16 October 2018

Европейская экономическая комиссия ООН

Конференция европейских статистиков
Рабочая сессия по статистике миграции
Женева, Швейцария
24-26 октября 2018 года
Пункт 3 предварительной повестки дня
Интеграция данных переписей, административных источников и обследований для измерения миграции

Оценка численности беженцев из данных обследований с учетом административных данных

Записка Бюро переписи населения США*

Резюме

Поток беженцев и соискателей приобретает все большее значение в международной миграции. Несмотря на осведомленность, в Соединенных Штатах мало известно об этих мигрантах после их въезда в страну. Бюро переписи населения США не собирает данные в своем крупнейшем обследовании - Обследовании Американского общества (ACS), которое может напрямую идентифицировать беженцев и соискателей. Оценка численности этих групп населения получена из модели логистической регрессии с помощью набора данных, который содержит ограниченную демографическую информацию о лицах, получивших юридический статус постоянного жителя в Соединенных Штатах. Эта модель затем используется для прогнозирования вероятности того, что лицо иностранного происхождения будет выявлено в ACS как беженец или соискатель. Для создания двоичной переменной присвоений, указывающих статус беженца/соискателя применяется алгоритм повторного отбора-отказа. Используя эти присвоения, можно использовать ACS для более детального изучения демографических характеристик беженцев/соискателей. Этот документ служит демонстрацией того, как многочисленные наборы административных данных могут использоваться для получения большей информации из меньшего, но более детального набора исследований.

* Доклад подготовлен Michael Bowerman, Статистик, Отдела международного миграционного прироста, Бюро переписи США.

I. Введение

1. После принятия Закона о беженцах 1980 года Государственный департамент США начал публиковать данные о беженцах, прибывших в Соединенные Штаты. С того времени Государственный департамент зарегистрировал почти три миллиона беженцев. В период между 1990 и 2016 годами более 600 000 человек получили убежище в Соединенных Штатах [1]. Согласно Программе оценки численности населения Бюро переписи, Соединенные Штаты приняли чистый поток мигрантов в 1,1 миллиона человек в период с 1 июля 2016 года по 1 июля 2017 года [2]. В 2016 финансовом году примерно 105 000 беженцев и соискателей получили законный статус постоянного жителя в Соединенных Штатах.

2. Хотя такое большое количество беженцев и соискателей въехало в Соединенные Штаты, мало известно об этом населении после их приезда. Бюро переписи населения США не включает информацию о статусе беженца или соискателя в большинство своих обследований и переписей, за исключением проведения обследования доходов и участия в государственных программах (SIPP). К сожалению, SIPP не имеет достаточно данных для надежной оценки группы, размер которой мал по сравнению с общей численностью населения Соединенных Штатов. Из-за ограниченной информации результаты пребывания этих людей в Соединенных Штатах не известны. Мотивация этого исследования заключается в том, чтобы иметь возможность использовать обследование Американского общества (ACS), которое имеет гораздо больший размер выборки данных, чем SIPP, для оценки контингента беженцев и соискателей в Соединенных Штатах. Для достижения этой цели при подготовке модели логистической регрессии был использован большой административный набор данных, содержащий всех лиц, получивших статус постоянного жителя в Соединенных Штатах в период с 2010 по 2015 год, с тем чтобы предсказать вероятность того, что согласно ACS какое-либо лицо прибыло в качестве беженца или соискателя. После того, как эта прогнозируемая вероятность была получена из модели логистической регрессии, был использован алгоритм повторного отбора-отказа для присвоения статуса беженца как двоичной переменной. Исходя из этого присвоения можно было бы оценить данные о группе населения беженцев и соискателей в Соединенных Штатах.

