

Международная лаборатория исследований населения и здоровья
Семинар «Современная демография»

Метод декомпозиции различий между двумя значениями кумулятивной демографической меры

Е.М. Андреев

22 марта 2017 г.

Представленный доклад опирается на следующие публикации

Андреев Е.М. 1982, Метод компонент в анализе продолжительности жизни. - Вестник статистики, № 9. С. 42-47

Shkolnikov V., Valkonen T., Begun A., Andreev E. 2001. Measuring inter-group inequalities in length of life. Genus, LVII(3-4), 33-62.

Andreev E. M., Shkolnikov V. M., Begun A.Z. 2002. Algorithm for decomposition of differences between aggregate demographic measures and its application to life expectancies, healthy life expectancies, parity-progression ratios and total fertility rates. Demographic Research 7, 499–522.

Shkolnikov V. M., Andreev E. M., Begun, A. Z. 2003. Gini coefficient as a life table function. Computation from discrete data, decomposition of differences and empirical examples. Demographic Research, 8, article 11:305–58. doi:10. 4054/DemRes. 2003. 8. 11.

Shkolnikov V. M., Andreev, E. M. 2010 Age-decomposition of a difference between two populations for any life-table quantity in Excel. MPIDR Technical Report TR-2010-002.

Shkolnikov V. M., Andreev E. M., Zhang Z., Oeppen J. E., Vaupel, J. W. 2011. Losses of expected lifetime in the United States and other developed countries: methods and empirical analyses. Demography, 48:1, 211-239. DOI:10. 1007/s13524-011-0015-6

Andreev E. M., Shkolnikov V. M. 2012. An Excel spreadsheet for the decomposition of a difference between two values of an aggregate demographic measure by stepwise replacement running from young to old ages MPIDR Technical Report TR-2012-002.

Простой вопрос

Допустим, мы имеем демографическую функцию равную $F(\alpha, \beta) = \alpha \cdot \beta$. В момент 0 она равна $a \cdot b$, в момент 1 она равна $A \cdot B$. Можно рассуждать так: если бы менялось только α , то функция стала бы больше на $(A - a) \cdot b$. Это есть вклад α в рост $F(\alpha, \beta)$. Если бы менялось только β , то функция стала бы больше на $a \cdot (B - b)$. Это есть вклад β в рост $F(\alpha, \beta)$. Если менялись обе переменные, после того как изменилось β вклад α в рост $F(\alpha, \beta)$ можно оценить как $(A - a) \cdot B$, а после того как изменилось α вклад β в рост $F(\alpha, \beta)$ к $A \cdot (B - b)$. На наш взгляд, обе оценки правильные: нет оснований отдать предпочтение началу или концу периода. Значит вклады α и β логичнее всего принять равными средней арифметической двух способов оценки, или

$$\text{вклад } \alpha = (A - a) \cdot \frac{B + b}{2}$$

(1)

$$\text{вклад } \beta = \frac{A + a}{2} \cdot (B - b).$$

Легко видеть, что

$$A \cdot B - a \cdot b = (A - a) \cdot \frac{B + b}{2} + \frac{A + a}{2} \cdot (B - b).$$

Метод декомпозиции Эвелин Китагава

То, что я описал, есть первый метод декомпозиции, предложенный Эвелин Китагава (Evelyn M. Kitagawa) в 1955 г. Она сравнивала значения двух общих коэффициентов смертности. Роль α играл вектор (ряд чисел) описывающий возрастную структуру населения ($\theta_0, \dots \theta_{85+}$), а β - вектор возрастных коэффициентов смирности ($M_0, \dots M_{85+}$). Сложение следует понимать как сложение векторов, а умножение – как их скалярное произведение, то есть сумму произведений координат.

$$\begin{aligned} CDR^2 - CDR^1 &= \sum_x (M_x^2 \cdot \theta_x^2 - M_x^1 \cdot \theta_x^1) = \\ (2) \quad &= \sum_x (M_x^2 - M_x^1) \cdot \frac{(\theta_x^1 + \theta_x^2)}{2} + \sum_x \frac{(M_x^2 + M_x^1)}{2} \cdot (\theta_x^2 - \theta_x^1) \\ &\quad \underbrace{\qquad\qquad\qquad}_{\text{Компонент смертности}} \qquad \underbrace{\qquad\qquad\qquad}_{\text{Компонент возрастной структуры}} \end{aligned}$$

За 100 лет между 1801 и 1901 общий коэффициент смертности населения Швеции снизился с 26.0 до 16.1 на 1000. Метод показывает, что вклад изменений смертности равен -11.9, а вклад изменений структуры равен 2.9 на 1000.

Общая постановка задачи

Пусть T некоторая кумулятивная характеристика демографического процесса, являющаяся функцией n переменных. Например, продолжительность жизни есть функция более чем ста возрастных коэффициентов смертности

$$T(\theta_1, \theta_2, \dots, \theta_n)$$

Рассмотрим два значения T , соответствующие двум точкам в n -мерном пространстве

$$T^1 = T(\theta_1^1, \theta_2^1, \dots, \theta_n^1)$$

$$T^2 = T(\theta_1^2, \theta_2^2, \dots, \theta_n^2)$$

Эти два значения могут быть связаны с изменением независимых переменных во времени и/или пространстве, и/или в зависимости от пола и/или иной группы населения. Допустим, мы хотим узнать вклад каждой из переменных в разность двух значений T так, чтобы

$$T^2 - T^1 = \sum \delta_i(\theta^1, \theta^2)$$

Каждое слагаемое $\delta_i(\theta^1, \theta^2)$ измеряет вклад различия между θ^1 и θ^2 в разность $T^2 - T^1$. Естественно ожидать, что если $\theta_i^1 = \theta_i^2$, то $\delta_i(\theta^1, \theta^2) = 0$ и что $\delta_i(\theta^1, \theta^2) = -\delta_i(\theta^2, \theta^1)$

Существование решения

Чтобы доказать, что задача имеет решение, мы просто явно построим это решение.

$$(3) \quad \begin{aligned} & [T(\theta_1^2, \theta_2^2, \dots, \theta_i^2, \dots, \theta_n^2) - T(\theta_1^1, \theta_2^2, \dots, \theta_i^2, \dots, \theta_n^2)] + \\ & + [T(\theta_1^1, \theta_2^2, \dots, \theta_i^2, \dots, \theta_n^2) - T(\theta_1^1, \theta_2^1, \dots, \theta_i^2, \dots, \theta_n^2)] + \dots + \\ & + [T(\theta_1^1, \theta_2^1, \dots, \theta_i^2, \dots, \theta_n^2) - T(\theta_1^1, \theta_2^1, \dots, \theta_i^1, \dots, \theta_n^2)] + \dots + \\ & + [T(\theta_1^1, \theta_2^1, \dots, \theta_i^1, \dots, \theta_{n-1}^1, \theta_n^2) - T(\theta_1^1, \theta_2^1, \dots, \theta_i^1, \dots, \theta_{n-1}^1, \theta_n^1)] \end{aligned}$$

Очевидно, что каждый член, кроме первого и последнего входит в данную сумму дважды, первый раз со знаком "-", второй раз со знаком "+". Таким образом, сумма в точности равна разности первого и последнего члена, то есть как раз $T^2 - T^1$.

