

Восстановление повозрастных коэффициентов рождаемости детей различных последовательностей

© 2017 Панфиль Любовь Александровна
Российский экономический университет имени Г.В. Плеханова
117997, г. Москва, Стремянный пер., д. 36
E-mail: Lybov5@mail.ru

Рассмотрены подходы к восстановлению отсутствующей статистической информации, необходимой для анализа количественных характеристик исследуемого объекта. Выявлены их основные преимущества и недостатки, разработан модифицированный метод, учитывающий особенности восстановления данных повозрастных коэффициентов рождаемости по последовательности рождения детей.

Ключевые слова: демография, демографическая политика, коэффициенты рождаемости, стимулирующие меры, эффективность.

Депопуляция населения России, начавшаяся с 1990 г. и выраженная в увеличении коэффициентов смертности и снижении коэффициентов рождаемости, привела страну к демографическому кризису¹. Отсутствие обеспечения естественного воспроизводства населения влечет за собой проблемы экономического и политического характера. По прогнозам Организации Объединенных Наций, на территории РФ в 2017 г. проживает 1,9 % населения мирового сообщества, к 2050 и 2100 гг. численность населения страны сократится и составит 1,4 и 1,1 % мирового населения². Это ставит под угрозу обеспечение целостности государства. Активные действия по решению данной проблемы предпринимались правительством страны с 2006 г.: были увеличены ежемесячные пособия по уходу за ребенком, введены родовые сертификаты, а также материнский капитал.

Анализ эффективности вводимых стимулирующих мер демографической политики в области рождаемости невозможен без изучения повозрастных коэффициентов рождаемости по последовательности рождения детей (ПКРП).

Под повозрастным коэффициентом рождаемости по последовательности рождения детей понимается показатель, отражающий количество рожденных детей определенной последовательности (первые, вторые, третьи и последующие) на 1000 женщин фертильного периода (15-49 лет) или отдельной их возрастной группы в регионе г в году t.

Сбор полноценной и достоверной информации за длительный временной период может быть затруднен частичным или полным ее отсутствием в некоторых регионах. В соответствии с Федеральным законом об актах гражданского состояния от 15 ноября 1997 г. № 143-ФЗ фикса-

ция порядкового номера рождения ребенка в актовой записи его рождения не предусматривается, но в 38 регионах сохранился учет порядкового номера рождения ребенка. Данная информация дает возможность оценивать эффективность пронаталистской политики РФ в этих регионах и по стране в целом. Желательно на местном и региональном уровнях иметь системы управленческого учета и планирования, которые себя хорошо зарекомендовали и успешно используются в организациях микроэкономики³. Среди этих 38 регионов данные ПКРП за один год в течение периода 1993-2016 гг. отсутствуют в 9 регионах, данные за два соседних года отсутствуют в 2 регионах.

Одним из способов восстановления информации является вычисление среднего значения из соседних коэффициентов в соответствии с выражением⁴ (1):

$$k_{i_p}^r = \frac{k_{i-1_p}^r + k_{i+1_p}^r}{2}, i = \overline{1,7}, r = \overline{1,38},$$

$$t = \overline{2,23}, p = \overline{1,2,3}, \quad (1)$$

где $k_{i_p}^r$, $k_{i-1_p}^r$, $k_{i+1_p}^r$ - количество рожденных детей последовательности p (1 - первые дети, 2 - вторые, 3 - третьи и последующие) в регионе r на 1000 женщин i-й возрастной группы (i = 1 - возрастная группа женщин 15-19 лет, i = 2 - 20-24 года, i = 7 - 45-49 лет) в году t, t - 1 и t + 1 (t = 2 - 1994 г., t = 3 - 1995 г., a t = 23 - 2015 г.), соответственно.

Другой способ восстановления данных - взятие среднего значения из нескольких соседних ПКРП. Возможно усреднение значений с неравными весовыми коэффициентами, определяемыми экспертным путем, по выражению (2):

$$k_{i_{tp}}^r = \sum_{j=1}^J v_j k_{i_{t-jp}}^r + \sum_{j=1}^J v_l k_{i_{t+1p}}^r, i = \overline{1,7},$$

$$r = \overline{1,38}, t = \overline{(1+J), (24-l)}, \quad (2)$$

где $k_{i_{tp}}^r, k_{i_{t-jp}}^r, k_{i_{t+1p}}^r$ - количество рожденных детей последовательности p в регионе r на 1000 женщин i -й возрастной группы в году $t, t-j$ и $t+1$, соответственно, v_j и v_l и - весовые коэффициенты, суммарное значение которых должно составлять единицу.

