probabilities between records were obtained. All pairs of records that had an overall matching probability greater than or equal to 0.85 were saved.

53. On page 9 of the article "Using a Probabilistic Model to Assist Merging of Large-Scale Administrative Records" (Romántico, 2019), the author identified that from a similarity probability of 85% or greater, they can be considered to be equal.

V. Probabilistic sex prediction by name

54. In the REBPM consolidation process, those variables with invalid, inconsistent, or missing values were identified, so it is necessary to treat them to transform them and improve their quality. Next, the procedures applied to the variables of the Administrative Records to be improved for statistical use will be mentioned, thus facilitating their link between the available information sources.

55. In particular, the administrative records of Migración Colombia do not contain the gender variable, however, within the integrated registry there are some individuals, apart from those that come from the Migración Colombia databases, who do not have information on gender, to complete this missing information. the methodology was proposed consisting of:

- Use machine learning techniques or (Machine learning)
- Explore the validated information of names and surnames of the records that have this data.

56. For the imputation of the gender variable in the records without information, in this case 1,835,093 records without information were considered; For this procedure, a cleaning phase is carried out, which is described below:

- People with names of 5 or more words were eliminated.
- Special characters such as (\,*,^,...) were removed, which are attributed to typing errors and if not removed, generate bias in the results.
- The digits from 0 to 9 were removed.

- Some articles were deleted, such as: “*el*”, “*del*”, “*de*”, “*los*”, “*las*”.
- Names with less than 3 letters were removed.
- All lowercase records are kept.

57. The conformation of the final base for model training has the following characteristics:

- The observations that have only one name are selected; this will be quite important in the model since it allows discriminating the records of people who only report a name in any of the records.
- People with 2 or more names were concatenated in a single word, the arrangement of names like this has several advantages since in this way the model is given the necessary tools to differentiate the different combinations of names that can be formed from personal names.

58. The choice of this structure is not random, by concatenating the names into a single word and putting them in the same bag with the others, variability is added to the model. What provides more information and evidence to classify a person who is called "María José", for example, and who can be correctly classified as a woman.

59. Once all the necessary elements for training and prediction are available, the new variables generated from the “name” variable are defined, which will be used to train the model. For this, they identified:

    i. Two variables, each with the first 3 and 5 letters of each name.

    ii. Two variables, each with the last 3 and 5 letters of each name.

    iii. A variable that is identical if the name ends in a vowel.

    iv. A variable with the character length of each name.

    v. A variable that identifies whether the name ends in an open vowel, especially useful for differentiating males from females.

60. Additionally, a weight variable “weight” is created that allows solving the problem of class imbalance that occurs when having fewer combinations of names for men, that is, a higher weight is given to the category with less frequency; There are other methodologies that can be used to compensate for the imbalance problem, such as taking a subsample of the category with the highest frequency in such a way that the categories are balanced, however, these were not considered.

61. As a result, there is evidence of a greater combination of compound names (2 names), corresponding to women, while men have mainly a single name or fewer name combinations.

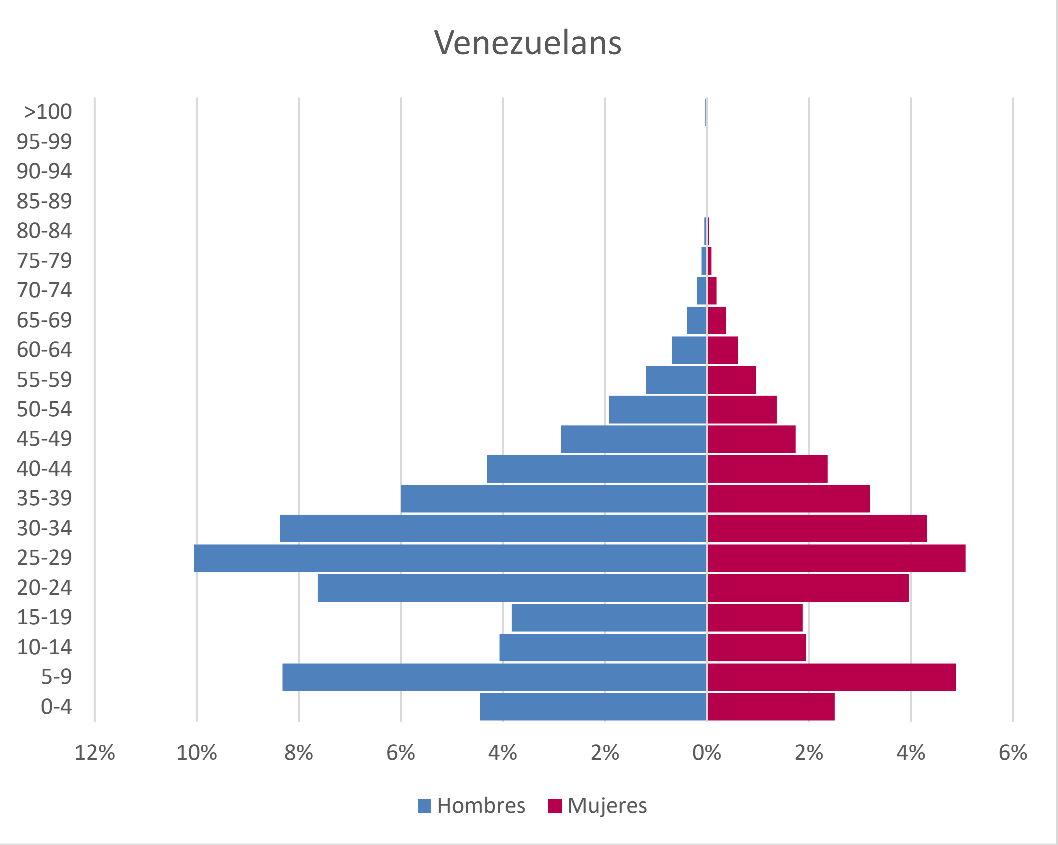
A. Prediction of the gender variable

62. Finally, the imputation of the sex variable was carried out for 1,835,093 individuals, whose results were 627,177 women and 1,207,916 men. It is evident that there is a higher estimate of men and the validation of these results will occur once Migración Colombia shares the information of the individuals with the missing variable.