Каждая квадратная скобка есть результат замены для очередного i θ_i^2 на θ_i^1 , где $i=1, \dots, n$. Таким образом мы можем отождествить каждую квадратную скобку с соответствующим $\delta_i(\theta^1, \theta^2)$.

Итак, разложение существует и далеко не единственное. Достаточно изменить порядок замен (перенумеровать переменные) и мы получим новое разложение.

Слишком много решений и другие проблемы

Возникает желание, как и в случае двух переменных перебрать все возможные варианты и взять в качестве решения среднюю арифметическую. Ясно, что число вариантов равно $n!$, так что полный перебор невозможен. За разумное время мне удалось сделать перебор для $n = 10$.

К идее последовательной замены почти одновременно пришли три

n	$n!$	n	$n!$
3	6	8	40 320
4	24	9	362 880
5	120	10	3628 800
6	720	11	39 916 800
7	5 040	12	479 001 600

демографа: Pressat (Франция) в 1985 г., Arriaga (США) в 1984 и Андреев в 1982 г. В 1980-ые общая идея перебора еще не была сформулирована. Авторы предлагали конкретные формулы для отдельных показателей. Но при выводе формул все трое двигались по возрасту от младших возрастов к старшим. На этом же принципе работают современные алгоритмы перебора вариантов порядка замены.

Другая, вначале неочевидная, проблема выявилась в процессе экспериментов. Оказалось, что если есть три набора $\theta^1, \theta^2, \theta^3$, то в общем случае

$$\delta_i(\theta^1, \theta^2) + \delta_i(\theta^2, \theta^3) \neq \delta_i(\theta^1, \theta^3)$$

В этой связи представляется целесообразным сравнивать ближайшие данные (скажем, соседние годы), а более отдаленные получать как суммы.

Метод пошаговой замены

Метод для определения вклада каждой переменной в изменение кумулятивной характеристики на основе формулы (3) получил название метода (или алгоритма) пошаговой замены (algorithm of stepwise replacement).

Во многих случаях практически невозможно перебрать все существующие упорядочения множества независимых переменных и при каждом упорядочении провести пошаговые замены.

С учетом этого сложилась следующая практика.

Для независимых переменных, зависящих от возраста, проводить замены только от младших возрастов к старшим. Если же несколько переменных относятся к одному и тому же возрасту, то желательно перебрать все возможные перестановки таких независимых переменных, если только их немного, скажем, меньше 10. Часто встречающийся пример: разложение изменений продолжительности жизни по возрастным группам и 7 группам причин смерти.

Во многих случаях удалось, применяя метод пошаговой замены к формальным переменным, вывести формулу, позволяющую осуществить декомпозицию без всяких замен. Самый удачный пример – формула для разложения по возрастам разности двух продолжительностей жизни.

Возрастные компоненты разности двух продолжительностей жизни.

Результаты пошаговых замен

Вклад возраста
 X в разность

$$\delta_x^{2-1} = l_x^2(e_x^2 - e_x^1) - l_{x+1}^2(e_{x+1}^2 - e_{x+1}^1) \leftarrow e_0^2 - e_0^1 \Big|_x$$

$$\delta_x^{1-2} = l_x^1(e_x^1 - e_x^2) - l_{x+1}^1(e_{x+1}^1 - e_{x+1}^2) \leftarrow e_0^1 - e_0^2 \Big|_x$$

$$\delta_x = \frac{\delta_x^{2-1} - \delta_x^{1-2}}{2}$$

Среднее значение

Здесь l_x^i есть значение функции дожития в возрасте x в населении i , а e_x^i - ожидаемая продолжительность жизни в том же возрасте в том же населении. Вклад возраста и причины смерти может быть определен по ниже приведенной приближенной формуле, в которой $M_x^i, M_x^{\{j\}i}$ есть возрастные коэффициенты смертности от всех и от j причины смерти

$$\delta_x^{\{j\}} = \frac{(M_x^{\{j\}2} - M_x^{\{j\}1})}{(M_x^2 - M_x^1)} \cdot \delta_x$$

Непрерывные модели (1)

В основе непрерывных методов декомпозиции лежит следующее свойство полного дифференциала функции в n -мерном пространстве. Пусть $T(\theta_1, \theta_2, \dots, \theta_n)$ -такая функция. Ее полный дифференциал по определению есть

$$dT(\theta_1, \theta_2, \dots, \theta_n) = \sum_{i=1}^n \frac{\partial}{\partial \theta_i} T(\theta_1, \theta_2, \dots, \theta_n) d\theta_i$$

Допустим, функция определена на некоторой линии соединяющей точки $\bar{\theta}^1 = (\theta_1^1, \theta_2^1, \dots, \theta_n^1)$ и $\bar{\theta}^2 = (\theta_1^2, \theta_2^2, \dots, \theta_n^2)$, и линия описана вектор-функцией $\bar{\theta}(t)$, $0 \leq t \leq 1$. Тогда интеграл дифференциала функции вдоль этой линии равен разности ее значений в конце и начале линии (точка в формуле означает производную по t)

$$T^2 - T^1 = \sum_{i=1}^n \int_0^1 \frac{\partial}{\partial \theta_i} T(\theta_1, \theta_2, \dots, \theta_n) \dot{\theta}_i dt$$

и i -ое слагаемое есть вклад i -ой переменной в изменение кумулятивной характеристики. Мне представляется, что первым этот подход описал Nathan Keyfitz в середине 1970-х годов.

Непрерывные модели (2)

По мнению Horiuchi, Wilmoth и Pletcher, тот факт, что в методе пошаговой замены не удается реализовать все возможные перестановки, делает его уязвимым. Поэтому они предложили свой собственный алгоритм расчета, основанный на непрерывной модели*. Непрерывные алгоритмы предлагались и ранее, но были как правило привязаны к одной кумулятивной переменной или одному типу разложения. Новый алгоритм почти столь же универсален, как метод пошаговых замен.

Проблема в том, что демографические данные дискретны. Если сравниваются два точки в жизни одного населения, то можно представить последовательность наблюдений с шагом 1 год, связывающим эти точки, но более мелкие деления есть условные построения. Если сравниваются два разных населения, то вся линия условна. То, что интеграл от полного дифференциала не зависит от выбора кривой вовсе не означает, что от выбора кривой не зависят *и* слагаемых.

