Сетевые Модели

Рогоза Ярослав Э308
r.yaroslav1w@gmail.com
Московский государственный университет имени М. В. Ломоносова - Экономический факультет

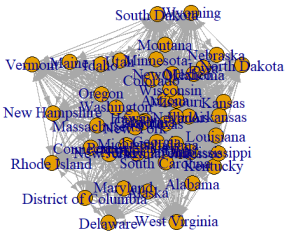
1 Введение

В своём проекте мы хотим провести анализ миграционных потоков среди штатов США. В качестве данных для своего проекта мы выбрали базу данных Мичиганского государственного университета “State Networks”. В ней хранятся данные о различных демографических, социальных, экономических и других показателях между штатами США [1]. Основным показателем для нашего исследования возмём “ACS_Migration” - миграцию из “State 2” в “State 1” за 2017 год. Также поставим наш исследовательский вопрос следующим образом - Выявление наиболее привлекательного штата США для миграции. В процессе анализа мы также рассмотрим дополнительные факторы, которые могут влиять на этот процесс. В качестве пособия по R был использован учебник “Introduction to R and network analysis” [2].

2 Исследование на всех данных о миграции

1. Для начала на основе наших данных осуществим визуализацию в виде графа, рассчитаем его характеристики, построим различные модели на основе параметров нашего графа и проведем сравнительный анализ.

   В качестве признака вершины графа возьмем название штата, а поток мигрантов будет представлен направленным и взвешенным ребром. Давайте построим граф:



   Исходя из внешнего вида графа можно сделать вывод о том, что он является перегруженным, из-за чего имеет слабую информативность.

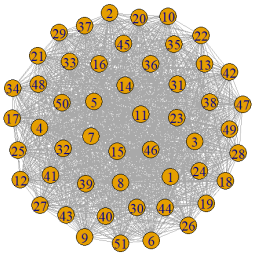
2. Рассчитаем основные характеристики графа:

   - число вершин графа: 51
   - число рёбер графа: 2550
   - степени всех вершин графа для каждого штата: 100
   - диаметр графа (не взвешенный): 1
   - взвешенный диаметр графа: 1141
   - радиус графа: 1
   - плотность графа: 1

   На основе полученных характеристик и внешнего вида графа можно сделать вывод о том, что он является полным, то есть каждая его вершина имеет связь между всеми остальными в графе.

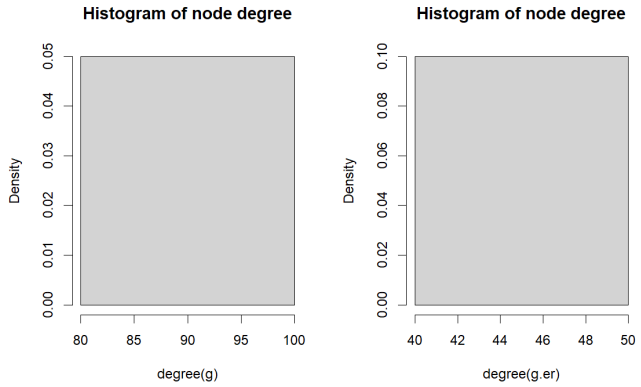
   Далее попробуем смоделировать графы разными методами и сравнить с реальными характеристиками.

   С применением модели Эрдёша-Реньи смоделируем граф, исходя из вероятности наличия ребра между двумя вершинами, которая равна плотности исходного графа, учитывая также количество его вершин:



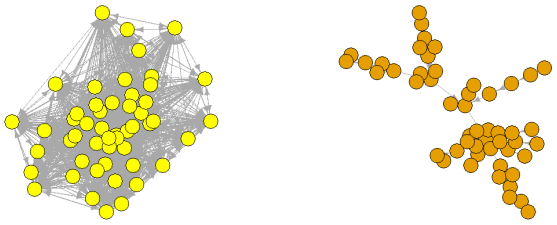
   - число вершин графа: 51
   - число рёбер графа: 1275
   - степени всех вершин графа: 50
   - диаметр графа: 1
   - радиус графа: 1
   - плотность графа: 1

