Distr.: General
20 октября 2022 г. 14:57:28

Русский

Европейская экономическая комиссия

Конференция европейских статистиков

Группа экспертов по статистике миграции

Женева, Швейцария, 26-28 октября 2022 года
Пункт A предварительной повестки дня
Положительные изменения в использовании административных данных для статистики миграции

Базы данных для формирования статистического регистра мигрантов

Записка Национального статистического управления (DANE)

Аннотация

С целью содействия реализации Комплексной миграционной политики, в основе которой лежит признание колумбийцев за границей и эффективное осуществление прав иммигрантов и репатриантов, Отраслевая группа статистики миграции в структуре Национального статистического управления (DANE) установила необходимость внедрения Информационной системы по статистике миграции – SIEM (аббревиатура на испанском языке). Целью этой системы является организация, консолидация и распространение информации с тем, чтобы ее можно было использовать и обрабатывать эффективно, своевременно и без значительных затрат. SIEM состоит из государственных и частных организаций, которые являются производителями или пользователями данных, правил, стандартов, технических процессов и инфраструктуры, участвующими в управлении информацией, связанной с миграцией. Исходя из идеи о необходимости надежной, качественной информации из первоисточника о международных иммигрантах, DANE провело консультации с несколькими учреждениями, производящими информацию, для получения административных записей о лицах,

*Подготовили Андрес Фелипе Копете Мартинес, Рафаэль Андрес Уррего Посада, Хуан Себастьян Овьедо Мозо, Мариана Франсиска Оспина Бохоркес

идентифицированных как иностранцы, для характеристики этого населения. Таким образом было получено 11 баз данных административных записей, в которых есть информация об иностранцах (по стране рождения) и колумбийцах, проживающих за границей. Целью данной статьи является описание методологии интеграции регистров и демонстрация ее возможностей с точки зрения характеристики международных мигрантов и включения их в регистр населения. Для интеграции регистров были выполнены следующие шаги: I) Очистка данных: процесс, состоящий из стандартизации форматов и длин переменных. Очистка данных включала удаление специальных символов, не позволяющих корректно считывать информацию в программах для статистического анализа (такие переменные как адрес проживания или имена и фамилии). II) Кодирование категорий переменных: унификация числовых кодов таких категорий переменных как пол, страна рождения или гражданство (ISO Alpha 3166-3), тип документа, коды департамента или муниципалитета. III) Удаление детерминированных дубликатов (по номеру документа) в каждой административной записи. IV) Удаление вероятностных дубликатов (сходства имен и фамилий) в каждой административной записи. V) Формирование единого регистра на основе отдельных регистров. VI) Удаление вероятностных дубликатов одиночной записи. Эта методологическая база позволяет продемонстрировать значительный прогресс в создании статистического регистра международных мигрантов, который позволяет формировать статистические данные о местонахождении и демографических характеристиках мигрантов, необходимые для выработки государственной политики.

I. Введение

1. С 2015 года наблюдается массовая иммиграция из Венесуэлы, которая недостаточно задокументирована в Колумбии. Причем следует учитывать, что регистр въездов и выездов из страны не содержит информации о передвижении через границу с Венесуэлой. По этой причине появилась необходимость задокументировать, проанализировать и описать это явление, что важно не только для характеристики иммигрантов, но и для выработки государственной политики, которая позволила бы им интегрироваться в общество и пользоваться своими правами. Несмотря на усилия колумбийского правительства по оказанию гуманитарной помощи, такие как внедрение Специального разрешения на постоянное проживание (PEP, аббревиатура на испанском языке), которое открывает доступ к таким услугам, как образование и здравоохранение, или характеристика населения посредством Административного регистра венесуэльских мигрантов (RAMV, аббревиатура на испанском языке), растущее число иммигрантов превышает институциональные и административные возможности, позволяющие интегрировать граждан Венесуэлы.
2. Таким образом, создается документ CONPES 3950, который, среди прочих положений, ориентирует государственные учреждения на совершенствование информационных систем, содержащих данные о мигрантах. В этом контексте DANE предприняло шаги по созданию Базового статистического регистра мигрантов (REBPM, аббревиатура на испанском языке), который служит источником справочной информации для учреждений и на основе которого формируется государственная политика, касающаяся мигрантов.

3. В настоящем документе описана методика, которая использовалась для создания REBPM на основе различных административных записей о мигрантах, предоставленных некоторыми национальными органами. Документ разделен на три раздела, в первом приводится описание и анализ полноты и качества переменных каждой из административных записей; во втором дается информация об очистке и нормировании переменных; а в третьем описываются процессы удаления дубликатов детерминированными и вероятностными методами и импутация переменной пола.

4. Важно отметить, что вся уточненная информация о международных иммигрантах скорее всего будет включена в Статистический регистр населения. Таким образом будет обеспечена структура, в которую включено все население, проживающее на территории Колумбии, и, появится справочная информация за много лет о различных демографических явлениях, особенно имеющих отношение к мигрантам. Ведь именно этой информации не хватает больше всего.