В методе пошаговых замен камнем преткновения является вопрос, в каком порядке следует заменять переменные. В непрерывной модели – вопрос о соотношении скоростей изменения отдельных независимых переменных. Мне представляется, эти вопросы в чем то очень похожи.

Эксперименты показывают, что результаты использования двух методов практически не различаются.

* Horiuchi S., Wilmoth J. R.. Pletcher S. D. 2008. A Decomposition Method Based on a Model of Continuous Change. Demography Vol. 45, No. 4, 785-801.

Декомпозиция путем случайной пошаговой замены

Очень заманчивой представляется идея декомпозиции путем последовательности случайных перестановок независимых переменных. Первым эту идею высказал Д.А.Жданов. Он же сформулировал важное ограничение: чтобы использовать случайные перестановки надо быть уверенным, что с ростом числа пошаговых замен, средний вклад каждой переменной стремиться к пределу, равному тому среднему вкладу, который был бы получен после реализации всех последовательностей пошаговых замен. Можно согласится и в том случае, если это условие выполняется не всегда, но с некоторой высокой вероятностью. Нам пока не удалось найти решение этой задачи.

Практическое решение задачи декомпозиции

Метод пошаговой замены был сформулирован в 2002 г. В том же году была написана первая VBA программа для декомпозиции, а через 10 лет, после неоднократной доработки мы поместили ее на сайте MPIDR для свободного использования*.

http://www.demogr.mpg.de/en/projects_publications/publications_1904/mpidr_technical_reports/an_excel_spreadsheet_for_the_decomposition_of_a_difference_between_two_values_of_an_aggregate_4591.htm

Программа оформлена как книга Excel, куда пользователь помещает исходные данные – независимые переменные за периоды подлежащие сравнению, и где он должен средствами Excel написать программу для расчета зависимой переменной на основе независимых. Число возрастных групп не превосходит 120, а в каждом возрасте может быть использовано до 10 независимых переменных (причин смерти, социальных групп, территорий и т.д.)

Там же представлены 5 конкретных примеров использования метода декомпозиции.

*Andreev E. M., Shkolnikov V. M. 2012. An Excel spreadsheet for the decomposition of a difference between two values of an aggregate demographic measure by stepwise replacement running from young to old ages
MPIDR Technical Report TR-2012-002

Объект декомпозиции

Хотя в современной литературе декомпозиция используется весьма широко, список показателей, которые становятся объектом декомпозиции, довольно короток. Я сталкивался с декомпозицией показателей смертности

1. Общий и стандартизованный коэффициент смертности, от всех причин и от некоторых причин.
2. Ожидаемая продолжительность жизни в том или ином возрасте.
3. Джини коэффициент и абсолютное межиндивидуальное расстояние (AID) табличного распределения умерших по возрасту.
4. Годы потерянной жизни e^{\dagger}
5. Средний возраст смерти от некоторой причины.

В анализе рождаемости я сталкивался с декомпозицией вероятностей увеличения семьи (parity progression ratio) и коэффициента суммарной рождаемости.

Метод декомпозиции, с одной стороны, позволяет указать в какой мере изменения каждой из исходных переменных повлияли на итоговый показатель и, с другой стороны, открывают способ сравнения изменений в терминах кумулятивного показателя.

Я хочу показать некоторые примеры использования декомпозиции продолжительности жизни и стандартизованного коэффициента смертности.

Рост продолжительности жизни между концом 19 и началом 21 века

Возраст	Россия		Англия и Уэльс	
	Мужчины	Женщины	Мужчины	Женщины
Всего	39,37	46,70	33,60	33,91
В том числе за счет возрастов				
0 - 9	32,73	32,49	17,33	16,22
10 - 19	1,62	1,98	1,27	1,37
20 - 29	1,44	2,21	1,74	1,84
30 - 39	0,87	1,97	2,12	2,19
40 - 49	0,85	1,76	2,55	2,40
50 - 59	0,71	1,94	2,65	2,53
60 - 69	0,62	2,34	2,76	3,00
70 +	0,53	2,01	3,18	4,35

Взаимоотношение между составом населения и уровнем смертности

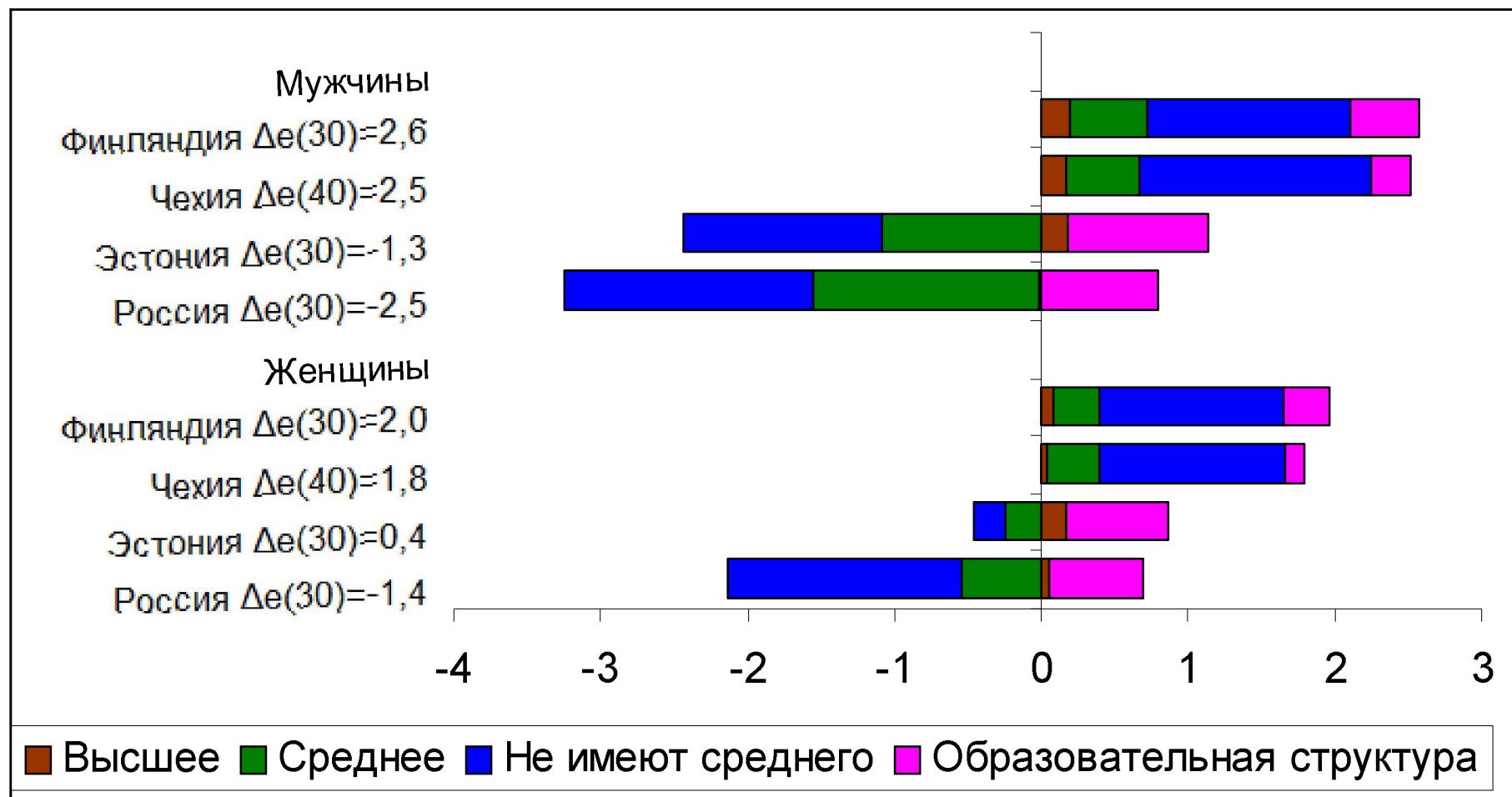
Я хотел бы продемонстрировать несколько достаточно старых результатов и один новый результат применения метода декомпозиции, связанных с оценкой влиянием структуры населения на смертность. Известно, что продолжительность жизни тем выше, чем выше уровень образования, а смертность женатых мужчин (это справедливо только если брак зарегистрирован) существенно ниже, чем не состоящих в браке. В наших работах мы попытались оценить влияние этих фактов на смертность населения, учитывая, что уровень образованности растет, а доля мужчин в зарегистрированном браке снижается во времени.