II. Обследование американского общества (ACS)

3. Основным источником данных Бюро переписи, используемым для составления ежегодной оценки численности чистого потока мигрантов США, является ACS, представляющее собой ежегодный опрос домохозяйств, размер выборки которого составил пять миллионов человек в 2016 году [3]. Для идентификации иммигрантов через ACS обычно используются четыре переменные: место рождения, место жительства год назад, год въезда в Соединенные Штаты и статус гражданства. Например, лица иностранного происхождения, имеющие гражданский статус натурализованного гражданина или негражданина, считаются частью иностранного контингента. Поток иностранного происхождения оценивается путем отбора лиц, идентифицированных в ACS как лица иностранного происхождения, и чье место жительства год назад было за границей или те, которые указали предыдущий год как год въезда в Соединенные Штаты. В ACS есть множество переменных, которые могут представлять интерес для изучения контингента беженцев, включая подробную демографическую, экономическую и географическую информацию. Тем не менее, в ACS нет вопроса, который непосредственно идентифицирует беженца или соискателя. Чтобы получить доступ к

этому богатому источнику данных при оценке населения беженцев, необходимо использовать внешний источник данных для присвоения статуса беженца в выборке ACS.

III. Анкета лица, законно получившего вид на жительство (LPR)

4. Когда физические лица получают вид на жительство в Соединенных Штатах либо путем прямого въезда в качестве законных постоянных жителей посредством подачи заявления в Государственный департамент, либо путем перехода от временного к постоянному статусу посредством подачи заявления в Министерство внутренней безопасности (DHS), данные о них заносят в LPR-анкету DHS, которую Бюро переписи получает из Отдела иммиграционной статистики. Однако, данные об этих лицах неполные; LPR содержит мало переменных, и есть проблемы с отсутствующими данными. Переменные, включенные в LPR, подробно описаны в таблице 1.

Таблица 1. Переменные анкеты LPR

Имя переменной	Детали
Вид разрешения на въезд	Вид визы, по которой въехал человек, - используется для разделения лиц на беженцев/соискателей и других
Страна рождения	Место рождения, включая отдаленные территории США, переведено в коды FIPS*
Страна гражданства	Коды этой переменной почти идентичны переменной "Страны рождения"
Страна последнего места жительства	Коды этой переменной почти идентичны переменной "Страны рождения"
Профессия	Слабозаполненная переменная
Пол	
Семейное положение	Записан как холост, женат, проживающий отдельно, разведен или овдовевший
По адресу	Почтовый индекс и название государства, куда переехал мигрант
Дата рождения	Год рождения
Дата въезда	Дата въезда в Соединенный Штаты

*Федеральный стандарт обработки информации (FIPS), уникальный код страны. Подробности здесь: https://www.census.gov/geographies/reference-files/2016/demo/popest/2016-fips.html

5. Данные LPR-анкеты накапливаются каждый финансовый год из тех, которые собраны Государственным департаментом США по выданным постоянным визам и DHS по корректировкам из временного статуса. В этом анализе использовались анкеты LPR с 2012-2015 финансовых годов, включающие 4,1 млн человек. Используя в LPR переменную "Вид разрешения на въезд", можно легко идентифицировать беженцев и соискателей в наборе данных.

6. Как было сказано ранее, LPR содержит незначительные сведения о лицах в анкете и, кроме того, здесь не хватает значительного количества данных для некоторых переменных. Во-первых, значительная часть переменной профессии отсутствует. Поскольку большая часть переменной профессии отсутствует, ее нельзя использовать для моделирования вероятности того, что человек является беженцем, что является неудачным, поскольку эта переменная может быть сопоставлена с ACS. Не хватает также 0,2% переменной пола; эти записи будут удалены из анализа, потому что кажутся случайными. Мы также удалим людей с отсутствующим возрастом или семейным

положением, так как они составляют небольшую часть анкеты. Других отсутствующих записей для переменных, представляющих интерес, не было выделено.