II. Административные записи, предоставленные другими органами

5. В связи с растущей потребностью в информации об иммигрантах DANE через Межведомственный совет по миграции обратилось к организациям, входящим в его состав, с просьбой предоставить административные данные о лицах, родившихся не в Колумбии, а в другой стране. Таким образом, четыре учреждения, среди которых Колумбийский институт благосостояния семьи (ICBF), Миграционная служба Колумбии, Министерство иностранных дел и Министерство труда, предоставили 11 наборов административных записей. Эти административные записи содержат как минимум идентификационные переменные, такие как тип и номер документа, имена и фамилии, страна и дата рождения, а также переменные характеристик (по теме каждой записи) и, в некоторых случаях, переменные местоположения.

A. Административные регистры ICBF

6. Колумбийский институт благосостояния семьи (ICBF) ведет два административных регистра. 1) «Cuéntame» - информационная система, предназначенная для поддержки управления и сбора информации об услугах, предлагаемых Управлением раннего детства ICBF на территории страны. А также 2) Административные процессы восстановления прав (PARD) - инструмент, гарантирующий реализацию прав детей и подростков в случае их несоблюдения, угрозы им или их нарушения. Эти записи разделены на иностранное население и население Венесуэлы. Количество записей, содержащихся в каждой из баз данных, показано ниже.

Административный регистр	Год	Количество регистров
Cuéntame - иностранцы	2018	70,299
	2019	116,359
	2020	113,140
	2021	95,263
Cuéntame - венесуэльцы	2018	66,270
	2019	112,220
	2020	109,182

	2021	87,146
PARD - иностранцы	2018	1,310
	2019	2,421
	2020	2,957
	2021	3,133
PARD - венесуэльцы	2018	1,310
	2019	2,421
	2020	2,957
	2021	3,133

Таблица 1. Административные регистры ICBF

B. Административные регистры миграционной службы Колумбии

7. В административных отчетах Миграционной службы Колумбии есть три типа регистров. 1) Иммиграционные свидетельства: документы, выдаваемые иностранцам, прошедшим административную процедуру и намеревающимся постоянно проживать в Колумбии. 2) Специальное разрешение на постоянное проживание (PEP): Разрешение, предоставленное выходцам из Венесуэлы в качестве механизма облегчения миграции граждан Венесуэлы, что позволит сохранить внутренний и общественный порядок, избежать трудовой эксплуатации этих иностранцев и обеспечить их постоянное проживание в достойных условиях в стране (Резолюция 5797 от 2016 года[1]). И 3) Единый регистр венесуэльских мигрантов (RUMV): «направлен на сбор и обновление информации в качестве исходной информации для формулирования и разработки государственной политики, а также на выявление мигрантов-граждан Венесуэлы[2] (...)».

8. Эти записи содержат идентификационные данные, такиекак тип и номер документа, имена и фамилии, страна рождения, национальность и дата рождения. Как видно, переменная пола не была предоставлена, и ее необходимо было импутировать. Метод, использованный для импутации, и полученные результаты будут представлены позже.

Административный регистр	Количество регистров
Удостоверения иностранца	100 390
Специальное разрешение на проживание (PEP)	187 870
Единый регистр мигрантов из Венесуэлы (RUMV)	1 048 001

Таблица 2. Административные регистры миграции в Колумбии

C. Административные регистры Министерства иностранных дел

9. Одной из функций Министерства иностранных дел является выдача паспортов и ведение регистра колумбийцев за границей (Консульский регистр). Оба эти регистра ориентированы только на граждан Колумбии, поэтому они не содержат информации

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

[2] Указ 2016 2021 года

об иностранцах в Колумбии. Однако эта информация позволяет охарактеризовать лиц, подающих заявление на получение паспорта или проживающих за границей. С другой стороны, Министерство иностранных дел предоставило административные записи о визах, выданных гражданам других стран, разрешающих въезд и пребывание в Колумбии. Количество записей в каждой базе данных представлено ниже.

Административный регистр	Количество регистров
Консульская регистрация	423 682
Паспорта	3 832 930
Визы	338 629

Таблица 3. Административные регистры Министерства иностранных дел

D. Административный регистр Министерства труда

10. Министерство труда ведет административный регистр - Единый регистр иностранных работников в Колумбии (RUTEC), в котором содержится информация об иностранных работниках, связанных с Колумбией или официально трудоустроенных в Колумбии. Эти данные заполняются работодателем, в них содержится информация о трудовой деятельности и идентификационных переменных, демографических и социально-экономических переменных. Далее приводится количество записей, составляющих базу данных.

Административный регистр	Количество регистров
RUTEC	51 465

Таблица 4. Административный регистр Министерства труда

III. Очистка и стандартизация информации

11. Учитывая, что административные записи, предоставленные органами власти, не предназначены для целей статистики, необходимо очистить их таким образом, чтобы обеспечить интеграцию всех регистров и, таким образом, оптимально выполнять запросы. В этом разделе будет представлен порядок нормирования представляющих в данном случае интерес переменных, таких как имена, фамилии, типы документов и пол.