Shkolnikov V. M., Andreev E. M., Jasilionis D., Leinsalu M., Antonova O.I., McKee M. The changing relation between education and life expectancy in central and eastern Europe in the 1990s Journal of Epidemiology and Community Health (2006) 60:10, 875-881.

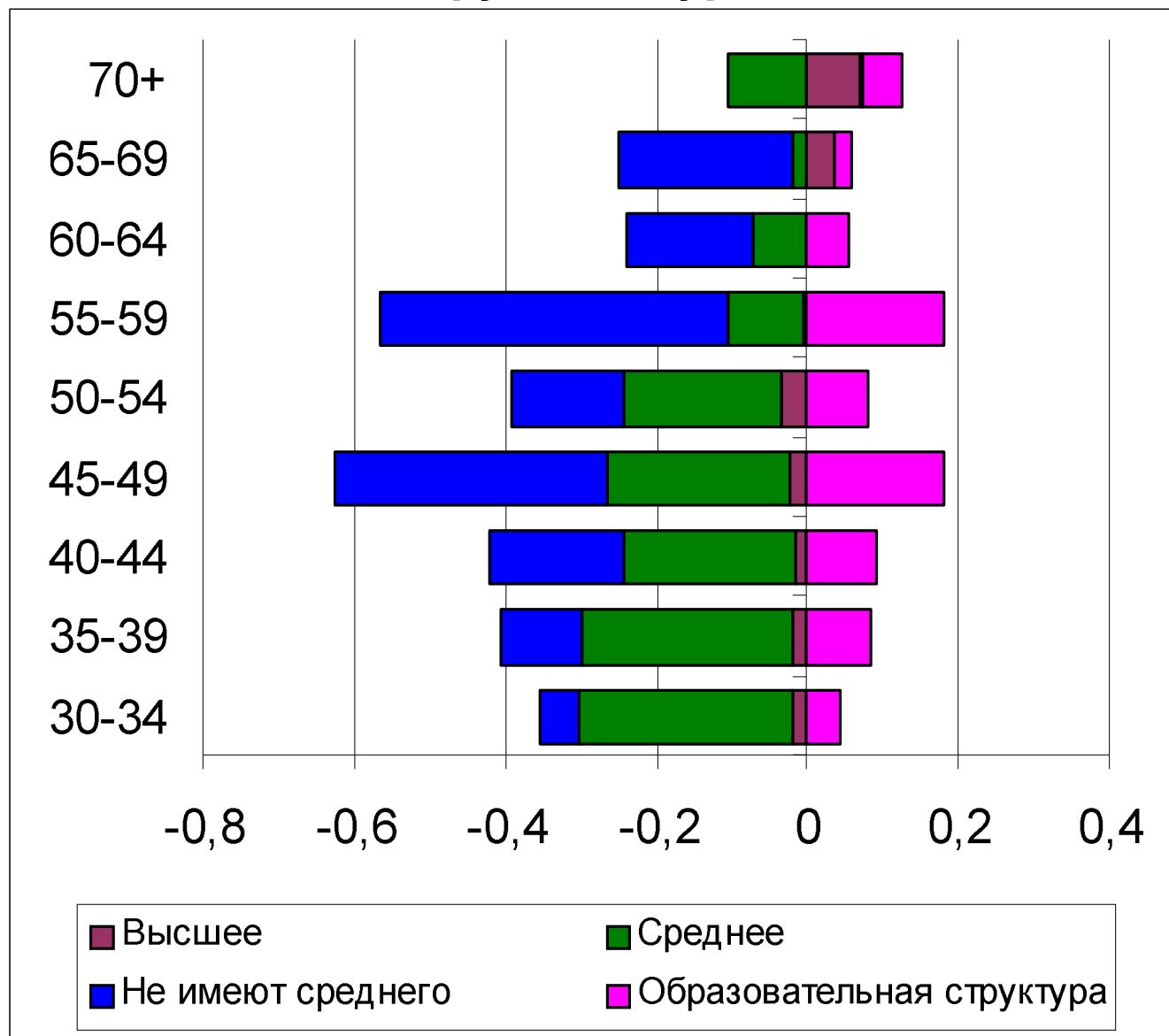
Jasilionis, D.; Andreev, E. M.; Kharkova, T. L.; Kingkade, W. W.: Change in marital status structure as an obstacle for health improvement: evidence from six developed countries European Journal of Public Health (2012) 22:4, 602-604.

Надо отметить, что мы использовали для анализа показатели смертности, основанные на несвязанных данных о живущих и об умерших, поэтому оценки смертности по образованию и брачному статусу могут содержать некоторые ошибки, что не меняет качественные выводы.

