**Бихевиоральная модель возрастной аккумуляции в демографических данных и ее приложение к устранению следов аккумуляции в российских и международных данных**[[1]](#footnote-1)

**Д.М. Эдиев**

ФГБОУ ВО «Северо-Кавказская государственная академия»

dalkhat@hotmail.com; ediev@ncsa.ru

Ниже дается краткое описание математической модели, использованной для очищения от аккумуляции эмпирических демографических данных, результаты которого доступны в виде открытой электронной базы данных (Приложение 1. «Результаты апробации модели возрастной аккумуляции») по ссылке: [http://ncsa.ru/upload/doc/db\_uPx\_Ps\_5\_3linear.csv](http://ncsa.ru/upload/doc/regMxProj.csv).

**Исходные данные.**

Возрастная аккумуляция, т. е. сосредоточение численности населения в отдельных возрастах, возникает из-за склонности людей в ответах на вопрос о возрасте (при проведении переписи или обследования) назы­вать не точный, а приблизительный, округлённый возраст [1]. Чаще всего возрастная аккумуляция возникает в возрастах оканчивающихся на 0 и 5. Впрочем, интересно отметить, что анализ исторических данных по горским народам Кавказа, практиковавшим двадцатеричный счет, выявил необычную аккумуляцию в возрастах, кратных двадцати («круглых» в двадцатеричной системе счета) [2].

Хотя статистика населения в последние десятилетия (а в развитых странах – уже в прошлые столетия) сделала большой скачок в качестве и надежности данных, возрастная аккумуляция остается важным феноменом, ограничивающим точность данных о населении и по сей день. В ходе изучения аккумуляции возрастов и методов ее устранения, авторами сформирована обширная эмпирическая база данных о возрастных структурах различных популяций, источниками для которой послужили: база данных переписей ООН [3], международная база данных по смертности Human Mortality Database [4], данные первой Всеобщая перепись населения Российской империи 1897 года, переписей населения СССР за различные годы и Всероссийских переписей населения [5], а так же архивные данные о населения некоторых народов Кавказа (автор выражает благодарность Хатуеву Р.Т. и Псху А.В., любезно оказавшим содействие). Сформированная база данных содержит 33 332 индивидуальные записи, охватывающие годы с 1756 по 2019 и 249 стран/территорий. Иллюстрацией распространенности проблемы возрастной аккумуляции является динамика индекса Уипла аккумуляции возрастов (см. детали расчетов ниже), которая показана на рисунке 1 (по данным переписей с 1950 года). Индекс Уипла выше 100 указывает на наличие аккумуляции. Как видно из результатов расчетов, представленных на рисунке 1, несмотря на некоторое улучшение качества данных переписей, даже в последнее десятилетие, переписные данные по многим странам содержат следы возрастной аккумуляции. Более того, после 1990-х заметно ухудшение качества данных переписей (это, в частности, произошло в данных переписей постсоветских стран).

Индексы Уипла измеряют возрастную аккумуляцию в возрастах, оканчивающихся на "0" и "5" по степени отклонения усредненной численности населения в таких возрастах от средней численности в возрастах 23-62 года:

, (1)

, (2)
где – численность населения в возрастной группе лет. Индекс Уипла больше 100 свидетельствует о наличии возрастной аккумуляции. Индекс измеряет степень аккумуляции в возрастах, кратных пяти (25, 30, 35 и т. д. вплоть до 60 лет), а индекс измеряет степень аккумляции в возрастах, кратных десяти.

Альтернативным традиционным показателем для измерения возрастной аккумуляции, является «смешанный метод» Миерса. Для каждой возможной последней цифры возраста, рассчитывается отклонение значений долей населения в каждом из возрастов, оканчивающихся на «0», «1», ..., «9», от 10% (после особой поправки на естественное убывание численности населения с возрастом [1]) и, таким образом, получают индексы предпочтения возраста, оканчивающегося на соответствующую цифру. Таким образом, в отличие от метода Уипла, метод Миерса дает возможность оценивания предпочтения для каждой конечной цифры возраста. Сводный индекс аккумуляции Миерса выводится как половина суммы абсолютных отклонений для всех конечных цифр возраста от 10%.