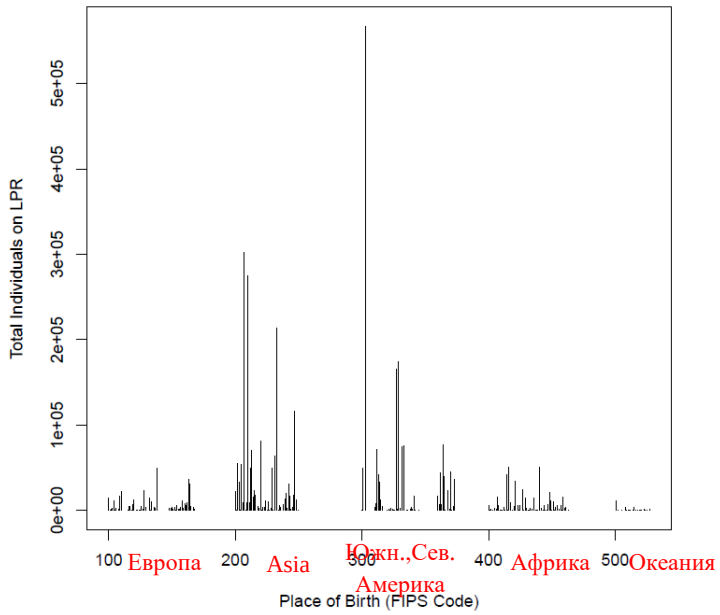
IV. Идентификация беженцев/соискателей в ACS с использованием LPR

7. В LPR существует пять переменных, которые могут быть информативными при прогнозировании статуса беженца или соискателя: возраст, пол, семейное положение, место рождения и год въезда. Другие переменные оказались малоинформативными, как в случае с профессией или даже неразличимыми, как данные о месте рождения в случае указания гражданства и страны предыдущего проживания. Используя эти переменные, модель логистической регрессии была приспособлена к данным LPR для прогнозирования вероятности того, что человек является беженцем или соискателем. Затем бинарный результат лица, являющегося беженцем или соискателем, моделируется с использованием алгоритма повторного отбора-отказа. Однако, прежде чем пытаться использовать данные LPR, необходимо обработать данные переменных.

A. Место рождения

8. В LPR представлено 216 разных стран рождения. Гистограмма LPR, показывающая количество лиц по месту рождения, представлена на рисунке 1. Данные о стране рождения, указанные в LPR, были слишком незначительными, чтобы вызвать озабоченность относительно репрезентативности выборки. Например, количество записей беженцев и соискателей было менее тридцати у 75 стран в LPR. Из-за этого использование переменной страны рождаемости при моделировании вероятности того, что лицо иностранного происхождения попало в Соединенные Штаты как беженец или соискатель, могло ввести в заблуждение.

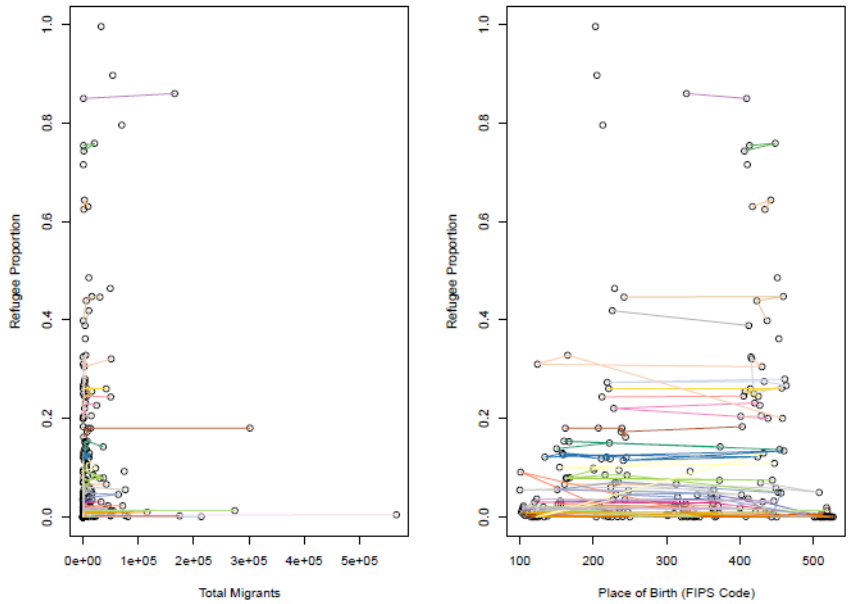
Рисунок 1. Гистограмма количества случаев по месту рождения в LPR, по FIPS-коду



9. Вместо того, чтобы использовать исходное место переменной рождаемости, в анкете LPR мы сгруппировали страны на основе общего количества записей LPR из каждой страны в сочетании с количеством записей беженцев из этой страны, используя полупараметрическую бета-биномиальную модель с предшествующим распределением Дирихле. Получение последующего образца проводилось с использованием R package DPpackage [4].