12. В регистрах Cuéntame и PARD от ICBF в 2018 году использовалась кодировка M для мужчин и F для женщин. Однако в последующие годы кодировка меняется: M означает женщин, а H означает мужчин. Эта переменная нормирована как 1 для мужчин и 2 для женщин.

13. С другой стороны, нормированы такие категории переменных «Тип документа», «Этническая принадлежность» и «Класс». Коды показаны в следующих таблицах.

ИМЯ_ТИП_ОБОЗНАЧЕНИЕ	ДОКУМЕНТ_ТИП_ОБОЗНАЧЕНИЕ
ГРАЖДАНСКАЯ РЕГИСТРАЦИЯ	1
УДОСТОВЕРЕНИЕ ЛИЧНОСТИ (дети)	2
КАРТА ГРАЖДАНСТВА	3
УДОСТОВЕРЕНИЕ ИНОСТРАНЦА	4
ПАСПОРТ	5
ДИПЛОМАТИЧЕСКАЯ КАРТОЧКА	6
НОМЕР НАЛОГОПЛАТЕЛЬЩИКА	7

СВИДЕТЕЛЬСТВО О РОЖДЕНИИ DANE	8
ОХРАННОЕ СВИДЕТЕЛЬСТВО	9
СПЕЦИАЛЬНОЕ РАЗРЕШЕНИЕ НА ПОСТОЯННОЕ ПРОЖИВАНИЕ	10
НЕУСТАНОВЛЕННЫЙ ВЗРОСЛЫЙ	11
НЕУСТАНОВЛЕННЫЙ НЕСОВЕРШЕННОЛЕТНИЙ	12
ВИЗА	13
КАРТА ПОГРАНИЧНОЙ МОБИЛЬНОСТИ	14
РАЗРЕШЕНИЕ НА ВРЕМЕННУЮ ЗАЩИТУ	15
ИНОСТРАННОЕ УДОСТОВЕРЕНИЕ	98
НЕТ ТИПА ДОКУМЕНТА	99

Таблица 5. Переменная кодирования типа документа

14. Имена и фамилии могут содержать специальные символы, такие как знаки ударения, тильды, умлауты и так далее, что усложняет считывание данных в различных программах обработки информации и, в конечном итоге, не позволяет сопоставлять людей, поскольку их имена пишутся по-разному. Чтобы устранить это неудобство, производится нормирование «обычных выражений». Кроме того, необходимо преобразовать все строки символов в прописные буквы, таким образом, различные административные записи могут быть связаны с использованием имен людей в качестве ключа (в случае отсутствия типа и номера документа).

15. Переменные департаментов, муниципалитетов и стран нормированы с помощью колумбийских кодов политико-административного деления (DIVIPOLA), а коды страны рождения и национальности - в соответствии с кодировкой ISO 3166-3.

IV. Выявление дубликатов

A. Детерминированные дубликаты

16. Выявление детерминированных дубликатов осуществлялось на основании информации об идентификационных переменных (вид и номер документа). Запрос делается сначала по типу документа, а затем по номеру. Таким образом, были выявлены повторяющиеся данные и сохранены первые из них. Данная процедура была проведена в каждом регистре, и были получены следующие результаты:

Исходные данные	Административный регистр	Случаи	Детерминированные дубликаты
Migración Colombia	PEP	187 870	3602
	RUMV	1 048 001	1
	Удостоверения иностранца	100 390	218
ICBF	Cuéntame - иностранцы	395 061	0
	Cuéntame - венесуэльцы	374 818	0
	PARD - иностранцы	9821	34
	PARD - венесуэльцы	8742	25
Министерство иностранных дел	Паспорта	3 832 930	100 966
	Визы	338 629	116 436

	Консульский регистр	423 682	0
Министерство труда	RUTEC	51 465	35

Таблица 6. Детерминированные дубликаты, выявленные в каждом административном регистре

B. Детерминированные дубликаты

17. Несмотря на использование базового сопоставления записей через такие переменные, как имена, фамилии и даты рождения, всегда ищут атрибуты, которые делают экземпляр записи максимально уникальным. Таким образом невозможно определить уровни сходства, поскольку это метод идентифицирует только точно такие же записи.

18. Метод, описанный в работе Фелледжи и Сантера [1969], используется для выполнения задачи связывания записей с использованием статистических и вероятностных принципов. Первым шагом является установление правил связывания, которые априори формируют наборы записей с высокой степенью сходства, умеренно похожих и отличающихся друг от друга. Метод требует знания этих параметров для классификации, то есть знания распределения трех наборов, образованных по правилу связывания.

19. В процессе выявления достаточно похожих записей используется процесс вычисления, и это метод Джаро-Винклера, который состоит в вычислении количества изменений, которые необходимо произвести, чтобы двухсимвольные строки стали равными, в процессе вычисления используется расстояние Левенштейна и Q-граммы.