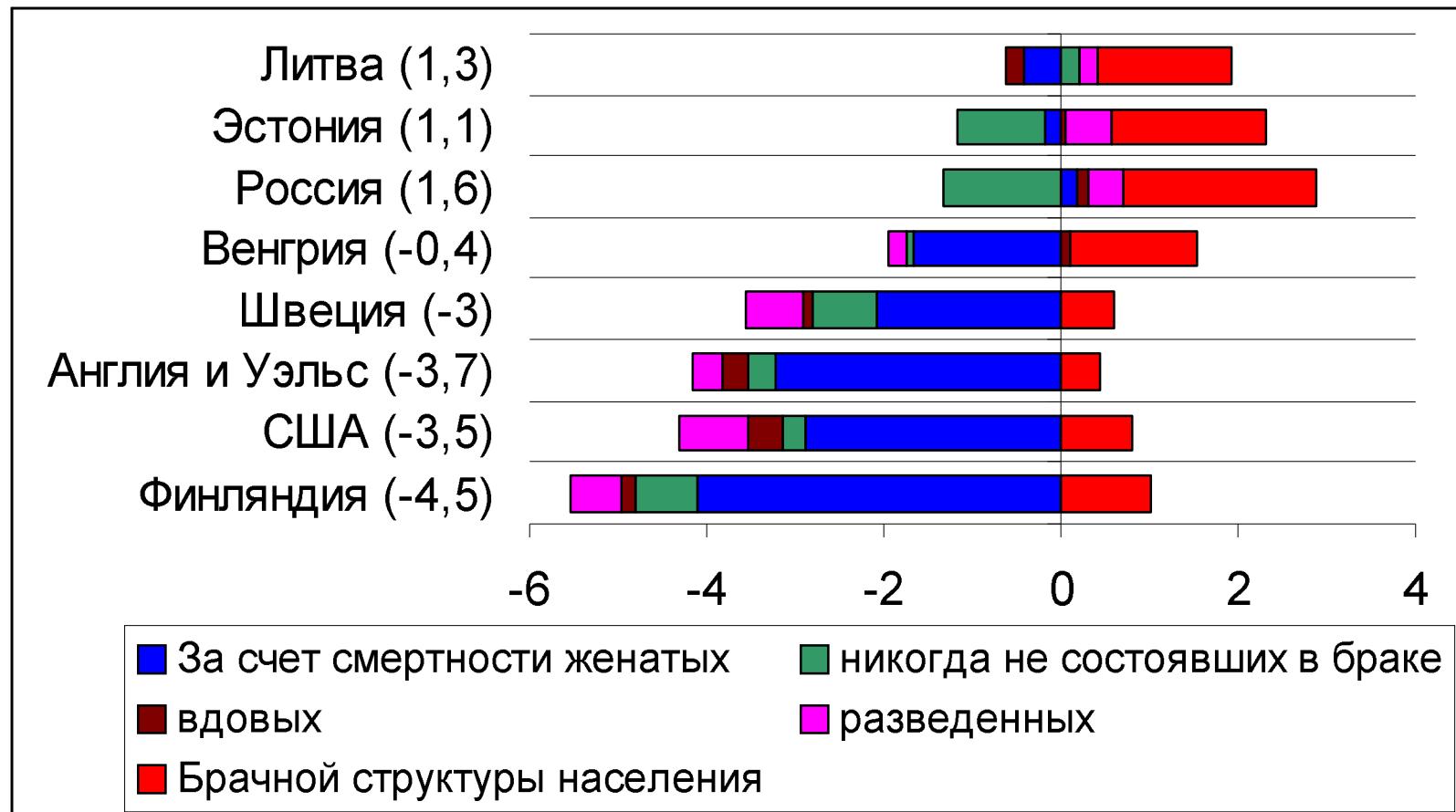
Вклад в рост продолжительности жизни взрослых в 4 странах между ~ 1990 и ~ 2000 годами изменения смертности групп и состава населения по образованию



Разложение изменения продолжительности жизни в 30 лет мужчин в России между 1989 и 1998 гг. по возрастным группам и уровню образования



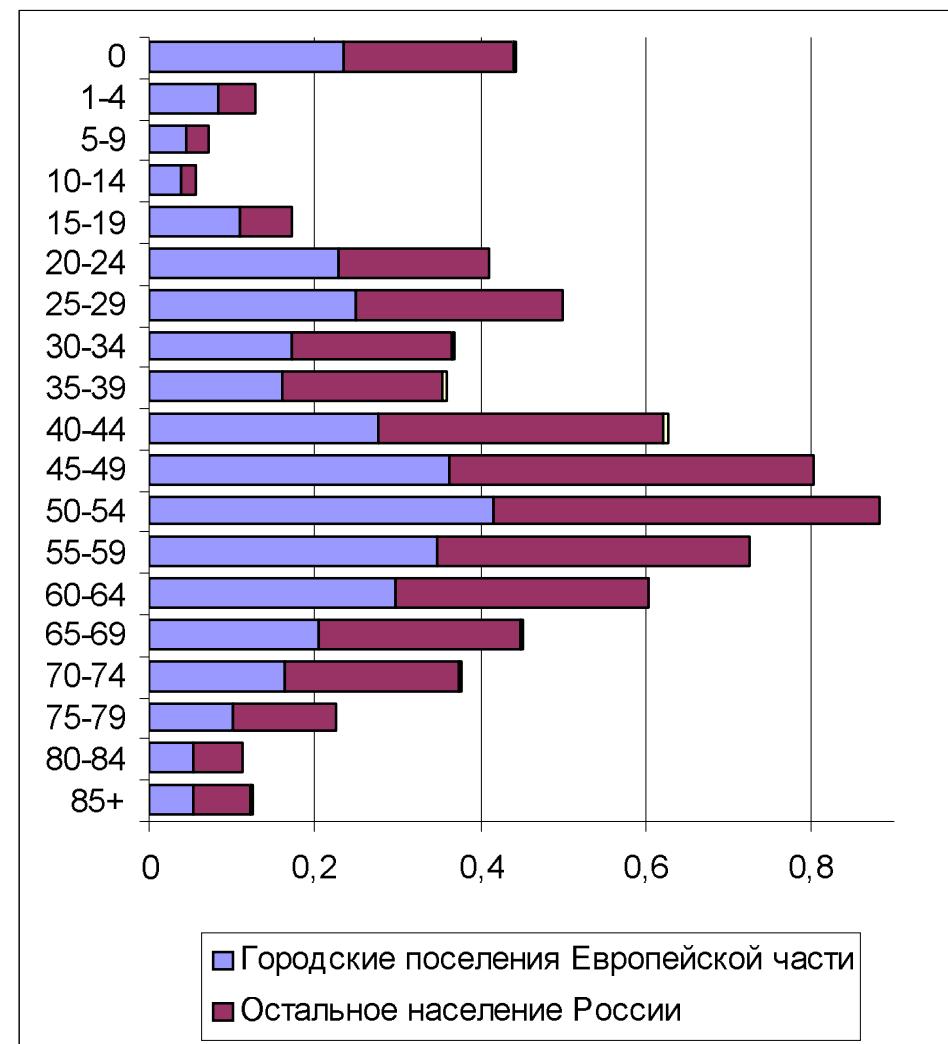
Декомпозиция изменения стандартизованного коэффициента смертности мужчин в возрасте 30-69 лет между ~1980 и ~ 2000 годом



В скобках указано изменение стандартизированного коэффициента смертности за период

Вклад двух групп территорий и возрастных групп в рост продолжительности жизни мужчин в России в 2003-2015 гг.