   Исходя из полученной информации можно сделать вывод о том, что при моделировании степени

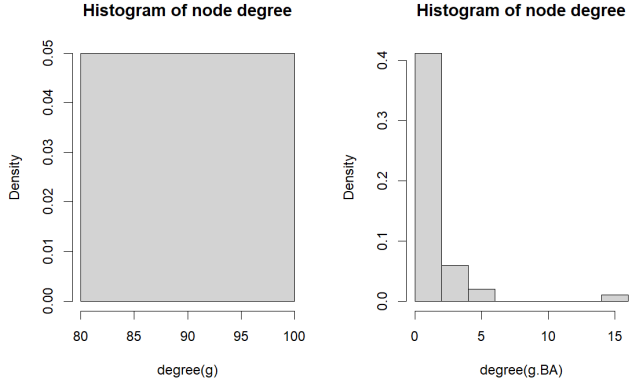


вершин графа уменьшились вдвое, что показывает высокую плотность исходного графа. Также изначально был дан полный граф, а следовательно вероятность выпадения ребра равна единице, что означает бессмысленность данного типа моделирования.

Воспользуемся моделью Барабаши — Альберта. Взяв в качестве параметра “n число вершин”, а за “power” единицу, для того чтобы обеспечить линеный рост.

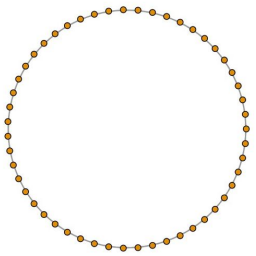


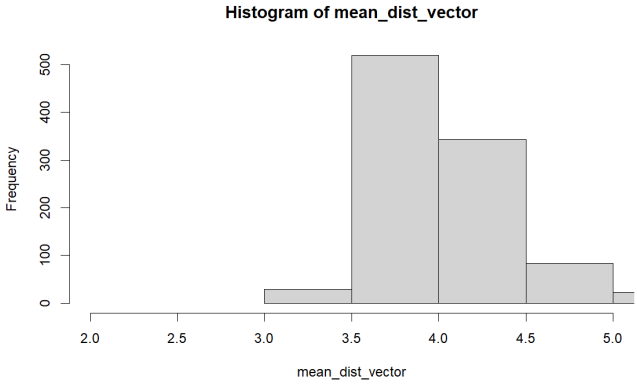
Как видно из рисунка выше смоделированный граф имеет меньше ребер из-за того, что не учитывает факт полноты изначального графа. Так-



же можно заметить меньшую плотность степеней вершин.

Перейдем к модели “Малый мир” Уоттса-Строгатца с регулярной решеткой при параметре вероятности равной нулю, а также размерностью равной единицы для каждого ребра размера 51 число вершин.





Модель Уоттса-Строгатца сама по себе изучает связь локальных групп (малый мир) с общей сетью. Так как мы передаем характеристики полного графа, можно сказать, что и в модели не будет локальных групп, поэтому нет особого смысла ее строить.

3. Далее в исследовании проведем расчет и анализ различных мер центральности.

   Начнем с Центральности по степени вершин. Эта центральность измеряет количество вершин, которые соединены с данной вершиной в графе, то есть это показатель, определяющий сколько связей у вершин. В нашем полном графе - подобная степень у каждой вершине одна и та же, и она равняется 100. То есть каждая вершина имеет связь со всеми остальными вершинами.

   Вторым возьмем центральность по близости. Мера центральности по близости измеряет, насколько близко (кратчайший путь) данная вершина к остальным вершинам в графе. Вершины с близкой центральностью находятся ближе к остальным вершинам и могут быть более центральными в смысле быстрого доступа к другим вершинам. Тут уже наибольшая центральность у штата Вермонт. Так как веса наших вершин - это миграция из одного штата в другой, то можно сделать

предположение, что это штат с наименьшей миграцией в обе стороны.

Третьим будет центральность по посредничеству. Центральность по посредничеству измеряет, насколько часто данная вершина находится на кратчайших путях между другими вершинами в графе. Вершины с высоким посредничеством могут выступать в роли посредников или связующих элементов между различными частями графа. Тут также самое большое значение принадлежит штату Вермонт. Но так как у нас полный граф - вершина с этим штатом пролегает через любой путь всех других вершин, плюс значения на ребрах у него наименьшие, значит и через нее и будет наименьший путь у всех ребер.