20. Алгоритм ECM (оценка условного максимального правдоподобия) использовался для оценки параметров, установленных в правиле связывания, делая сравнение между атрибутами независимым при неизвестном состоянии связи или привязки, получая в результате оптимальные свойства сходимости.

21. Алгоритм в основном основан на выборе атрибутов, обеспечивающих высокую мощность множества записей, следовательно, для работы выбираются фамилии людей. Это связано с тем, что ранее на частотном графике можно было выделить большее количество групп. В отличие от имен, группы в переменной «фамилии» встречались чаще, и, следовательно, сила разграничения для правила связывания была снижена.

22. Используя декартово произведение, была получена вероятность сходства между записями, классифицированными по правилу связывания. Наконец, те записи, которые показали вероятность сходства выше 85%, подверглись оценке для определения их качества как дубликатов (Enamorado, 2019).

23. Связывание записей из разных источников информации, в которых отсутствует переменная-идентификатор (ID), приводит к созданию сложных процедур, позволяющих выявить сходство между парой записей путем сравнения таких атрибутов, как имя, фамилия и дата рождения. Таким образом, для идентификации вероятностных дубликатов был использован алгоритм связывания записей для определения количества совпадений в наборе данных, которые были классифицированы как дубликаты.

C. Связывание записей

24. Связывание записей — это термин, используемый для обозначения процесса объединения записей из двух или более источников информации, которые считаются принадлежащими одному и тому же объекту, или для поиска дубликатов в одном источнике информации. Этот метод обычно используется для объединения записей, для которых неизвестен уникальный идентификатор. В данном случае данные связываются с использованием таких атрибутов, как имя, фамилия, пол, дата рождения или муниципалитет проживания.

25. В процессе связывания записей сравнение всех пар записей (по всем атрибутам) может потребовать значительных вычислительных ресурсов, поэтому было разработано несколько методов для стратегического выбора записей для сравнения. Эти методы основаны на том, что многие записи не принадлежат одному и тому же объекту и обычно называются процессами индексации.

26. Метод индексации состоит в поиске пар записей, которые, возможно, принадлежат одному и тому же объекту. После их идентификации сравнение между записями производится только по возможным подходящим претендентам, а остальные пары записей в процессе не рассматриваются. Учитывая его важность, индексирование должно выполняться тщательно, потому что, если пара записей не является подходящими претендентами, они никогда не могут быть связаны. Кроме того, если многие записи являются подходящими претендентами, вычислительные затраты остаются высокими.

27. Наиболее широко используемый метод индексации известен как стандартная индексация или блокирование. При использовании этого метода пары записей сравниваются по одному атрибуту записи, называемому ключом блокировки. Все записи, совпадающие с ключом блокировки, относятся только к одному блоку, поэтому полученные блоки являются взаимоисключающими. Настоятельно рекомендуемыми атрибутами для использования в качестве ключа блокировки являются атрибуты с очень небольшим количеством ошибок, небольшим количеством отсутствующей информации и стабильными в контексте времени.

28. После создания каждого блока сравнение записей выполняется с использованием коэффициентов подобия. Если две записи идентичны, коэффициент подобия между ними равен 1, наоборот, если атрибуты двух записей совершенно разные, коэффициент подобия будет равен нулю. Коэффициенты подобия и процесс связывания записей, используемые Фелледжи и Сантером, описаны ниже.

D. Теоретическая база по Фелледжи и Сантеру

29. В 1969 году Фелледжи и Сантер предположили наличие двух популяций A и B, включая особый случай, когда A=B, для выявления дубликатов в одном источнике информации. Каждый элемент совокупностей имеет определенное количество атрибутов, например возраст, пол и дату рождения.

30. Запись a, принадлежащая совокупности A, может представлять тот же объект (физическое лицо, компанию и т. д.), что и запись b, принадлежащая совокупности B. Идея состоит в том, чтобы сопоставить записи a и b и решить, представляют ли они один и тот же объект. Следовательно, создается множество A x B (декартово произведение между A и B), в котором находятся все возможные комбинации между записями совокупности A и совокупности B.

31. Набор AxB делится на два непересекающихся подмножества: подмножество, в котором все пары записей (a,b) представляют один и тот же объект (или одного и того же человека), называемое набором связей, и другое подмножество, в котором все пары записей (a,b) не представляют один и тот же объект, которое называется набором без связей. Каждая пара записей (a,b) имеет истинный статус связи M, который считается случайной величиной и принимает значение 1 во всех парах, принадлежащих набору связей, и значение 0, если пары принадлежат набору без связей. Поскольку истинное состояние соответствия между записями неизвестно, цель состоит в том, чтобы оценить M для каждой пары записей.

32. Каждая пара записей (a,b) сравнивается с помощью коэффициента подобия (или функции сравнения), которая сравнивает k атрибутов между записями. Различия между каждым из атрибутов хранятся в k-мерном векторе, обозначаемом y, называемом вектором сравнения. Далее предполагается, что вектор сравнения является реализацией k-мерного случайного вектора Y, который представляет истинные (неизвестные) различия между каждой парой записей.