Значимыми преимуществами приведенных способов восстановления ПКРП являются их простота и высокая скорость реализации. Недостаток - искажение реальной информации, высокая погрешность⁵.

Еще одной альтернативой восстановления данных о ПКРП является метод k ближайших соседей. При помощи этого метода возможно восстановление $k_{i_{tp}}^r$ в регионе r в году t в определенной i -й возрастной группе женщин на основе аналогичных данных в других регионах, значения ПКРП в которых доступны за весь период.

Метод k ближайших соседей многошаговый и подразумевает реализацию следующих этапов:

1. Отбор регионов, в которых информация о ПКРП в определенной i -й возрастной группе женщин не имеет пропусков за весь изучаемый период. Множество $R^1 = \{r^1\}$ причем $R^1 < 38$.

2. Вычисление расстояний от регионов группы R^1 до региона r , в котором имеется пропуск данных в году t у i -й возрастной группы женщин в соответствии с выражением (3):

$$D_{r^1, r} = \sqrt{\sum_{j=1}^{t-1} (k_{i_{t-jp}}^{r^1} - k_{i_{t-jp}}^r)^2 + \sum_{j=t+1}^T (k_{i_{t-jp}}^{r^1} - k_{i_{t-jp}}^r)^2}. \quad (3)$$

Количество рассчитываемых расстояний определяется набором элементов множества R^1 .

3. Формирование группы регионов $R^2 = \{r''\}$, причем $R^2 \leq R^1$, состоящей из k элементов. В ее состав входят такие регионы, расстояние от которых до восстанавливаемого региона минимально. Количество их k устанавливается экспертным путем.

4. Вычисление весовых коэффициентов для каждого из элементов множества R^2 производится по выражению (4):

$$V_{r''} = \frac{1}{1 + D_{r^1, r}}. \quad (4)$$

5. На завершающем этапе восстановления данных по методу k ближайших соседей вычисляется недостающее значение ПКРП $k_{i_{tp}}^r$ в регионе r в году t у определенной i -й возрастной группы женщин по формуле (5)

$$k_{i_{tp}}^r = \frac{\sum_{r'' \in R^2} r''^2 \left[k_{i_{tp}}^{r''} \cdot V_{r''} \right]}{\sum_{r'' \in R^2} r''^2 V_{r''}}. \quad (5)$$

Относительно предыдущих методов восстановления пропущенных данных метод k ближайших соседей чаще остальных демонстрирует высокую точность при подборе недостающего значения, но его многошаговость значительно замедляет скорость импутации данных.

В научной литературе также приведены другие способы восстановления отсутствующих данных, как, например, на основе регрессионного моделирования⁶, при помощи использования инструментария Data mining⁷, на основе методов нечеткого моделирования⁸ и пр.

В настоящее время актуальным становится восстановление статистических данных на основе проектирования и обучения нейронных сетей. Преимуществом этого метода является получение высокоточной информации, однако реализация процедуры затруднена необходимостью наличия вспомогательного программного обеспечения и умением обращаться с ним.

Все вышеперечисленные методы служат адаптацией общих методов восстановления исходной информации к восстановлению данных ПКРП, однако ни один из них не учитывает особенностей показателя. В связи с этим автором разработан модифицированный метод, который учитывает специфику показателя, позволяет производить расчеты на доступном программном обеспечении и демонстрирует высокую точность в подборе пропущенных данных.

Спецификой данных ПКРП является наличие взаимосвязи этого показателя с рождаемостью всех детей без учета последовательности их рождения. Статистическая информация о численности рожденных детей без учета последовательности их рождения на 1000 женщин имеется во всех регионах, по всем возрастным группам женщин фертильного периода с 1993 по 2016 г. Необходимо учитывать, что коэффициент рождаемости определенной последовательности рождения детей не может превышать коэффициента рождаемости без учета последовательности, а сумма коэффициентов рождаемости по всем последовательностям должна быть равна коэффициенту рождаемости без учета последовательности.