И последним возьмем по собственному значению - учитывает не только количество связей вершины, но и важность соседей этой вершины. Это означает, что вершина с высоким собственным значением центральности будет более центральной, если она связана с другими центральными вершинами. Здесь с наибольшей центральностью является штат Флорида - у него значение равняется 1, но на самом деле Rstudio увеличивает самое большое значение до 1, поэтому на самом деле оно может быть меньше, но все же наибольшим. Это значит, что по миграции - Флорида является наибольшим.

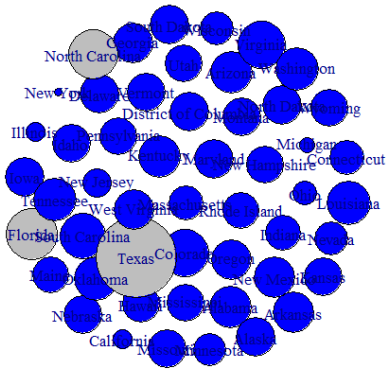
Но наилучшим в данном графе является центральность по собственному значению, так как с учётом того, что у нас полный граф и на ребрах находится поток миграции, то все остальные центральности показывает не то, что надо для нашего исследовательского вопроса.

4. Проверим, есть ли в графе предпочтительное присоединение.

   Поскольку наш граф является полным, то расчет ассортативности не является возможным, так как все вершины соединены друг с другом. Добавив дополнительную информацию о региональной структуре США, как Запад, Юг, Северовосток, Средний запад. Удалось посчитать диадичность равную 0.23, которая говорит нам о том, что 23% рёбер соединяют штаты из одного и того же региона и гетерофильность равную 0.77, означающую то что 77% всех рёбер в сети соединяют вершины разных категорий.

5. Кластеризация и выявление обществ в графе.

   При нашем полном графе кластеризация происходит неудачно, поэтому невозможно выделить общества в графе. Но можем выделить лидеров по чистому притоку мигрантов. Для этого зададим эту характеристику вершинам и отобразим их:



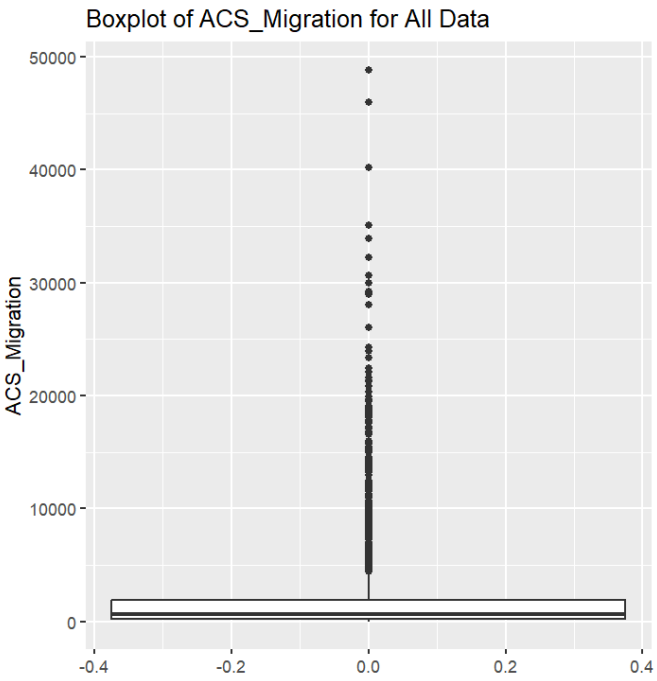
   Исходя из полученной информации делаем вывод о том, что лидером по чистому притоку являются Техас, Флорида и Северная Каролина

6. Общий вывод:

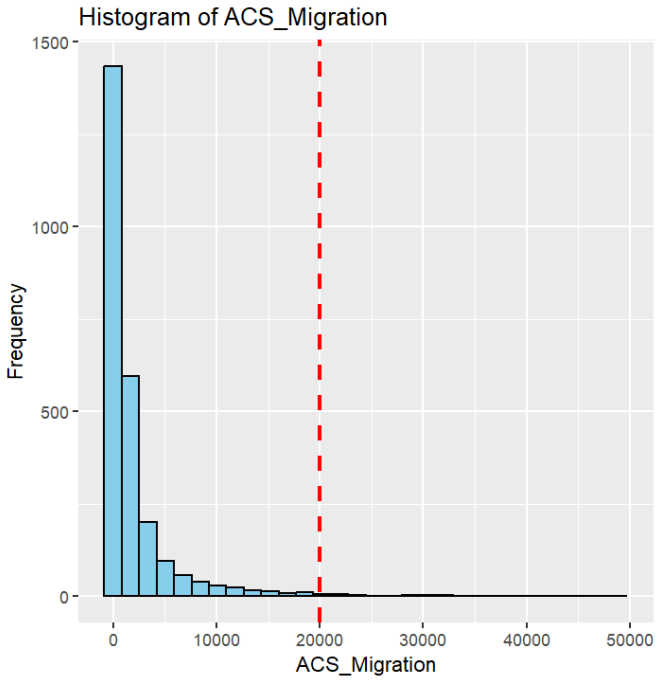
   Поскольку полный граф сложно анализировать, мы вернемся к его обсуждению в 4 пункте и сейчас будем рассматривать только крупные миграционные потоки и их поведение.

3 Исследование только на крупных миграциионных потоках

Изучим данные о миграции:

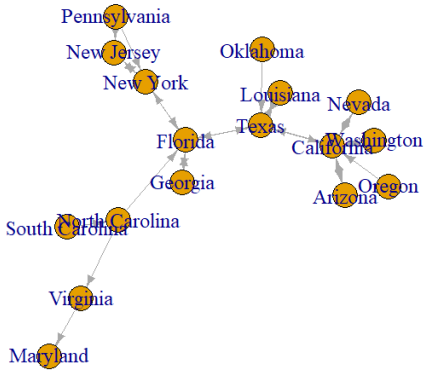


Исходя из ящика с усами видно, что основные потоки миграции сконцентрированы в районе тысячи человек. Для выделения крупного миграционного потока мы посчитали 99 квантиль и ограничили по нему. В дальнейшем исследовании мы будем считать, что если из штата въезжает или выезжает больше 19620 человек, то будем называть это крупным миграционным потоком.



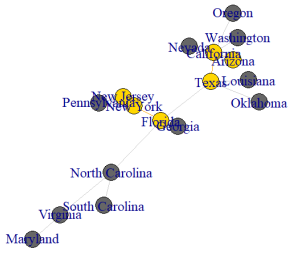
На гистограмме также видно, что основное распределение сконцентрировано приблизительно до 20000 человек, при его удалении мы сконцентрируемся на его хвосте и посмотрим крупнейшие миграционные потоки.

1. В качестве признака вершины графа возьмем название штата, а поток мигрантов будет представлен направленным ребром. Давайте построим граф:



2. Рассчитаем основные характеристики графа:

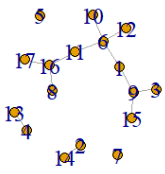
   - число вершин графа: 17
   - число рёбер графа: 26
   - степени всех вершин графа для каждого штата: Arizona - 2, California - 9, Florida - 7, Georgia - 2, Louisiana - 2, Nevada - 2, New Jersey - 3, New York - 5, North Carolina - 3, Oklahoma - 1, Oregon - 1, Pennsylvania - 2, South Carolina - 1, Texas - 7, Virginia - 2, Washington - 2, Maryland - 1
   - диаметр графа (не взвешенный): 5
   - взвешенный диаметр графа: 174698 Виузализируем его:

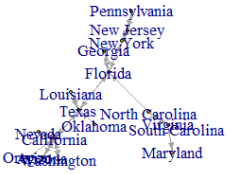


   - радиус графа: 3
   - плотность графа: 0.1

Попробуем смоделировать данный граф:

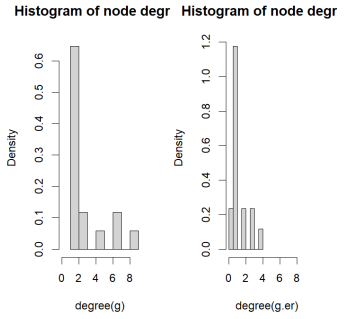
Воспользуемся моделью Эрдёша-Реньи с параметрами: n равному 17 и p равному 0.1.