33. Для каждой пары записей у нас есть случайный вектор (Y, M). Реализация Y может наблюдаться через y; однако истинное состояние соответствия M является ненаблюдаемой скрытой переменной, которая связана с различиями между атрибутами y.

34. Фелледжи и Сантер формулируют свою теорию на основе правил связывания. Эти правила используются для классификации каждой из пар записей (a,b) как относящихся к подмножеству набора связей или набора без связей. Предлагаемые ими правила тесно связаны с правилами, установленными в теории принятия решений, и основаны на следующих условных вероятностях:

    - $m(y) = P(Y = y \,/\, M = 1)$ Вероятность того, что вектор y является реализацией Y при условии, что он принадлежит набору связей (это истинная связь).
    - $u(y) = P(Y = y \,/\, M = 0)$ Вероятность того, что вектор y является реализацией Y при условии, что он принадлежит набору без связей.

35. Основная цель авторов состоит в том, чтобы найти оптимальную и мощную решающую функцию для разграничения распределений истинных связей (пар записей, принадлежащих набору связей) и истинных не-связей (пар записей, принадлежащих набору без связей). Правило связывания основано на следующем отношении подобия:

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

36. Для получения более подробной информации об определении правила связывания и всех доказательствах того, почему это самое сильное правило, (см. Fellegi and Sunter, 1969, p. 1201-1207).

37. В модели, предложенной Фелледжи и Сантером, представляющими интерес параметрами являются вероятности m, u и π, где π=P(M=1) известно как частота связывания.

    Авторы используют теорему Байеса для нахождения выражений, включающих три интересующих параметра, и облегчают применение итеративных алгоритмов оценки. Найдены выражения:

- Вероятность того, что это связь (успешное совпадение) при заданном векторе различий между признаками Y:

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

- Вероятность того, что это не связь, при заданном векторе различий между признаками Y:

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

38. Джонатан де Брюин предлагает, чтобы для оценки параметров модели использовался алгоритм EM (Максимизация ожидания), поскольку после сравнения различных методов оценки, таких как логлинейные модели, байесовские сети и оригинальный метод оценки, предложенный Фелледжи и Сантером, Алгоритм EM показал лучший результат.

39. Алгоритм EM — это итерационный алгоритм, используемый для вычисления оценок максимального правдоподобия, в основном в задачах с неполными данными. В контексте связывания записей истинный статус связи M является скрытой переменной и рассматривается в алгоритме как неполные данные. Алгоритм позволяет оценить вероятности m(y) и u(y) из итерационных выражений, найденных по теореме Байеса.

E. Выявление вероятностных дубликатов в административных записях о миграции

40. Алгоритм, разработанный Фелледжи и Сантером и оптимизированный Брэйном, использовался для выявления вероятностных дубликатов среди миграционных административных записей, перечисленных ниже:

Таблицы	*Регистры*	*Дубликаты*	*% дубликатов*
Удостоверения иностранца	100 172	78	0,08%
Cuéntame - иностранцы	395 061	134 131	33,95%
Cuéntame - венесуэльцы	374 818	127 384	33,99%
PARD - иностранцы	9 787	207	2,12%
PARD - венесуэльцы	8 717	189	2,17%
Паспорта	3.731.964	1 192	0,03%
PEP	184 268	434	0,24%
Консульская регистрация	423 682	3 015	0,71%
RUMV	1.048.000	10 885	1,04%
RUTEC	51 465	1 193	2,32%
Визы	222 193	13 188	5,94%

Таблица 7. Список миграционных RRAA с числом дубликатов

41. Впоследствии все таблицы были объединены вертикально для окончательного поиска дубликатов детерминированными и вероятностными методами. После очистки базы данных путем выявления и извлечения дубликатов по типу документа и номеру документа дубликаты идентифицируются с использованием вероятностного метода. Сводные результаты представлены в таблице Table 8. В приложении Annex 1 представлен алгоритм, который был реализован в среде для совместной работы Google.

Обработка таблицы миграции	*Регистры*
Объединенная таблица	6 258 191
Детерминированные дубликаты	-256 584
Вероятностные дубликаты	-469 813
Все записи о миграции	**5 531 794**