Эти особенности не учтены в большинстве вышперечисленных методов. Для того чтобы учесть их, восстановление отсутствующего коэффициента рождаемости по последовательности рождения детей в регионе г в году t у определенной i-й возрастной группы женщин целесообразно производить с учетом соседних значений ПКРП, а также коэффициентов рождаемости всех детей без учета последовательности их рождения по формуле (6):

$$k_{i_{tp}}^r = \begin{cases} 0, & k_{i_{t-1p}}^r + k_{i_{t+2p}}^r = 0 \\ 0,5 \cdot \left(\frac{k_{i_{t-1p}}^r}{k_{i_{t-10}}^r} + \frac{k_{i_{t+2p}}^r}{k_{i_{t+20}}^r} \right) \cdot k_{i_{t0}}^r, & k_{i_{t-1p}}^r + k_{i_{t+2p}}^r \neq 0, \end{cases} \quad (6)$$

$$i = \overline{1,7}, r = \overline{1,38}, t = \overline{2,23}, p = 1,2,3,$$

где $k_{i_{tp}}^r, k_{i_{t-1p}}^r, k_{i_{t+2p}}^r$ - количество рожденных детей последовательности p (1 - первые дети, 2 - вторые, 3 - третьи и последующие) в регионе г на 1000 женщин i-й возрастной группы ($i = 1$ - возрастная группа женщин 15-19 лет, $i = 2$ - 20-24 года, $i = 7$ - 45-49 лет) в год, t - 1 и t + 1 ($t = 2$ - 1994 г., $t = 3$ - 1995 г., а t = 23 - 2015 г.), соответственно;

$k_{i_{t0}}^r, k_{i_{t-10}}^r, k_{i_{t+20}}^r$ - количество всех рожденных детей без учета последовательности их рождения в регионе г на 1000 женщин i-й возрастной группы в году t, t - 1 и t + 1, соответственно.

Модифицированные коэффициенты рождаемости ПКРП рассчитываются как доли от коэффициента рождаемости без учета последовательности рождения, что ограничивает значение $k_{i_{tp}}^r$ сверху, не позволяя ему превысить значение $k_{i_{t0}}^r$, а также делает сумму показателей $k_{i_{tp}}^r$ по всем последовательностям p равной $k_{i_{t0}}^r$.

Восстановление повозрастных коэффициентов рождаемости по последовательности рождения детей в регионе г в году t у i-й возрастной группы женщин в соответствии с модифицированным методом возможно только при наличии аналогичных значений ПКРП в регионе г в году t у i-й возрастной группы женщин в предшествующем и последующем годах. Если необходимо восстановить данные за 2 соседних года $k_{i_{tp}}^r$ и $k_{i_{t+1p}}^r$, то следует воспользоваться формулами (7) и (8):

$$k_{i_{tp}}^r = \begin{cases} 0, & k_{i_{t-1p}}^r + k_{i_{t+2p}}^r = 0 \\ 0,5 \cdot \left(\frac{k_{i_{t-1p}}^r}{k_{i_{t-10}}^r} + \frac{k_{i_{t+2p}}^r}{k_{i_{t+20}}^r} \right) \cdot k_{i_{t0}}^r, & k_{i_{t-1p}}^r + k_{i_{t+2p}}^r \neq 0 \end{cases} \quad (7)$$

$$i = \overline{1,7}, r = \overline{1,38}, t = \overline{2,22}, p = 1,2,3$$

$$k_{i_{t+1p}}^r = \begin{cases} 0, & k_{i_{tp}}^r + k_{i_{t+2p}}^r = 0 \\ 0,5 \cdot \left(\frac{k_{i_{tp}}^r}{k_{i_{t0}}^r} + \frac{k_{i_{t+2p}}^r}{k_{i_{t+20}}^r} \right) \cdot k_{i_{t+10}}^r, & k_{i_{tp}}^r + k_{i_{t+2p}}^r \neq 0 \end{cases} \quad (8)$$

$$i = \overline{1,7}, r = \overline{1,38}, t = \overline{2,22}, p = 1,2,3,$$

где $k_{i_{t+2p}}^r$ - количество рожденных детей последовательности p в регионе г на 1000 женщин i-й возрастной группы в году t + 2, $k_{i_{t+2p}}^r$ - количество всех рожденных детей без учета последовательности их рождения в регионе г на 1000 женщин i-й возрастной группы в году t + 2, остальные элементы выражений (7) и (8) идентичны элементам выражения (6).

Сравнительный анализ качества подбора данных по перечисленным методикам восстановления проведен на показателях повозрастных коэффициентов рождаемости Московской области в периоды 2011-2014 гг. для группы женщин 25-29 лет среди городского населения. Регион и возрастная группа были отобраны произвольно с учетом наличия полной информации о ПКРП в целях анализа отклонений моделируемых и реальных значений.