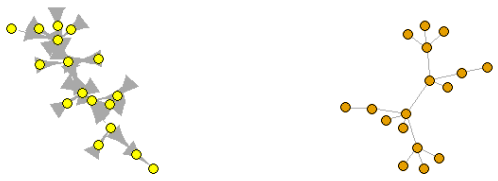


3.
- число вершин графа: 17
- число рёбер графа: 12
- степени всех вершин графа: 2, 1, 1, 1, 0, 4, 0, 1, 3, 1, 2, 1, 1, 1, 1, 3, 1
- диаметр графа: 6
- радиус графа: 0
- плотность графа: 0.09

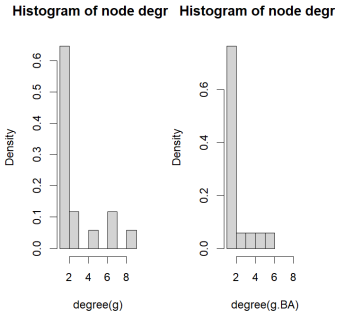
Исходя из полученных данных можно сделать о том, что при моделировании появляются изолированные вершины, в отличие от настоящего графа, также при моделировании меньше число ребер, нулевой радиус и больший диаметр. Данный тип моделирования не подходит по вышеописанным причинам, а также по степеням вершин, так как при моделировании число вершин со степенью 0 существуют тогда как в исходном графе таких вершин нет.



Воспользуемся моделью Барабаши — Альберта. Взяв в качестве параметра “n число вершин”, а за “power” единицу, для того чтобы обеспечить линеный рост.

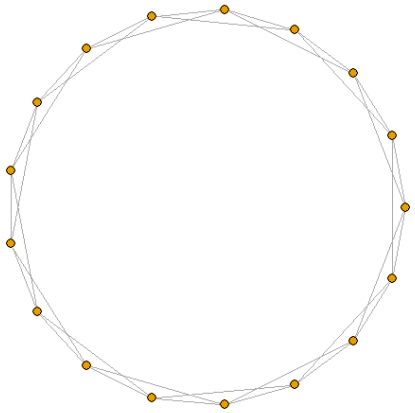


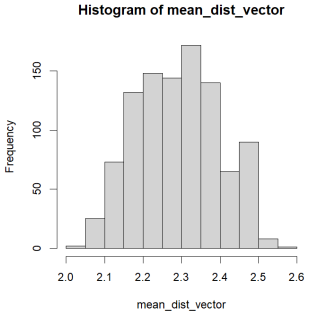
Как видно из рисунка выше граф неплохо моделируется и изначальная структура передана. Рассмотрим гистограмму степеней вершин:



По степеням вершин можно заметить также сходство, но в исходном графе присутствуют с большими степенями, чем в моделируемом

Перейдем к модели “Малый мир” Уоттса-Строгатца с регулярной решеткой при параметре вероятности равной нулю, а также размерностью равной единицы для каждого ребра размера 17 число вершин.





4. Далее в исследовании проведем расчет и анализ различных мер центральности.

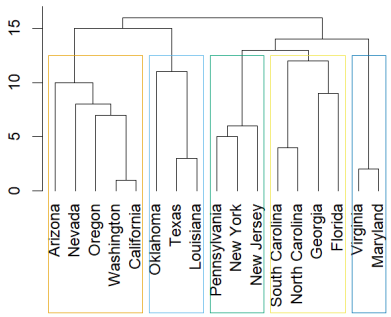
- Центральность по степени вершин: Самая большая степень у штата Калифорния равная 9, то есть у него больше всего связей со всеми другими - через Калифорнию происходит наибольшее количество миграционных потоков.
- Центральность по близости: Тут уже самая большая центральность у Верджинии, то есть у нее наименьшее расстояние до других штатов, то есть самая маленькая миграция.
- Центральность по посредничеству: Техас обладает наибольшим значением. То есть через нее чаще всего проходят пути миграции других штатов.
- Центральность по собственному значению: Тут также Калифорния на первом месте по значению. Значит показатель миграции наибольший тут. Но также и Техас недалеко по значению, значит он чуть менее централен.