На рисунках 2 и 3 представлены гистограммы распределения индексов Уипла и на эмпирических данных, собранных авторами*.* На рисунках 4 и 5 представлены гистограммы для индексов Миерса и. Как видно из представленных распределений, используемая база данных о возрастной структуре реальных населений вполне репрезентативна для населений с различными уровнями аккумуляции, измеряемой индексами как Уипла, так и Миерса. Распределение индекса , скошенное вправо, указывает на случаи аккумуляции в возрастах, оканчивающихся на цифру 0 (аналогично – ), а скошенные влево распределения для других цифр окончания возраста (кроме ) указывают на «избегание» таких, не «круглых», возрастов.


Рис. 1. Распределения («свечные» диаграммы) коэффициента Уипла в данных переписей населения, проведенных с 1950 года в зависимости от периода проведения переписи.


Рис. 2. Распределение записей сформированной базы данных эмпирических
возрастных структур по значению индекса Уипла


Рис. 3. Распределения записей сформированной базы данных эмпирических
возрастных структур по значению индекса Уипла


Рис. 4. Распределение записей сформированной базы данных эмпирических
возрастных структур по значению индекса Миерса


Рис. 5. Распределение записей сформированной базы данных эмпирических
возрастных структур по значению индексов Миерса . (в %)

**Модель возрастной аккумуляции и метод реконструкции истинной возрастной структуры населения**

В приложенных результатах апробирована модифицированная и обобщённая модель Эдиева [6]:

…

…, (3)
где – наблюдаемая возрастная структура населения, подверженная аккумуляции возрастов ( – переменная возраста); - истинная возрастная структура, которую необходимо реконструировать; – показатель силы аккумуляции в круглом возрасте ; – коэффициенты, характеризующие поведение людей, механизм округления возрастов, так что дает долю людей возраста типа (оканчивающемуся на заданную цифру в исходной модели, далее модель будет обобщена) , которые округляют свой возраст к ближайшему круглому возрасту типа *.*

В обощенной модели, используемой здесь, предполагается более общая струкутра округления возрастов (которая позволяет анализировать в т.ч. и ситуации округления в рамках двадцатеричной системы исчисления):

, (4)
здесь дает долю людей возраста , округляющих его до круглого возраста *.* Струкутру округления возрастов можно описать, вообще говоря, с помощью функции типа круглого возраста и рассчтояния между задейтсвованными возрастами:

, (5)
где – тип круглого возраста (в исходной модели, последняя цифра возраста). Ввиду мультипликативного вхождения параметров и в (4), требуется дополнительное условие масштабирования, которое в исходной модели формулировалось для параметров (). Здесь используется более удобное для общей постановки условие:

 (6)
для всякого круглого возраста .

При заданных параметрах модели, (4) является линейной системой и может быть решено в векторно-матричной форме:

, (7)
где, , **,**  – единичная матрица.

Для нахождения параметров модели (4)-(7), используется оптимизационная процедура, максимизирующая гладкость – в рамках модели – реконструируемой возрастной структуры :

, (8)
где,  **–** некий стандартный профиль, свободный от аккумуляции (нами используется сглаживающий сплайн, приложенный к исходному населения с числом степеней свободы 10% от количества возрастных групп).

В типичном случае, оптимизации в (8) подлежит 20 параметров () и 16 параметров (; ), всего 36 параметров. Это число возрастает при анализе более сложных случаев округления возрастов. Большая размерность задачи оптимизации и наличие множества локальных экстремумов могут быть препятствием к применению (8) на практике. В работе предлагается разбить процедуру оптимизации на этапы и упростить задачу, с тем, чтобы повысить устойчивость алгоритма. Предположим известный набор параметров (варианта выбора этих параметров описаны далее). Тогда модель (4), (7) линейна в относительно параметров . Кроме того, заменим неизвестную возрастную структуру , которая влияет на величину округления возрастов в правой части в (4), (7) на аппроксимацию :