Восстановление ПКРП производилось пятью методами:

- 1) вычислением средних величин из двух соседних значений ПКРП (см. (1));
- 2) вычислением средних из четырех соседних значений с равными весовыми коэффициентами. Использовались 2 значения до моделируемого периода и 2 значения после моделируемого периода (в целях реализации этого метода наиболее поздним из восстанавливаемых периодов был определен 2014 г.);
- 3) методом k ближайших соседей. Значение $k = 5$ отобрано экспертным путем;
- 4) модифицированным методом (см. (6));
- 5) регрессионным моделированием. Построены парные регрессионные зависимости ПКРП

**Восстановление данных повозрастных коэффициентов рождаемости
по последовательности рождения детей в Московской области в период 2011-2014 гг.
для возрастной группы женщин 25-29 лет**

Тип данных	Метод восстановления	Последовательность рождения	Значения ПКРП				Отклонения от реальных данных, %				Среднее значение модулей отклонений, %
			2011	2012	2013	2014	2011	2012	2013	2014	
Реальные данные	-	первые	52,1	56,6	57,0	59,1	-	-	-	-	-
		вторые	34,2	36,6	36,6	38,7	-	-	-	-	-
		третьи*	5,9	6,6	7,1	7,9	-	-	-	-	-
		всего	92,2	99,8	100,7	105,7	-	-	-	-	-
Моделируемые данные	Среднее двух соседних ПКРП	первые	55,4	54,5	57,8	59,6	6,3	-3,6	1,6	0,8	3,1
		вторые	35,2	35,4	37,7	40,0	3,1	-3,4	2,9	3,2	3,2
		третьи*	5,9	6,5	7,2	8,0	0,3	-1,0	1,7	1,8	1,2
		всего	96,5	96,4	102,7	107,6	4,7	-3,3	2,1	1,8	3,0
	Среднее четырех соседних ПКРП	первые	54,8	55,6	57,5	59,4	5,1	-1,7	1,0	0,5	2,1
		вторые	35,1	35,8	38,2	40,5	2,7	-2,2	4,4	4,5	3,4
		третьи*	5,9	6,5	7,3	8,2	0,5	-0,5	3,1	4,4	2,1
		всего	95,8	98,0	103,1	108,1	3,9	-1,8	2,4	2,2	2,6
	Пяти ближайших соседей	первые	54,6	57,7	59,1	56,5	4,7	2,0	3,7	-4,4	3,7
		вторые	34,7	37,2	38,6	39,0	1,7	1,6	5,4	0,7	2,4
		третьи*	5,5	6,6	7,4	7,7	-6,3	0,6	4,0	-1,8	3,2
		всего	94,9	101,5	105,0	103,3	2,9	1,8	4,3	-2,3	2,8
	Модифицированный метод	первые	52,9	56,4	56,7	58,6	1,5	-0,2	-0,5	-0,8	0,8
		вторые	33,6	36,6	36,9	39,2	-1,6	0,0	0,8	1,3	0,9
		третьи*	5,6	6,7	7,1	7,8	-4,4	2,2	-0,5	-0,3	1,9
		всего	92,2	99,8	100,7	105,7	0,0	0,0	0,0	0,0	0,0
	Регрессионное моделирование	первые	56,8	59,6	62,5	65,3	8,9	5,4	9,7	10,5	8,6
		вторые	35,3	36,7	38,1	39,4	3,3	0,1	3,9	1,8	2,3
		третьи*	5,4	5,6	5,9	6,1	-8,2	-13,9	-17,0	-22,1	15,3
		всего	97,5	101,9	106,4	110,9	5,7	2,2	5,7	4,9	4,6

* В данной категории учитывается рождаемость третьих и последующих детей.

Источник. Центр демографических исследований РЭШ. URL: http://www.demogr.nes.ru/index.php/ru/demogr_indicat/data. Расчеты автора.