Таблица 8. Объединение уточненных RRAA, Таблица миграции

F. Этапы обработки включают пять этапов:

42. Очистка переменных: Наиболее распространенные несоответствия данных связаны с вариантами написания имен, такими как прозвища (например, Роб и Роберт), форматами даты рождения, кодировкой данных и отсутствующей информацией. Методы, используемые для устранения этих несоответствий для преобразования данных в стандартные формы, включают: редактирование, нормирование, дедупликацию и сопоставление.
43. Нормирование — это форматирование элементов данных таким образом, чтобы они были единообразно представлены во всех наборах данных. Например, даты могут быть представлены в разных форматах, таких как ДД/ММ/ГГГГ или ММ/ДД/ГГГГ. Необходимо выбрать один формат и использовать его для всех наборов данных проекта.
44. Связывание данных; включает в себя множество сравнений записей. В идеале каждая запись в наборе данных A сравнивается с каждой записью в наборе данных B, чтобы определить, какие пары записей с наибольшей вероятностью являются связями. Однако если каждую запись сравнивать между двумя наборами данных, содержащими по 100 000 записей в каждом, потребуется 10 миллиардов сравнений. Даже при использовании хороших вычислительных мощностей это заняло бы значительное время.
45. Чтобы сэкономить время, используется «блокировка», позволяющая уменьшить количество сравнений записей, необходимого для поиска потенциальных пар записей. Например, при использовании пола для блокировки сравниваются только записи одного и того же пола (мужской или женский), что обычно вдвое сокращает количество необходимых сравнений. Однако пол не слишком полезен для блокировки, так как он просто разделяет набор данных на два больших блока, поэтому все еще требуется много сравнений.
46. В идеале стратегия блокировки должна генерировать небольшие блоки одинакового размера в каждом наборе данных. Например, использование месяца рождения приведет к 12 блокам (по одному на каждый месяц) и, как ожидается, будет иметь четное количество записей в каждом блоке. Распространенная стратегия заключается в том, чтобы поддерживать как можно меньший размер блока и, таким образом, сравнивать как можно меньше записей.
47. В нашем случае используются следующие переменные: Имена, Фамилии и Дата рождения. Просто взглянув на гистограмму задействованных переменных, можно было установить, что фамилии образовывали меньшие группы сравнения, таким образом, титаническая задача провести сравнение с использованием декартова произведения записи на n-1 регистры, оставшиеся в таблице превратилась в сравнение только тех записей, в которых одинаковая фамилия.

48. Конфигурация Splink; это было сделано в схеме Json. В этой конфигурации параметры по умолчанию в библиотеке Splink могут быть изменены. В этой схеме переменная «фамилия» устанавливается как условие, генерирующее блоки, указанные в предыдущем пункте (правило блокировки). В типе привязки выбран параметр "dedupe_only", который работает только для выявления дубликатов в пределах одного и того же RRAA, таким образом настраиваются столбцы сравнения, которыми в данном случае являются "ФИО" и "Дата рождения". Кроме того, для работы алгоритма необходима переменная-идентификатор, уникальная для каждой записи, таким образом можно определить соответствие по паре строк.
49. При настраивании сравнения по именам были установлены три уровня сравнения, где:
    - Уровень 2: Сравниваемые строки одинаковые
    - Уровень 1: Сравниваемые текстовые строки схожи
    - Уровень 0: Сходства текстовых строк нет.
50. Дата рождения не настраивается, но по умолчанию существует 2 уровня:

- Уровень 1: Сравниваемые текстовые строки одинаковые.
- Уровень 0: сравниваемая текстовая строка отличается.

51. Сравнение может быть выполнено многими различными методами для вычисления значений сходства для строки текста, числовых значений или дат. В нашем сценарии, где мы вычисляем показатель сходства для значений текстовой строки, использовался алгоритм Джаро-Винклера, который представляет собой гибрид между Q-граммами и расстоянием Левенштейна.
52. Установить порог классификации. После запуска алгоритма EM для оценки вероятностей m(y) и u(y) каждой пары записей в каждом блоке были получены вероятности совпадения по имени, вероятности совпадения по дате рождения и общие вероятности совпадения между записями. Все пары записей, общая вероятность совпадения которых была больше или равна 0,85, были сохранены.
53. На странице 9 статьи «Using a Probabilistic Model to Assist Merging of Large-Scale Administrative Records» (Romántico, 2019) автор определил, что при вероятности сходства 85% или выше их можно считать равными.

V. Вероятностное предсказание пола по имени

54. В процессе консолидации REBPM были идентифицированы переменные с недопустимыми, непоследовательными или отсутствующими значениями, поэтому их необходимо обработать, чтобы преобразовать и улучшить их качество. Далее будут упомянуты процедуры, применяемые к переменным административных записей, которые должны быть улучшены для статистического использования, что облегчит связь между доступными источниками информации.
55. В частности, административные записи миграционной службы Колумбии (Migración Colombia) не содержат переменной пола, однако в интегрированном регистре есть несколько лиц, помимо тех, которые поступают из баз данных миграционной службы, для которых нет информации о поле. Чтобы восполнить эту отсутствующую информацию, была предложена следующая методика:

- Использовать методы машинного обучения или (Машинное обучение)

- Изучить проверенную информацию об именах и фамилиях записей, в которых есть эти данные.

56. Для импутации переменной пола в записях без информации в этом случае было рассмотрено 1 835 093 записи без информации. Для этого выполняется этап очистки, который описан ниже:

- Отсеивались люди с именами из 5 и более слов.
- Специальные символы, такие как (\,*,^,...), были удалены, поскольку они связаны с опечатками и, если их не удалить, приводят к искажению результатов.
- Цифры от 0 до 9 были удалены.
- Некоторые артикли были удалены, например: *«эль», «дель», «де», «лос», «лас».*
- Имена, содержащие менее 3 букв, были удалены.
- Все строчные записи сохранены.