10. Эта модель объединяет страны рождения, основанные на общем числе лиц из этой страны в LPR и числе беженцев и соискателей из этой страны, по сути разделяя страны по вероятности приезда оттуда беженцев и соискателей в Соединенные Штаты. По модели был создан последующий образец, разделенный на 37 групп. Таким образом, вместо моделирования, основанного на месте рождения, для которого потребуется 215 дополнительных переменных, мы моделируем на основе группового присвоения, требуя только 36 дополнительных переменных. Присвоения по группам показаны на рисунке 2; точки на графике представляют каждую страну, представленную в LPR, а на линиях показаны страны, связанные в каждой группе.

Рисунок 2. Присвоение по группам стран рождения в LPR

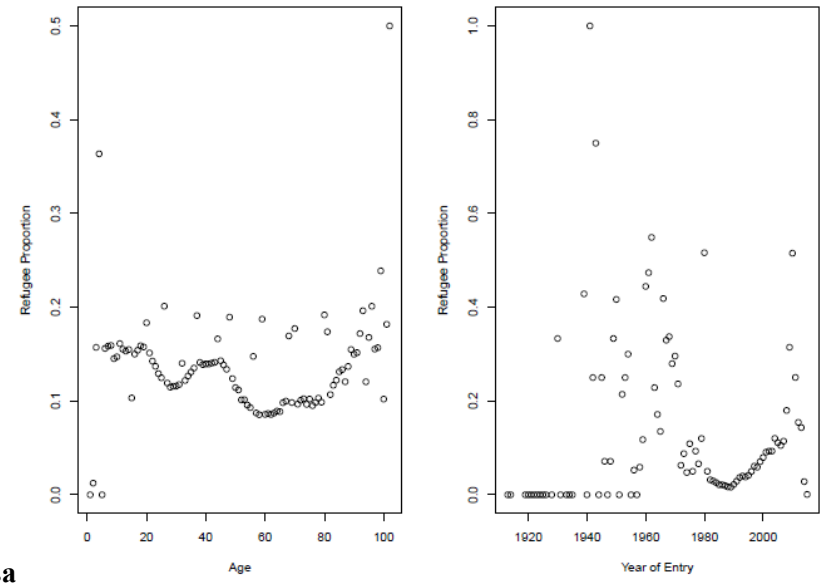


B. Возраст и год въезда

11. На рисунке 3 показаны графики взаимоотношений между возрастом и годом въезда и часть LPR, в которой указаны беженцы или соискатели. Модели с течением времени для этих связей не показывают четкую форму, которая может быть урегулирована путем преобразования данных. Простое моделирование года и возраста как переменных вида рискует привести к чрезмерной аппроксимации, а моделирование их как необработанных числовых переменных является неприемлемым - связь между возрастом или годом въезда и вероятностью быть беженцем/соискателем не может быть смоделирована по прямой. Вместо этого соответствующие отношения между возрастом, годом въезда и вероятностью того, что лицо является беженцем или соискателем, будут включены в модель с использованием натуральных кубических сплайнов с 20 узлами или кусочный полином в том же диапазоне, что и исходные переменные, где отдельный кубический

полином устанавливается между каждой из 20 узловых точек. Вместо того, чтобы требовать около 100 видов в год и 100 видов по возрасту, вероятность беженца/соискателя будет смоделирована на основе полинома, соответствующего короткому периоду года и возраста.