Около 47% населения России живет в городских поселениях ее Европейской части. В 2003 г там жило 47,5% мужчин, а к 2015 г их стало 46,4%. С 2003 по 2015 г продолжительность жизни мужчин выросла на 6,84 лет, в том числе за счет снижения смертности в городских поселениях Европейской части на 3,27 лет или 47,8% и на 3,57 лет – за счет остальных мужчин. Но вот что интересно, в возрастах до 25 лет вклад городских поселениях Европейской части существенно больше 50%, но после 25 лет – существенно меньше.



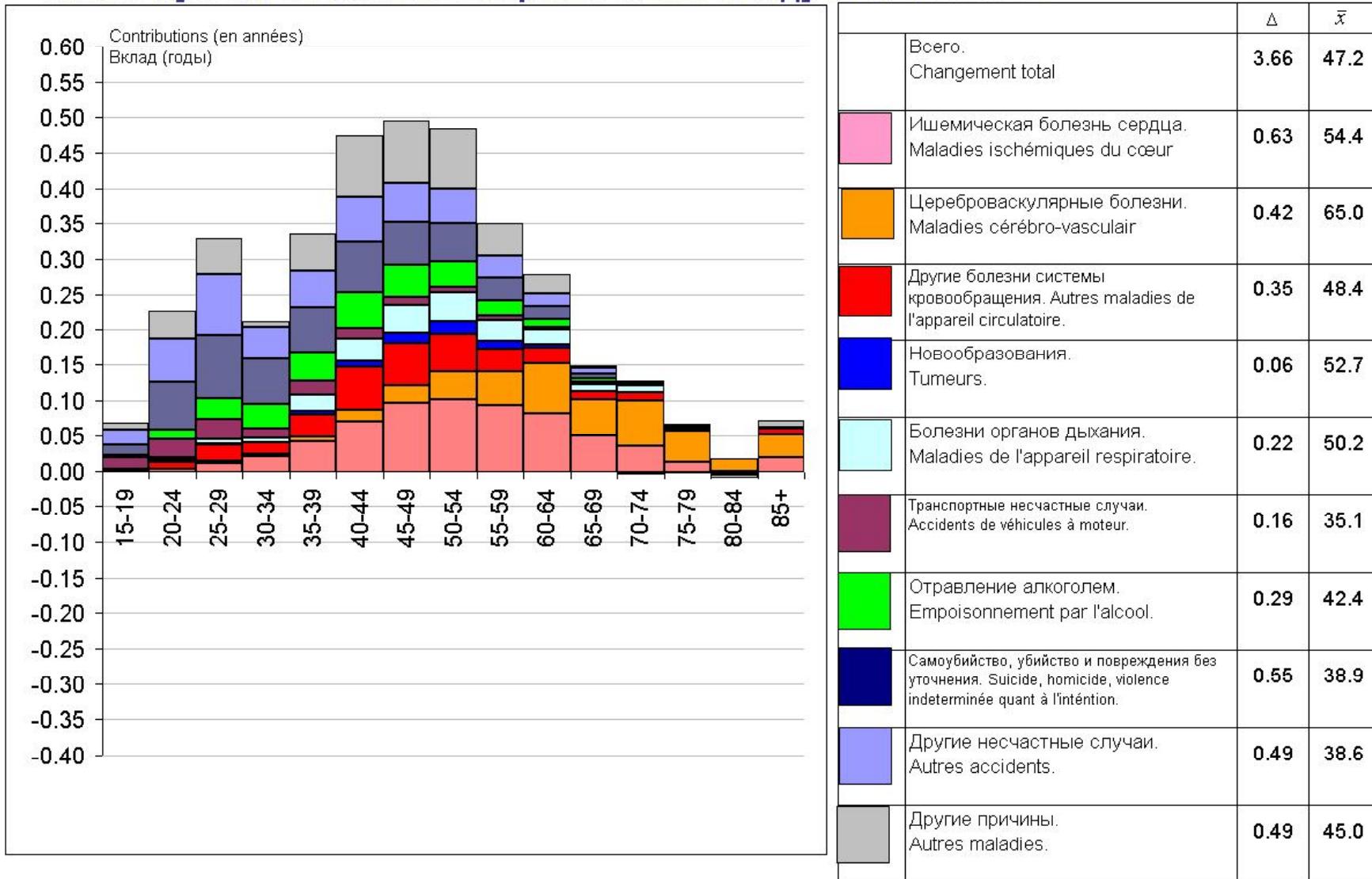
Декомпозиция для визуализации результатов анализа

В заключении я хочу показать несколько иллюстраций из моего доклада для русско-французского коллоквиума в ноябре 2010 г.

Их цель показать, что снижение смертности мужчин в России в 2003-2009 гг. очень похоже на снижение в 1984-1987 и 1994-1998 гг. и совсем не похоже на снижение в Чешской Республике.