57. Общий вид итоговой базы для обучения модели имеет следующие характеристики:

- Выбираются наблюдения, имеющие только одно имя; это будет достаточно важно в модели, поскольку позволяет различать записи людей, которые сообщают только имя в любой из записей.
- У людей с 2 или более именами они были объединены в одно слово. Такое расположение имен имеет несколько преимуществ, поскольку таким образом модель получает необходимые инструменты для различения различных комбинаций имен, которые могут быть образованы из личных имен.

58. Выбор этой структуры не случаен, объединение имен в одно слово и помещение их в один пакет с другими добавляет модели изменчивости, что дает больше информации и доказательств для классификации человека, которого зовут, например, «Мария Хосе», и которого можно правильно классифицировать как женщину.

59. Как только все необходимые элементы для обучения и прогнозирования будут доступны, будут определены новые переменные, сгенерированные из переменной «имя», которые будут использоваться для обучения модели. Для этого они определили:

    i. Две переменные, каждая с первыми 3 и 5 буквами каждого имени.
    ii. Две переменные, каждая с последними 3 и 5 буквами каждого имени.
    iii. Переменная, которая идентична, если имя заканчивается на гласную.
    iv. Переменная с длиной символа каждого имени.
    v. Переменная, определяющая, заканчивается ли имя открытой гласной, что особенно полезно для различения мужчин и женщин.

60. Дополнительно создается весовая переменная «вес», позволяющая решить проблему дисбаланса классов, возникающего при наличии меньшего количества сочетаний имен у мужчин, то есть больший вес придается категории с меньшей частотностью. Существуют и другие методологии, которые можно использовать для компенсации проблемы дисбаланса, например, взятие части выборки категории с наибольшей частотой таким образом, чтобы категории были сбалансированы, однако они не рассматривались.

61. В результате есть доказательства большего числа сочетаний составных имен (2 имени), что соответствует женщинам, в то время как мужчины имеют в основном одно имя или меньшее количество сочетаний имен.

A. Прогнозирование гендерной переменной

62. Наконец, импутация переменной пола была выполнена для 1 835 093 человек, в результате чего было получено 627 177 женщин и 1 207 916 мужчин. Очевидно, что численность мужчин выше, и подтверждение этих результатов произойдет после того, как Migración Colombia предоставит информацию о лицах с отсутствующей переменной.

VI. Некоторые результаты

63. Одним из наиболее важных аспектов текущей работы является возможность охарактеризовать иммигрантов на территории Колумбии. Визуализация этого населения для целей формирования государственной политики, которая позволит им интегрироваться в общество, должна обеспечиваться посредством своевременной и качественной информации. Для этой цели на странице DANE[3] было создано средство просмотра, в котором пользователи могут просматривать информацию, которую можно получить из административных записей.

64. Некоторые результаты, полученные с помощью Базового статистического регистра мигрантов, представлены ниже. Первый результат — это описание по возрасту и полу людей из Венесуэлы и других частей света. Отмечено, что среди иммигрантов больше мужчин трудоспособного возраста. При подробном анализе возрастно-половой пирамиды венесуэльцев видно, что группа от 5 до 9 лет является самой представленной среди лиц моложе 20 лет. Что касается группы лиц трудоспособного возраста старше 20 лет, то наиболее представленной является группа от 25 до 29 лет.