VI. Some results

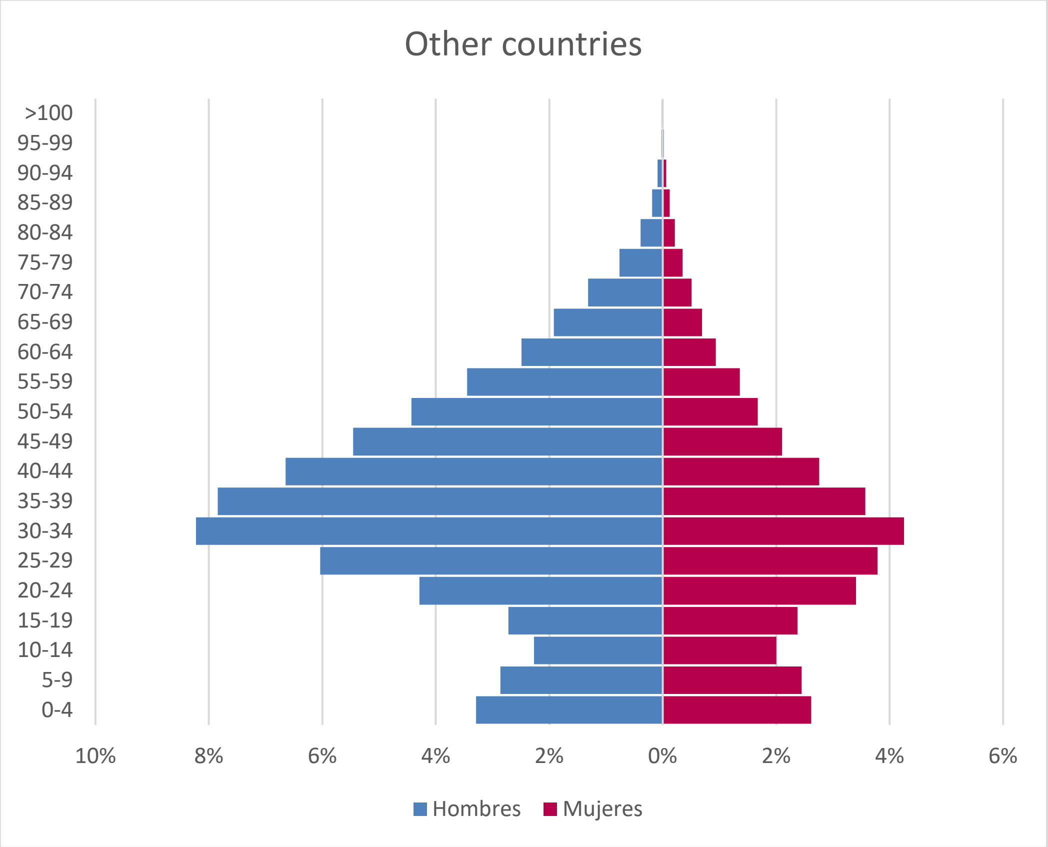
63. One of the most relevant aspects of this process that is being carried out is the possibility of characterizing the immigrant population in Colombian territory. The visualization of this population for purposes of generating public policies that allow integration into society must be given through timely and quality information. For this purpose, a viewer was created on the DANE[3] page in which users can consult information that can be obtained from the administrative records.

64. Some of the results obtained with the Migrant Population Base Statistical Registry are presented below. The first result is the description by age and sex of people from Venezuela and other parts of the world. It is observed that there is a greater immigration of men of working age. Analyzing in detail, in the case of the pyramid of Venezuelans, it is observed that the group from 5 to 9 years old is the one with the highest participation among those under 20 years of age. As for the group of people in working ages over 20 years, the group of 25 to 29 years is the one with the highest participation.



Graphic 1. Population distribution of Venezuelan immigrants

[3] https://geoportal.dane.gov.co/geovisores/sociedad/estadisticas-migracion/



Graphic 2. Population distribution of immigrants from countries other than Venezuela

VII. Annex 1

65. Algorithm to identify probabilistic duplicates

```
[('spark.executor.memory', '2g'),
 ('spark.driver.host', 'cf400d5312bd'),
 ('spark.driver.memory', '4g'),
 ('spark.executor.id', 'driver'),
 ('spark.sql.warehouse.dir', 'file:/content/spark-warehouse'),
 ('spark.driver.port', '35377'),
 ('spark.executor.cores', '10'),
 ('spark.jars', 'jars/scala-udf-similarity-0.0.8.jar'),
 ('spark.cores.max', '15'),
 ('spark.app.name', 'Spark Updated Conf'),
 ('spark.rdd.compress', 'True'),
 ('spark.serializer.objectStreamReset', '100'),
 ('spark.master', 'local[*]'),
 ('spark.submit.pyFiles', ''),
 ('spark.app.startTime', '1652709514996'),
 ('spark.submit.deployMode', 'client'),
 ('spark.driver.extraClassPath', 'jars/scala-udf-similarity-0.0.8.jar'),
 ('spark.ui.showConsoleProgress', 'true'),
 ('spark.app.id', 'local-1652709515160')]
```

66. The environment was configured in such a way that the full capacity of the nodes was used and they worked in parallel. For this configuration, the splink Robin [2020] library was used.