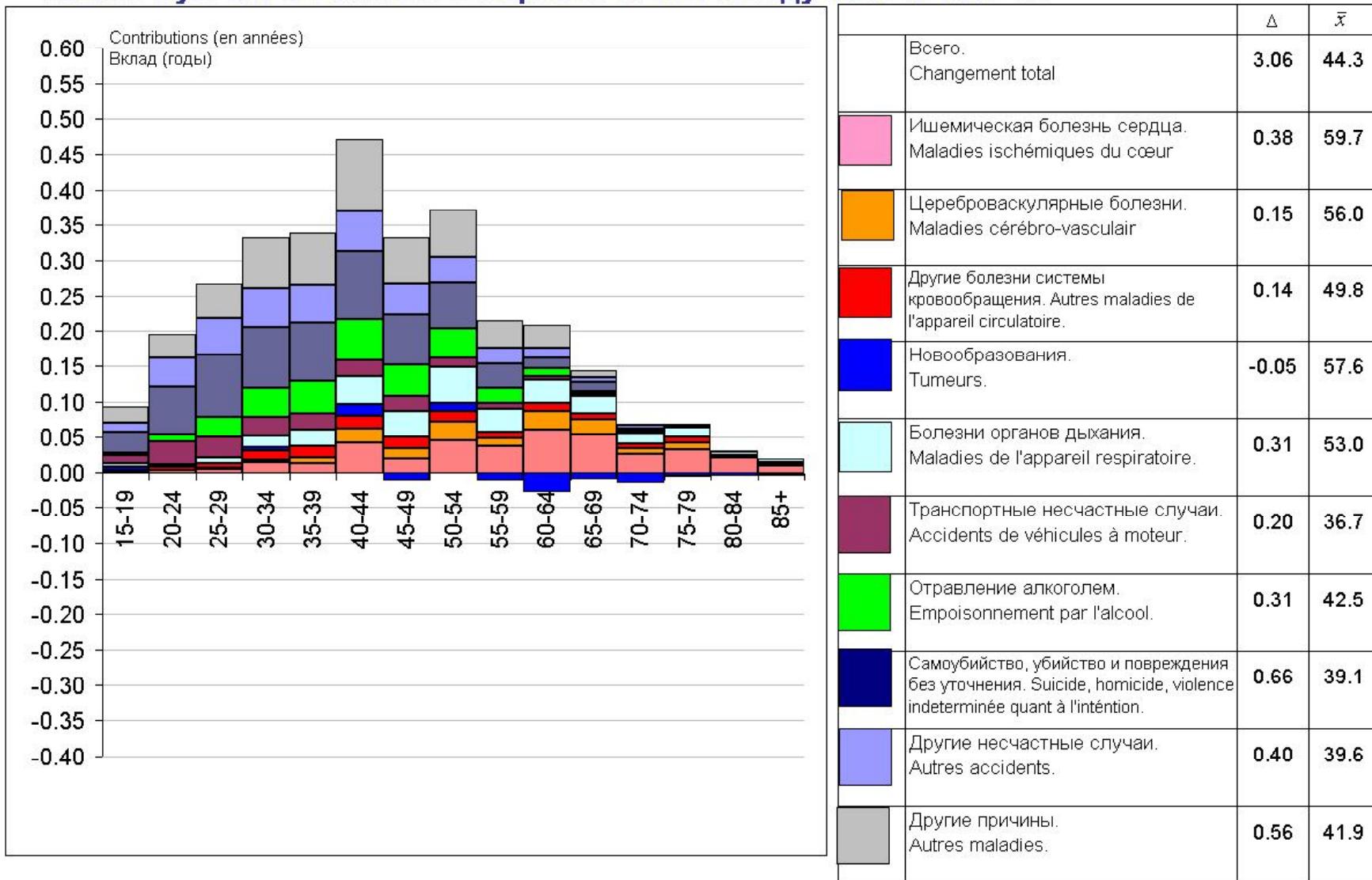
Poids des mortalités par âge et par cause dans aux gains ou pertes d'espérance de vie masculin à 15 ans en Russie entre 2005 et 2009.

Вклад отдельных возрастов и причин смерти в изменение продолжительности жизни мужчин в России в возрасте 15 лет между 2005 и 2009 гг.



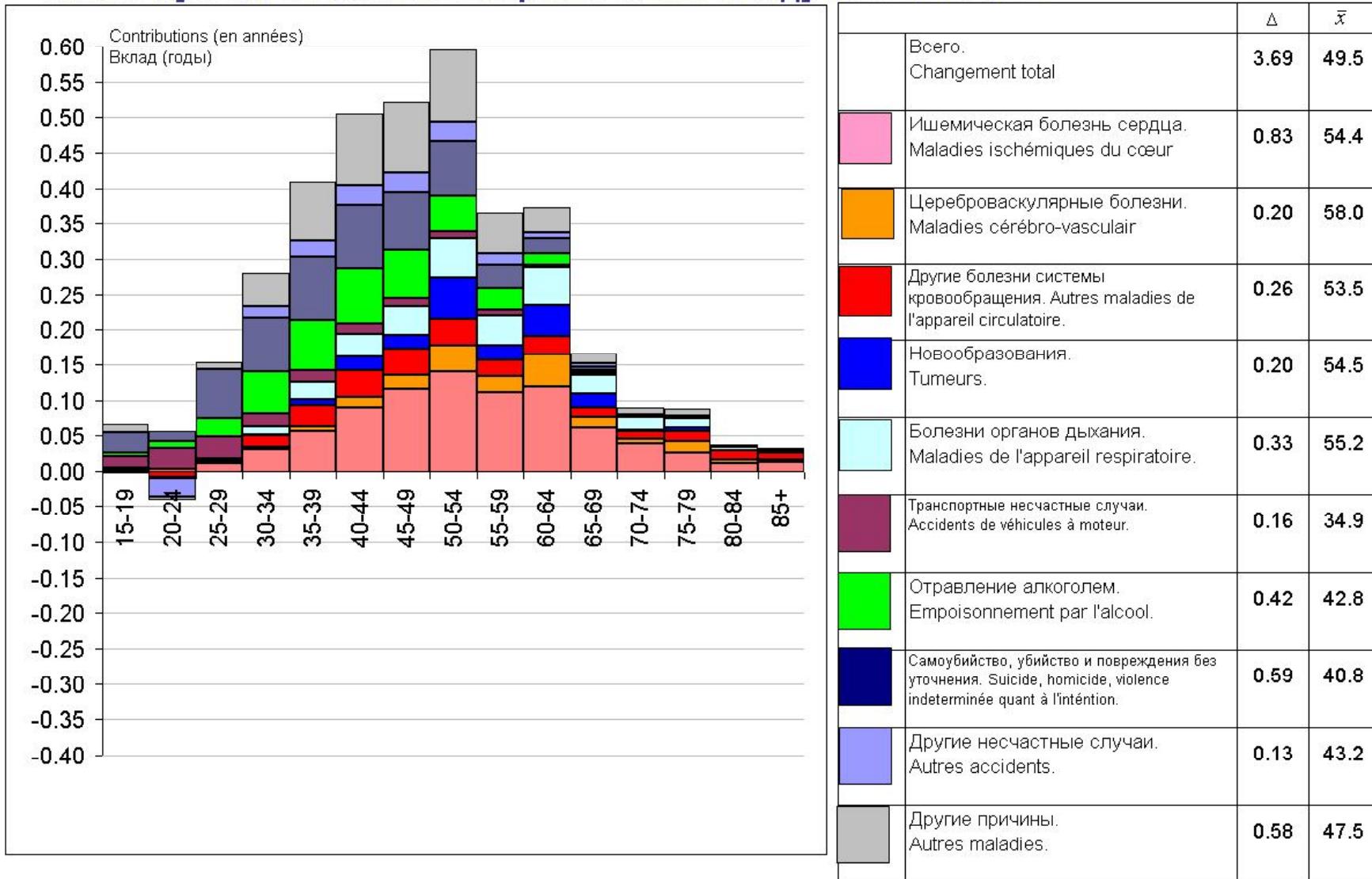
Poids des mortalités par âge et par cause dans aux gains ou pertes d'espérance de vie masculin à 15 ans en Russie entre 1984 et 1987.