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

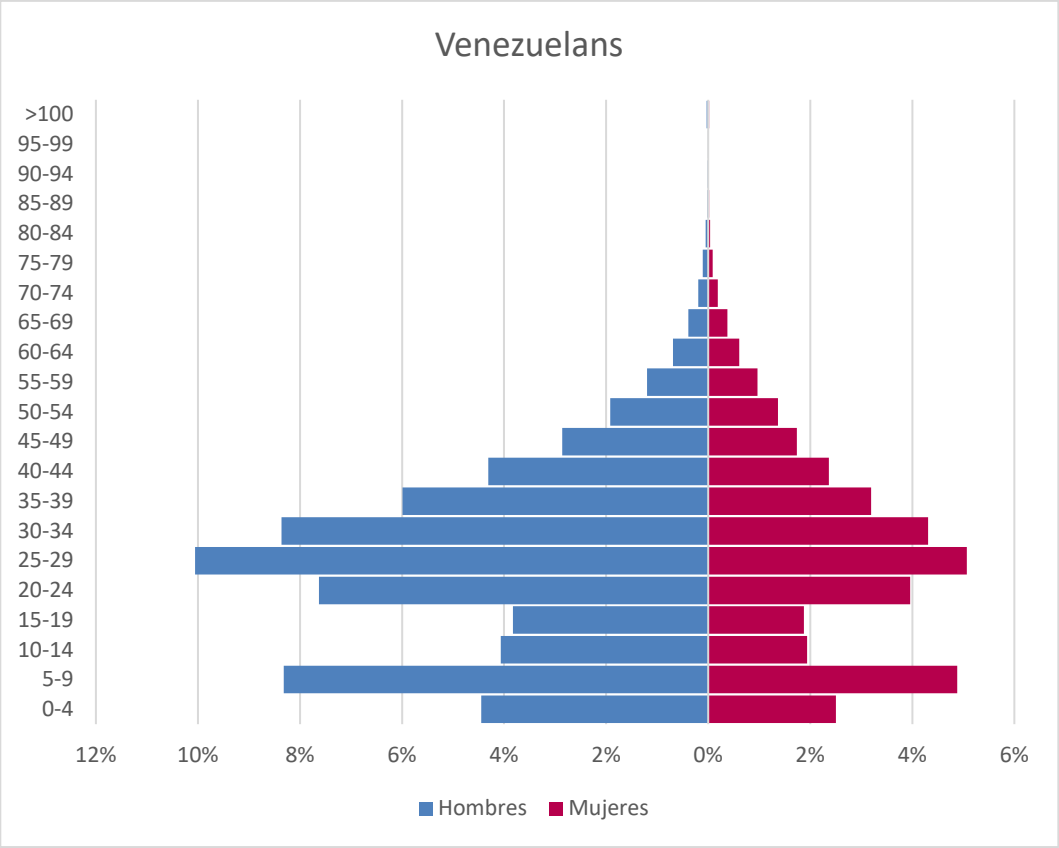

Рисунок 1. Распределение населения – иммигранты из Венесуэлы

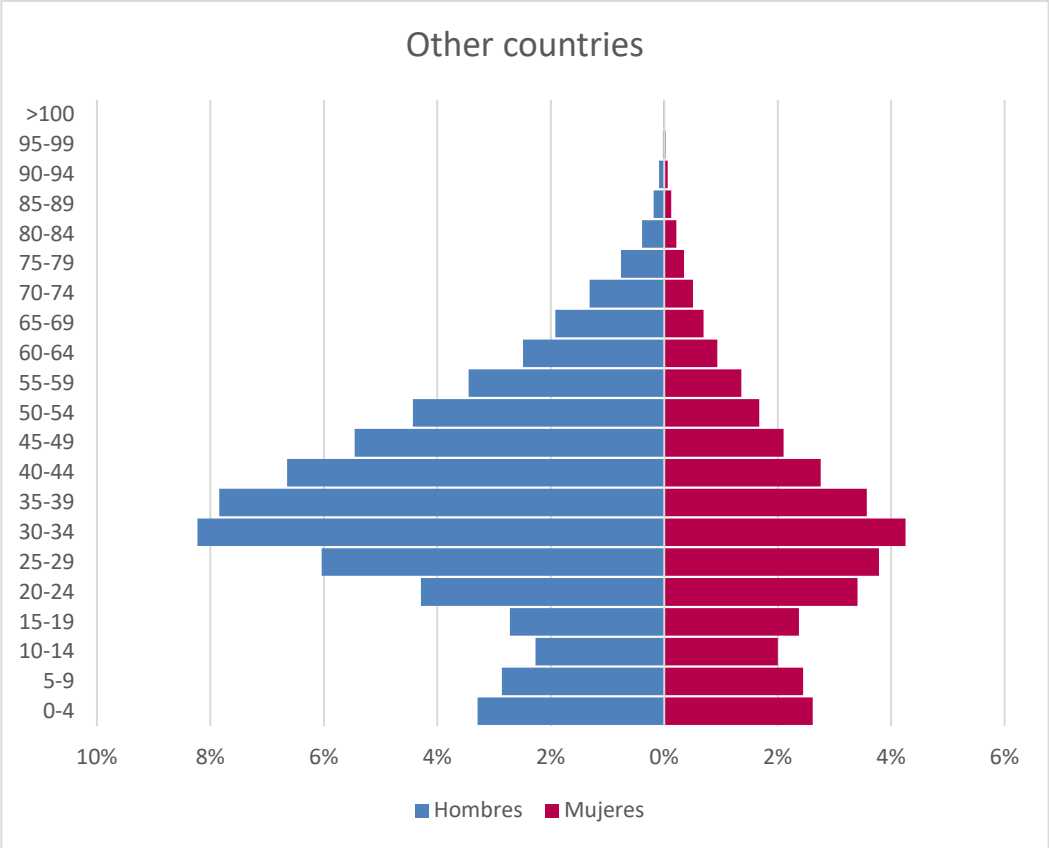

Рисунок 2. Распределение населения – иммигранты из стран кроме Венесуэлы

VII. Приложение 1

65. Алгоритм выявления вероятностных дубликатов

```
[('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. Среда была настроена таким образом, чтобы использовалась полная мощность узлов и они работали параллельно. Для данной конфигурации использовалась библиотека splink Robin [2020].