5. Проверим, есть ли в графе предпочтительное присоединение.

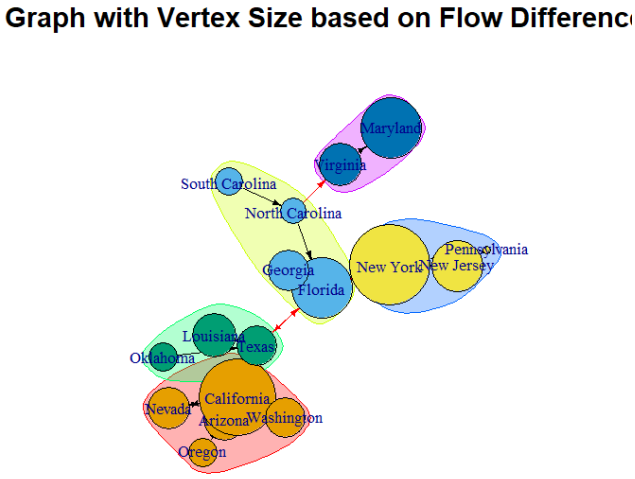
   Добавив в качестве дополнительных характеристик регион штата: Запад, Юг, Северо-восток, Средний запад. Рассчитаем ассоративность, которая равна 0.85, что означает наличие предпочтительного присоединения вершин из одного региона, то есть крупные миграционные потоки зачастую происходят внутри регионов США. Ассортативность же по степени равна -0.4, что указывает на дисассортативное соединение. Это означает, что в сети узлы с большим количеством связей имеют тенденцию соединяться с узлами с меньшим количеством связей и наоборот. То есть крупные "хабы"миграции соединяются с менее популярными пунктами назначения или источниками миграции.

   Удалось посчитать диадичность равную 0.85, которая говорит нам о том, что 85% рёбер соединяют штаты из одного и того же региона и гетерофильность равную 0.15, означающую то что 15% всех рёбер в сети соединяют вершины разных категорий.

6. Кластеризация и выявление обществ в графе. Проведем кластеризацию вершин с помощью метода реберного посредничества (edge betweenness). Этот метод определяет сообщества в сети, исходя из реберного посредничества, которое представляет собой меру, указывающую на количество кратчайших путей между всеми парами вершин, проходящих через данное ребро. В этом методе ребра с наивысшим значением посредничества удаляются последовательно и выделяются красным на графике, пока сеть не разделится на отдельные сообщества.



Как видно на рисунке выше удалось выделить 5 кластеров. Визуализируем их в виде графа и изменим размер вершины таким образом, чтобы чем больше был миграционный приток за вычетом оттока, тем больше была бы вершина.



Исходя из рисунка можно сделать выводы о том, что в каждом кластере явно виден один лидер по чистому притоку мигрантов. Такими центрами притяжения являются: Калифорния, Луизиана, Флорида, Мэриланд и Нью-Йорк.

7. Общий вывод:

   В ходе проведенного исследования сети крупных миграционных потоков между штатами США был проанализирован ряд ключевых характеристик графа. Было выяснено, что существует предпочтительное присоединение между штатами в пределах одного региона, что говорит о существенной роли региональных особенностей и взаимосвязей в процессе крупных потоков миграции.

   Ассортативность по степени указывает на тенден-

цию крупных "хабов"миграции к соединению с менее активными пунктами миграции. Это может быть связано с разнообразием экономических, социальных и культурных возможностей, которые предлагают крупные центры в сравнении с менее населенными районами.