Вклад отдельных возрастов и причин смерти в изменение продолжительности жизни мужчин в России в возрасте 15 лет между 1984 и 1987 гг.



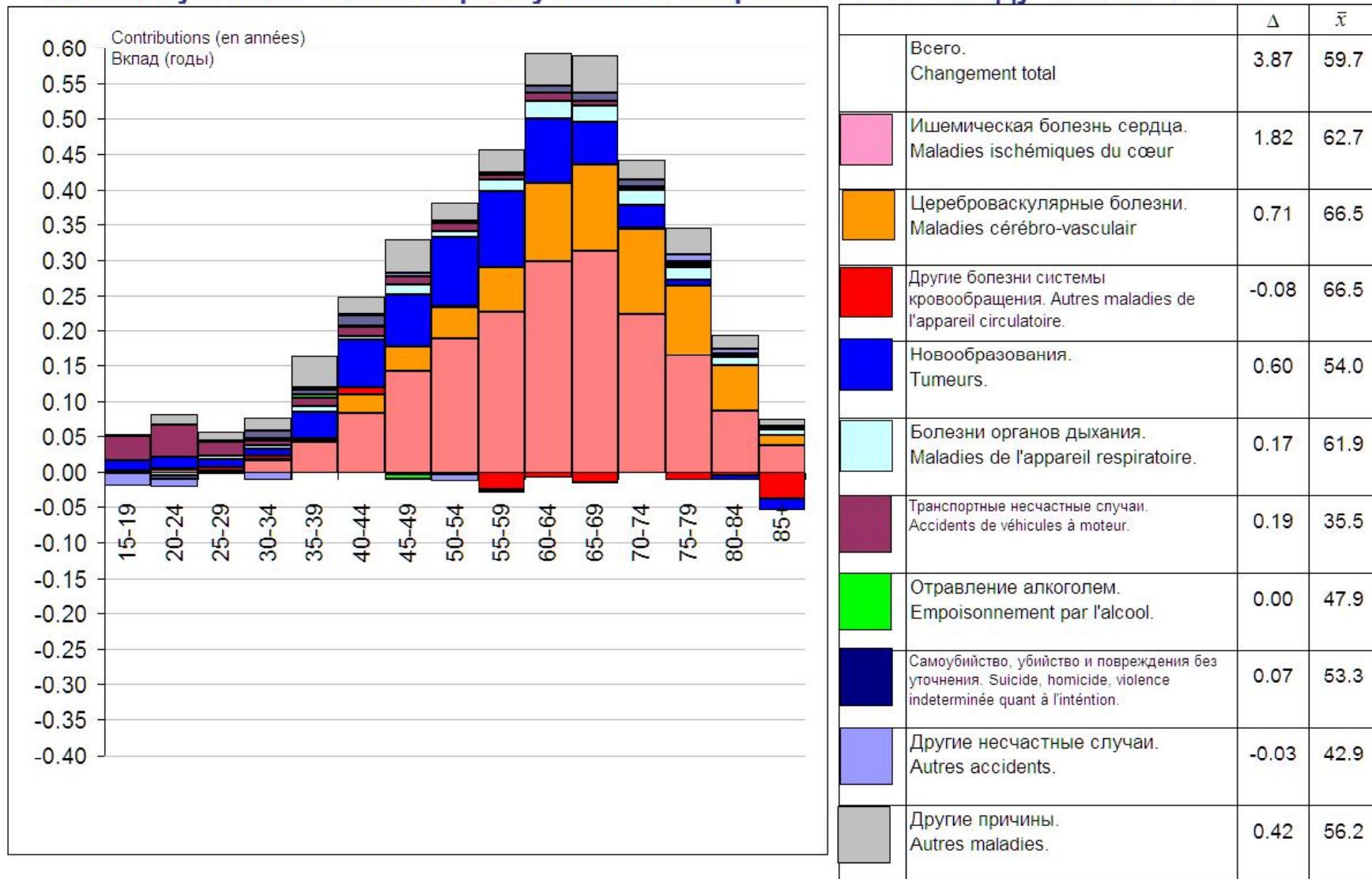
Poids des mortalités par âge et par cause dans aux gains ou pertes d'espérance de vie masculin à 15 ans en Russie entre 1994 et 1998.

Вклад отдельных возрастов и причин смерти в изменение продолжительности жизни мужчин в России в возрасте 15 лет между 1994 и 1998 гг.



Poids des mortalités par âge et par cause dans aux gains ou pertes d'espérance de vie masculin à 15 ans en République tchèque entre 1990 et 2003.

Вклад отдельных возрастов и причин смерти в изменение продолжительности жизни мужчин в Чешской республике в возрасте 15 лет между 1990 и 2003 гг.



Литература

- Arriaga E. 1984. Measuring and explaining the change in life expectancies. *Demography* 21(1), 83-96.
- Beltrán-Sánchez H., Preston S. H., and Canudas-Romo V. 2008. Cause-deleted life tables and decompositions of life expectancy. *Demographic Research*. 19, P. 1323-1350.
- Das Gupta, P. 1994. Standardisation and decomposing of rates from cross-classified data. *Genus*, L(3-4), 171-196.
- Das Gupta, P. 1999. Decomposing the difference between rates when the rate is a function of factors that are not cross-classified. *Genus*, LV(1-2), 9-26.
- Horiuchi S., Wilmoth J. R.. Pletcher S. D. 2008. A Decomposition Method Based on a Model of Continuous Change. *Demography* Vol. 45, No. 4, 785-801.
- Kitagawa, E. 1964. Standardized comparisons in population research. *Demography*, 1, 296-315.
- Pollard J. H. 1982. The expectation of life and its relationship to mortality. *Journal of the Institute of Actuaries*, 109, Part II, No 442, 225-240.
- Pollard J. H. 1988. On the decomposition of changes in expectation of life and differentials in life expectancy. *Demography*, 25, 265-276.
- Ponnappalli K. M. 2005. A Comparison of Different Methods for Decomposition of Changes in Expectation of Life at Birth and Differentials in Life Expectancy at Birth. *Demographic Research* Vol. 12, article 7:141–72.
- Pressat R. 1985. Contribution des écarts de mortalité par âge à la différence des vies moyennes. *Population*, 4-5, 766-770.
- Vaupel, J. W. ; Canudas Romo, V. 2003. Decomposing change in life expectancy: a bouquet of formulas in honor of Nathan Keyfitz's 90th birthday *Demography* 40:2, 201-216.
- Vaupel, J. W. and Canudas Romo, V. 2002. Decomposing demographic change into direct vs. compositional components. *Demographic Research*, 7-1,