. (9)
Аппроксимация может быть уточнена итеративно на основе текущих оценок (сглаженных для повышения устойчивости алгоритма). В прилагаемых результатах использовано 20 таких итераций, но близкие результаты дает совпадающее со сплайном . Уравнение (9) линейно относительно параметров :

, (10)
где – вектор параметров для предположительно круглых возрастов , – матрица произведений :

, (11)
где, . Подставляя (10) в (8), имеем:

. (12)
Дифференцируя (12) по , имеем:

. (13)
Это ссотношение можно решить, используя обощенную усеченную обратную матрицу:

. (14)
 По оцененным параметрам в (14), можно реконструировать неискаженную возрастную структуру населения (10):

, (15)
что завершает метод реконструкции.

**Опции для параметров**

Хотя оптимизационный подход для нахождения параметров может быть полезен и информативен в отдельных случаях, он не столь эффективен как универсальный алгоритм ввиду малой чувствительности (и наличия локальных минимумов) целевой функции к этим параметрам. В общем случае, хорошие результаты дает простое линейное приближение в (5):

 (16)
где, – предполагаемый лан для округления к круглым возрастам типа (в приложении, использованы лаги 5 и 3 для возрастов, оканчивающихся на 0 и 5, соответственно).

**Приложение 1.** Результаты апробации модели возрастной аккумуляции

См. отдельный файл, доступный по ссылке: [http://ncsa.ru/upload/doc/db\_uPx\_Ps\_5\_3linear.csv](http://ncsa.ru/upload/doc/regMxProj.csv)

**Поля, использованные в базе данных:**

| **Поле** | **Значение** |
| --- | --- |
| Country | Страна |
| Region | Регион |
| Type | Тип поселений |
| Comment | Комментарий |
| Year | Год |
| Sex | Пол |
| Indicator | Индикатор |
| Source | Источник данных |
| Link | Ссылка на источник |
| Whipple5 | Инндекс Уипла К5 |
| Whipple10 | Инндекс Уипла К10 |
| X0…Х100 | Реконструированные (выравненные) значения численности населения в возрастах 0…100 |
| K5.EDI | Индекс Уипла после применения выравнивания |
| Resid.EDI | Показатель отклонения выравненной возрастной струкутры от исходной |
| smoothPs | параметр алгоритма |
| svdPs | параметр алгоритма |
| iterate.target | параметр алгоритма |
| p.(y-x).x | Параметры модели ***p*** по типу круглого возраста и уклонению от него |
| k.x | Параметры модели ***k*** силы аккумуляции в возрасте ***x*** |

**Литература**

1. Shryock H.S., Siegel J.S. The Methods and Materials of Demography. Washington D.C.: United States Bureau of the Census, 1973.

2. Ediev D.M. Eight decades or just four score? An unusual case of age heaping // Le Hofrat. Vienna, 2019. P. 4–5.

3. United Nations Statistics Division. Demographic Statistics Database. Population by age, sex and urban/rural residence [Electronic resource]. 2020. URL: http://data.un.org/Data.aspx?d=POP&f=tableCode%3A22 (accessed: 12.11.2020).

4. University of California (Berkeley), The Max Planck Institute for Demographic Research. Human Mortality Database. Online database sponsored by University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany) [Electronic resource]. 2020. URL: www.mortality.org (accessed: 05.08.2020).

5. Демоскоп Weekly. Приложение Демоскопа Weekly [Electronic resource]. 2020. URL: http://www.demoscope.ru/weekly/pril.php (accessed: 12.11.2020).

6. Ediev D.M. Demographic Losses of Deported Soviet Peoples [Demograficheskiye Poteri Deportirovannykh Narodov SSSR]. Stavropol: AGRUS, 2003. 336 p.

1. Работа выполнена при финансовой поддержке Российского Фонда Фундаментальных Исследований (проект №18-01-00289 «Математические модели и методы устранения искажений показателей смертности и продолжительности жизни престарелого населения»). [↑](#footnote-ref-1)