Кластеризация показала наличие пяти ярко выраженных кластеров, в каждом из которых есть определенный штат-лидер, который привлекает наибольшее количество мигрантов. Эти штаты, такие как Калифорния, Луизиана, Флорида, Мэриланд, Нью-Йорк, играют ключевую роль в общенациональной миграционной динамике.

4 Уточнение исследовательского вопроса и построение предиктивной модели

Уточним исследовательский вопрос: Выявление привлекательного штата для миграции жителей США на примере всей информации о миграции за 2017 год и только на крупных миграционных потоках. В качестве предиктивной модели мы выбрали гравитационную модель, которая строится следующим образом:

$$\text{migr}_{i,j} = A\frac{(\text{ pop }_i \times pop_j)}{\text{dist }_{i,j}}$$

$\text{migr}_{i,j}$ - миграционный поток между регионами i и j;

A - константа (мера сходства между странами, контрольные переменные?);

pop_i, pop_j - население регионов i и j, соответственно;

$\text{dist}_{i,j}$ - мера расстояния между регионами.

Расширенная модель:

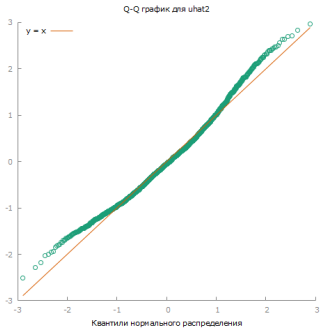
$$\begin{aligned}\ln\left(\text{migr}_{i,j}\right) = \beta_0 &+ \beta_1 \ln\left(\text{pop}_i\right) + \beta_2 \ln\left(\text{pop}_j\right) \\ &+ \beta_3 \ln\left(\text{dist}_{i,j}\right) \\ &+ \sum_k^K \beta_k^i X_i + \sum_k^K \beta_k^j X_j \\ &+ \varepsilon_{i,j}\end{aligned}$$

X_i и X_j - наборы регрессоров для регионов i и j соответственно

1. **Построим расширенную гравитационную модель для всех данных:**

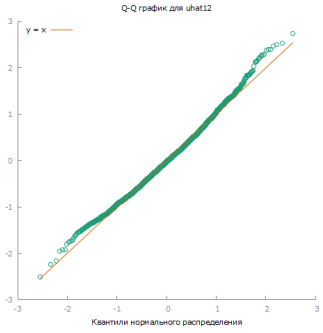
   Первая вариация модели включающая только информацию о населении в двух штатах и миграции:

$$\begin{aligned}\widehat{\text{l\_ACS\_Migration}} = &\underset{(0,5)}{-13,68} \\ &\underset{(0,02)}{+ 0,80}\,\text{l\_State1\_Pop} \\ &\underset{(0,02)}{+ 0,81}\,\text{l\_State2\_Pop} \\ &\underset{(0,02)}{- 0,56}\,\text{l\_Distance}\end{aligned}$$



$R^2 = 0,7$, а все регрессоры значимы. Также остатки регрессии идут практически под углом 45 градусов, что свидетельствует о их нормальности. Попробуем улучшить модель по показателю R^2:

$$\begin{aligned}\widehat{\text{l\_ACS\_Migration}} = &\underset{(0,42)}{-16,6} + \underset{(0,02)}{0,81}\,\text{l\_State1\_Pop} \\ &\underset{(0,01)}{+ 0,82}\,\text{l\_State2\_Pop} \\ &\underset{(0,02)}{- 0,25}\,\text{l\_Distance} \\ &\underset{(0,06)}{+ 1,63}\,\text{Border}\end{aligned}$$



$R^2 = 0,8$, а все регрессоры и константы значимы. Остатки регрессии распределены нормально.

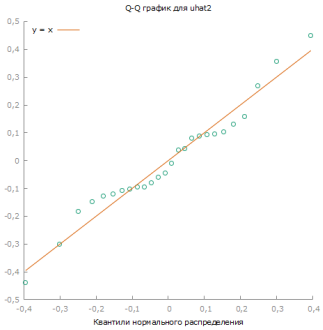
Проинтерпретируем результаты гравитационной модели:

- При увеличении на 1% населения первого штата, миграция увеличивается на 0,81%
- При увеличении на 1% населения второго штата, миграция увеличивается на 0,82%.