Рисунок 3. Возраст и доля беженцев слева; Год въезда и доля беженцев справа



C. Модель логистической регрессии

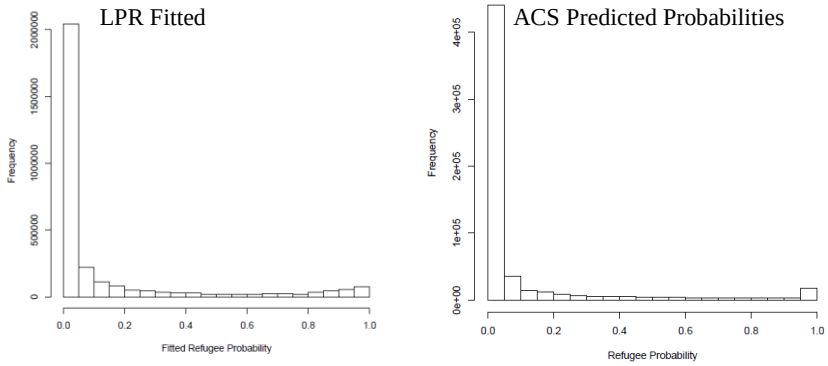
12. После завершения предварительной обработки данных модель логистической регрессии соответствовала данным LPR, моделируя вероятность беженца/соискателя по полу, семейному положению, групповому присвоению назначению в зависимости от страны рождения, возраста и года въезда в Соединенные Штаты:

$$logit(p_i) = sex_i + mar_i + cluster_i + S_{age}(age_i) + S_{year}(year_i) + \epsilon_i,$$

где p_i – это, возможно, лицо, i – это беженец или соискатель, $S_{age}(age_i)$ – это значение натурального кубического сплайна при значении age_i, возраст лица i, и $S_{year}(year_i)$ – это значение функции сплайн при $year_i$, и ϵ_i – это $N(0, \sigma^2)$ случайная ошибка. Эта модель была лучше, чем вложенные модели. Гистограмму установленных вероятностей, что каждый человек в LPR является беженцем/соискателем, можно найти на рисунке 4 (слева).

13. Эта модель логистической регрессии была применена к 1-летней ACS в 2015 году. Лица иностранного происхождения были выбраны из образца ACS, используя переменные гражданство и страна. Затем лица были сгруппированы по стране рождения в те же группы, которые использовались для сопоставления логистической модели с данными LPR. Используя модель, рассчитывалась вероятность того, что каждый человек в ACS является беженцем. Гистограмма с установленными вероятностями для ACS приведена на рисунке 4 (справа).

Рисунок 4. Установленные вероятности того, что человек является беженцем или соискателем по LPR (слева) и ACS (справа)



D. Повторный отбор-отказ

14. Предсказываемая вероятность того, что согласно ACS лицо является беженцем или соискателем, была получена из модели логистической регрессии, но решение по-прежнему должно быть принято независимо от того, является ли особа беженцем. Для этого использовался алгоритм повторного отбора-отказа. Это позволяет избежать простого отсечения при некоторых значениях, на что повлияет ошибка модели. Этот алгоритм выполняется следующим образом:

Для лица i :

1. Соотношение a получается из $Uniform(0,1)$
2. Если $\hat{p}_i > a$, где $\hat{p}_i$ – это is the установленная вероятность для лица i , образец принимается
3. Повторите шаги 1 и 2 для 1000 итераций
4. Если $\frac{number\ of\ accepted\ samples}{number\ of\ iterations} > .5$, лицо помечается как беженец/соискатель

15. Чтобы проверить этот подход, из LPR берется образец, используемый для соответствия исходной модели логистической регрессии. В этом случае данные об одном миллионе из 4,1 млн. человек были сохранены как тестовые. Модель логистической регрессии применялась снова с использованием оставшихся 3,1 миллионов индивидуальных обучающих наборов. После этой подгонки предсказанные вероятности вычислялись с использованием тестового набора, и алгоритм повторного отбора-отказа использовался для обозначения беженцев в тестовом наборе. Используя этот метод, было обнаружено, что только 92,9% лиц из тестового набора был присвоен статус беженца или не беженца правильно. Затем этот алгоритм проводился для каждого лица иностранного происхождения в ACS, чтобы присвоить статус беженца.

V. Результаты

16. В рамках одногодичного опроса в 2015 году среди лиц иностранного происхождения были выделены группы беженцев или не беженцев, используя метод повторного отбора-

отказа, и предсказана вероятность беженца с помощью модели логистической регрессии. Затем была произведена оценка численности в национальном масштабе с использованием весовых коэфициентов из ACS.