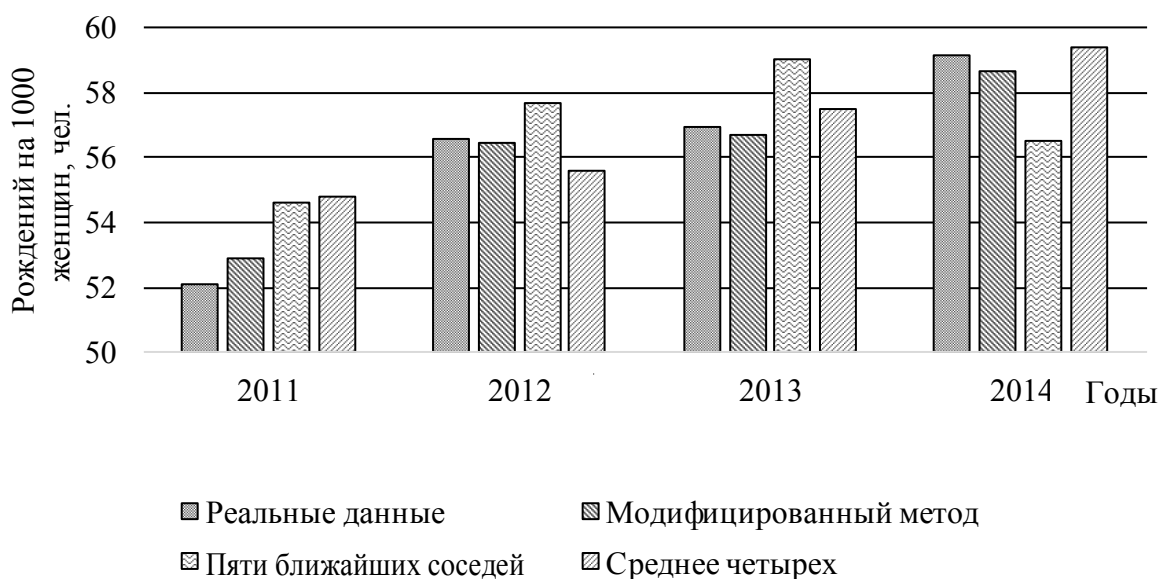


Рис. Сопоставительный анализ реального и моделируемых коэффициентов рождаемости первых детей среди женщин 25-29 лет Московской области

от времени (по данным 2006–2010 гг.), и вычислены прогнозные значения.

По результатам восстановления данных в Московской области установлено, что наиболее точный подбор повозрастных коэффициентов рождаемости по последовательностям рождения детей продемонстрировали высокую точность подбора информации, а также учет специфики этого показателя. Восстановление данных произведено при помощи программного обеспечения Microsoft Excel.

При моделировании рождаемости вторых детей значения, полученные при помощи модифицированного метода, также отличались наименьшими средними значениями модулей отклонений (0,9 %).

Наименьшая средняя погрешность восстановленных коэффициентов рождаемости третьих и последующих детей отмечена при использовании метода двух средних и составляет 1,2 %, при этом среднее значение модулей отклонений данных, полученных модифицированным способом, составляет 1,9 %.

В силу использования коэффициента рождаемости всех детей без учета последовательности их рождения в модифицированном методе сумма моделируемых значений ПРКП идентична реальному статистическому показателю рождаемости без учета последовательности рождения, тогда как использование других методов восстановления привело к отклонению моделируемых данных от действительных от -3,3 до 5,7 % в Московской области в период 2011–2014 гг. в группе женщин 25–29 лет.

Полученные результаты апробации модифицированного метода восстановления повозрастных коэффициентов рождаемости по последовательностям рождения детей продемонстрировали высокую точность подбора информации, а также учет специфики этого показателя. Восстановление данных произведено при помощи программного обеспечения Microsoft Excel.

¹ Панфиль Л.А. Подходы к оценке качества жизни населения // Путеводитель предпринимателя. 2016. □ 32. С. 183–199.

² Мировые демографические перспективы ООН. URL: <https://esa.un.org/unpd/wpp/Download/Standard/Population/>.

³ Алшина И.Ф. Управленческий учет для управленцев // Современные аспекты экономики. 2005. □ 3 (80). С. 78.

⁴ Тихомиров Н.П., Тихомирова Т.М., Урмаев О.С. Методы эконометрики и многомерного статистического анализа. Москва, 2010.

⁵ Карлов И.А., Кошур В.Д. Гибридный метод восстановления пропущенных данных с адаптивным управлением на основе нечеткой логики и нейронных сетей // Вестник КГПУ имени В.П. Астафьева. 2012. □ 4. С. 222–227.

⁶ Карлов И.А. Восстановление пропущенных данных при численном моделировании сложных динамических систем // Научно-технические ведомости СПбГПУ. Информатика. Телекоммуникации. Управление. 2013. □ 6 (186). С. 137–144.

⁷ Карлов И.А. Методы восстановления пропущенных значений с использованием инструментария Data mining // Вестник Сибирского государственного аэрокосмического университета имени академика М.Ф. Решетникова. Красноярск, 2011. □ 7 (40). С. 29–33.

⁸ Перемитин Т.О., Яценко И.Г., Лучкова С.В. Программный комплекс восстановления пропущенных значений в многомерных данных на основе методов нечеткого моделирования // Программные продукты и системы. 2014. □ 1 (105). С. 86–92.

Поступила в редакцию 02.08.2017 г.