- При увеличении на 1% расстояния между двумя столицами штатов, миграция уменьшается на примерно 0,25%, что говорит о том, что увеличение расстояния между штатами снижает миграционную активность.
- Если между двумя штатами есть граница (Border = 1), то миграция увеличивается на 163%

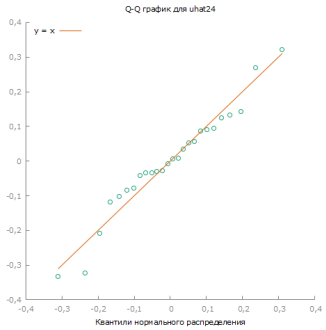
2. **Построим расширенную гравитационную модель для крупных миграционных потоков:**

$$\widehat{\text{l\_ACS\_Migration}} = \underset{(2,58)}{-0,36} + \underset{(0,08)}{0,35}\,\text{l\_State1\_Pop} + \underset{(0,1)}{0,36}\,\text{l\_State2\_Pop} - \underset{(0,09)}{0,2}\,\text{l\_Distance}$$



$R^2 = 0,45$ все регрессоры значимы, но константа не является значимой, а также Distance при 1% уровне значимости. Остатки распределены нормально. Попробуем улучшить результаты модели увеличив R^2 по средствам добавления регрессоров:

$$\widehat{\text{l\_ACS\_Migration}} = \underset{(2,63)}{1,20} + \underset{(0,08)}{0,29}\,\text{l\_State1\_Pop} + \underset{(0,1)}{0,27}\,\text{l\_State2\_Pop} - \underset{(0,08)}{0,04}\,\text{l\_Distance} - \underset{(7,26e-006)}{1,41e-006}\,\text{IncomingFlights} + \underset{(0,13)}{0,16}\,\text{Border} - \underset{(0,20)}{0,46}\,\text{IdeologyDif} - \underset{(0,002)}{0,005}\,\text{RaceDif} + \underset{(0,004)}{0,0029}\,\text{ReligDif}$$



$R^2 = 0,66$ значимыми являются только l_State1_Pop, l_State2_Pop, IdeologyDif и RaceDif. Остатки распределены нормально.

Проинтерпретируем значимы переменные в гравитационной модели с наибольшим R^2:

- При увеличении на 1% населения первого штата, миграция увеличивается на 0,29%
- При увеличении на 1% населения второго штата, миграция увеличивается на 0,27%.
- При увеличении на 1% идеологической разницы, миграция уменьшается на примерно 0,46%,
- При увеличении на 1% расовой разницы, миграция уменьшается на примерно 0,005%,

5 Выводы

Осуществив исследование на тему - Выявление наиболее привлекательного штата для миграции в США, мы выяснили, что по полному графу наиболее привлекательным штатом является Флорида по центральности по собственному значению . А по разнице входящих и исходящих миграционных потоков уже наиболее привлекательный стал штат Техас.

Ограничив далее данные по 99 квантилю и совершив исследование по такому графу, мы пришли к выводу, что наиболее лучший штат для миграции - является Калифорния по центральности по собственному значению. А по разнице входящих и исходящих - Нью-Йорк.

Далее в своем исследовании мы уточнили наш исследовательский вопрос. Мы использовали программу Gretl для построения гравитационной модели для наших данных - для полных и обрезанных. В полных данных мы получили, что популяция штатов, дистанция и границы между штатами влияют на изменение величины миграционных потоков. Увеличение популяции, наличие границ положительно влияет на изменение миграционных потоков, а увеличение дистанции влияет негативно.

В обрезанных данных у нас получилось, что регрессоры популяция населения, расовые и идеологические различия являются значимыми. Тут же популяция положительно влияет, а расовые и идеологические - отрицательно на изменения миграции.

Список литературы

[1] S. F. Olson, "State Networks Database v. 1.1," *East Lansing, MI: Institute for Public Policy and Social Research (IPPSR)*, 2019.

[2] K. Ognyanova, "Introduction to R and network analysis," *SCI Methods Workshop, Rutgers University*, 2018.