17. Некоторые интересные различия были отмечены в демографических характеристиках группы беженцев или соискателям по ACS и всех родившихся за рубежом. Некоторые из этих различий предсталены в таблице 2.

Таблица 2. Различия между группами, родившимися за границей по ACS 2015 года, приписанными к беженцам/соискателям и другим лицам иностранного происхождения

Процентов в категории										
	Пол		Возраст			Семейное положение				
	Мужской	Женский	<25	25 to 65	>65	В браке	Вдова/вдовец	В разводе	Проживающий отдельно	Не был в браке
Беженцы/соискатели	39%	61%	20%	38%	42%	39%	16%	9%	3%	33%
Другие лица иностранного происхождения	48%	52%	14%	70%	16%	60%	5%	7%	2%	25%

18. Руководствуясь ACS видно, что группа беженцев больше других состоит из женщин. Их распределение по возрасту также представлено более широко, тогда как подавляющее большинство в группе, не являющейся беженцами/соискателями, составляет от 25 до 65 лет. Беженцы и соискатели, скорее всего, овдовевшие, а не беженцы/соискатели скорее всего состоят в браке.

19. Из Ежегодника DHS по статистике иммиграции в 2015 году 590 000 человека получили убежище в Соединенных Штатах в период с 1990 по 2015 год. С 1990 года по 2015 год в Соединенные Штаты въехали 1 907 000 беженца, в общей сложности 2 517 000 беженцев и соискателей въехали в Соединенные Штаты за эти два периоды. Используя описанный выше метод и весовые коэффициенты в ACS, по оценка численнсти в 2015 году в Соединенных Штатах проживали 1 667 000 беженцев и соискателей, приехавших после 1990 года. Необходимо так же принимать во внимание смерность и низкий уровень эмиграции. Пользуясь данным методом при подсчетах, было вычснено, что в 2015 году в Соединенных Штатах в общей сложности проживали 2 869 000 беженцев и соискателей. Согласно нашей оценке, 6,6% контингента иностранцев, родившихся в Соединенных Штатах в 2015 году, по оценкам из ACS, были беженцами и соискателями.

20. Гистограмма года въезда людей, которые, как прогнозировалось, были беженцами или соискателями, согласно ACS, отображена на рисунке 5. На ней видны пики поступления беженцев и соискателей после Второй мировой войны и снова после последующих войн в Ираке и Афганистане.

21. Распределение стран рождения значительно отличается у двух пиков. До 1970 года большинство беженцев и соискателей приезжали из Европы, в частности из Германии, Италии, Австрии и Мексики. После 1980 года эта тенденция переместилась на большинство беженцев и соискателей, прибывающих с Ближнего Востока, после войн в Афганистане, Ираке и на Кубе. Заметно отсутствие беженцев и соискателей, которые прибыли с 1960-х до середины 1990-х годов, особенно таковых из Юго-Восточной Азии и бывших республик Советского Союза. Вероятно, это связано с тем, что они не представлены в версиях LPR, доступных для этого исследования. Пик, относящийся к 1950-м годам выглядит довольно странно, так как большинство людей в LPR прибыли в США после 2000 года. Есть две вероятные причины этого всплеска: во-первых, лица в LPR, которые прибыли до 1960 года, часто были беженцами и соискателями, соответственно лица, чей год въезда по ACS был до 1960 года, также были приписаны к

беженцам или соискателям; во-вторых, лица, эмигрировавшие из Европы, имели демографические профили, которые соответствовали беженцам и соискателям, приехавшим позже.

Рисунок 5. Год въезда в Соединенные Штаты лиц согласно ACS 2015 года, являющихся по прогнозам беженцами или соискателями; плотность[1] синим цветом

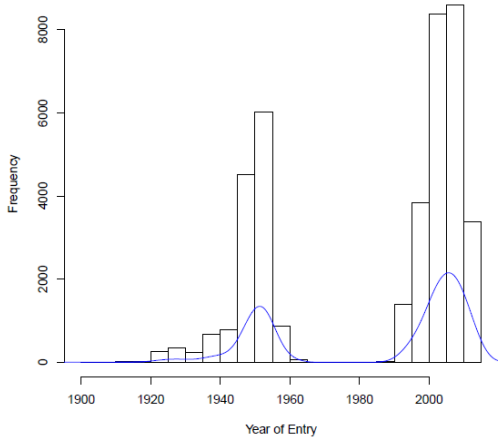
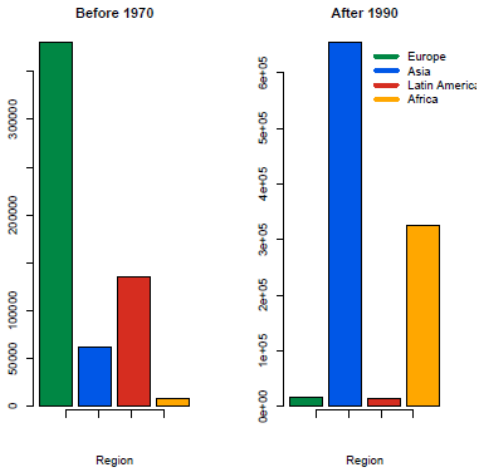


Рисунок 6. Место рождения лиц согласно ACS 2015 года, которым присвоен статус беженца или соискателя, по времени прибытия



22. Далее, на рисунке 7 показано сравнение двух графиков. На первом черным цветом показана плотность частоты регионов рождения лиц со статусом беженца, присвоенного ACS, по тому принципу, чей приезд пришелся на период между 2010 и 2013 годами - эти годы выбраны, потому что лучше всего отображают население иностранного происхождения из выборки годов LPR, используемой для соответствия модели логистической регрессии. Синяя кривая - это плотность частоты регионов гражданства лиц согласно Всемирной системе обработки данных о приеме беженцев (WRAPS)2

[1] Эмпирическая функция распределения вероятностей.

[2] WRAPS представляет собой компьютерную систему, используемую для обработки и отслеживания беженцев, которые переселены в Соединенные Штаты в рамках Программы приема беженцев в США.

которые были переселены в США в период с 2010 по 2013 год [6]. Распределения, опять же, соответствуют, и это, в свою очередь, предполагает, что метод дает достоверные оценки для контингента беженцев в Соединенных Штатах.

Рисунок 7. Плотность частоты регионов происхождения согласно WRAPS и ACS

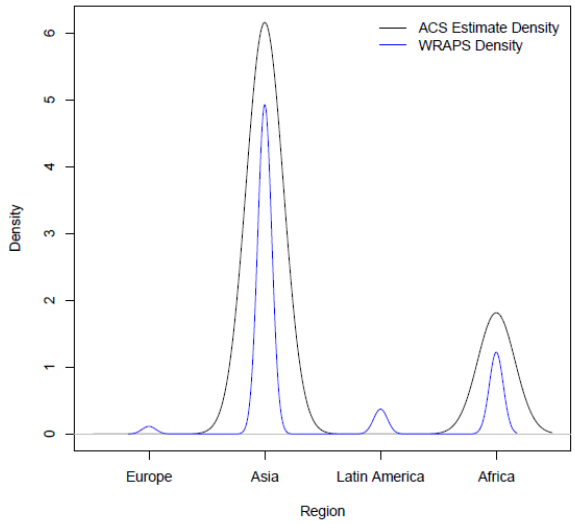


Таблица 3. Упорядоченная таблица наиболее распространенных мест рождения для оценок численности приписанных беженцев/соискателей и не беженцев/не соискателей по ACS

Топ мест рождения беженцев/соискателей	Количество беженцев/соискателей	Топ мест рождения не беженцев/не соискателей	Количество не беженцев/не соискателей
Куба	699,700	Мексика	11,561,000
Китай	178,300	Индия	2,387,000
Ирак	143,100	Филиппины	1,975,000
Мьянма	112,900	Китай	1,928,000
Германия	111,600	Эль Сальвадор	1,349,000
Мексика	91,140	Вьетнам	1,299,000
Эфиопия	74,040	Доминиканская Респблика	1,059,000
Сомали	70,330	Южная Корея	1,057,000
Таиланд	70,120	Гватемала	925,500
Италия	53,930	Канада	764,000

23. В таблице 3 показаны страны рождения самого большого числа беженцев и соискателей по сравнению со странами, не относящимися к беженцам, и не соискателям, оцененными данным методом по ACS. Эти два списка довольно разнородны, причем только Мексика и Китай появляются в обоих списках - хотя это, вероятно, связано с большим количеством людей, родившихся в Мексике и Китае, в соответствии с ACS. Некоторые заметные страны фигурируют в списке лиц, не являющихся беженцами/не соискателями, что оказалось неожиданным, в том числе Эль Сальвадор, Доминиканская Республика и Гватемала. По числу лиц из этих стран, ищущих убежище в Соединенных Штатах, наблюдались всплески в последние годы [5], поэтому мы ожидали увидеть их среди самых частых мест рождения беженцев/соискателей. Возможно, эти лица, как соискатели убежища, еще не появились в LPR, поскольку в наборе данных присутствуют только лица, получившие убежище. Они появятся в списках, так как им предоставлен статус соискателей в Соединенных Штатах, так что, метод оценки не прогнозирует, что лица, местом рождения которых является одна из этих стран, являются беженцами или соискателями. Опять же, Италия, Германия и Мексика являются одними из самых частых стран происхождения беженцев/соискателей, что может оказаться неверным. Необходимо

больше исследований, чтобы понять, правильно ли метод присваивает этим лицам статус беженца/соискателя. Для устранения ошибочной классификации может потребоваться включение большего количества наборов данных, чтобы выяснить, ошибочно ли они классифицированы.

VI. Вывод

24. Метод, описанный здесь, объединяющий модель логистической регрессии с использованием алгоритма повторного отбора-отказа, дает оценку популяции беженцев и соискателей из Обзора Американского общества, которая согласуется с данными о допуске беженцев и соискателей и данными о переселении беженцев в Соединенных Штатах. Особенная ценность состоит в том, что для получения оценок из гораздо более подробного ACS можно использовать редкий источник данных, такой как анкета лица, законно получившего вид на жительство, которая дает возможность сделать подробные демографические и экономические оценки численности населения беженцев и соискателей, которые в настоящее время проживают в Соединенных Штатах. Кроме того, этот метод дает оценку численности популяции беженцев и соискателей, тогда как LPR предоставляет только данные о потоках беженцев и соискателей.

25. Дальнейшие исследования могут изучить возможность получения доступа к другим наборам административных данных непосредственно из Государственного департамента для оценки других групп, которые не могут быть непосредственно опрошены посредством ACS. Более важным для этого документа, в частности, было бы включение ранних лет LPR для тестирования модели логистической регрессии - эта модель была тестирована на наборах данных LPR с 2012 по 2015 год, хотя наборы данных существуют и с предыдущих лет. Это может повысить точность оценок численности, особенно что касается волны беженцев после Второй мировой войны и в 1970-х годах, где было отмечено, что многим иностранцам из Италии, Германии и Мексики были присвоены статусы беженцев/соискателей, возможно, неправильно, а беженцы из Юго-Восточной Азии в 1970-х годах были упущены.

VII. Использованная литература

[1] Office of Immigration Statistics. 2016. *2015 Yearbook of Immigration Statistics*. Department of Homeland Security.

[2] Population and Housing Unit Estimates. 2018 https://www.census.gov/programs-surveys/popest/data/tables.html. US Census Bureau, Department of Commerce.

[3] United States Census Bureau. 2018. https://www.census.gov/programs-surveys/acs/about.html. Department of Commerce.

[4] Alejandro Jara, Timothy Hanson, Fernando Quintana, Peter Mueller, Gary Rosner. 2011. *DPpackage: Bayesian Semi- and Nonparametric Modeling in R*. Journal of Statistical Software, 40(5), 1-30.

[5] UNHCR Population Statistics. 2018. http://popstats.unhcr.org/en/asylum_seekers_monthly. The UN Refugee Agency.

[6] Refugee Processing Center. 2018. http://ireports.wrapsnet.org/. Department of State.