Математическая биология и биоинформатика
2016. Т. 11. № 2. С. 367–384. doi: 10.17537/2016.11.367

УДК:519.65

О восстановлении гладких распределений по сгруппированным данным

Авилов К.К.[a,b,1]

[a] *Институт вычислительной математики им. Г.И. Марчука РАН , Москва, Россия*

[b] *ЦНИИ Организации и информатизации здравоохранения МЗ РФ, Москва, Россия*

Аннотация. В статье предложен простой непараметрический метод для восстановления гладких распределений аддитивных величин по сгруппированным данным. Этот метод опирается на требование минимизации меры негладкости решения при условии точного соблюдения групповых сумм, что сводит задачу к задаче квадратичного программирования. Метод был протестирован на данных по возрастному распределению смертности, была показана его точность, сравнимая и превышающая точность метода других авторов. При тестировании на данных по заболеваемости раком были выявлены недостатки и ограничения чисто непараметрического подхода. К преимуществам предложенного метода относятся алгоритмическая и вычислительная простота, достаточная гибкость математической модели.

Ключевые слова: *сгруппированные данные, гладкие распределения, восстановление, гистограммы, задача квадратичного программирования, непараметрические методы.*

ВВЕДЕНИЕ

В биоматематических исследованиях часто возникают ситуации, когда распределение какой-либо величины, зависящей от непрерывного или мелкодискретного параметра (напр., заболеваемости или смертности как функции возраста) описывается сгруппированными данными (напр., количеством заболевших или умерших по 5-летним возрастным группам), а для включения этого распределения в математическую модель требуется именно непрерывный или мелкодискретный его вид. Очевидно, что в общем случае и без использования дополнительных предположений задача по восстановлению распределения по сгруппированным данным решения не имеет. В случае, если есть основания предполагать, что зависимость задается аналитической формулой определенного вида, задача восстановления распределения сводится к задаче параметрической регрессии.

Однако на практике вид распределения, как правило, заранее неизвестен, а потому распределение непараметрически описывается как табличная функция на достаточно мелкой дискретной сетке (напр., по 1-летним возрастным группам). Естественным требованием к значениям табличной функции является соответствие сгруппированным данным. Дополнительные требования, накладываемые на

[1] kkavilov@gmail.com

значения табличной функции, определяют физический смысл задачи оценки значений табличной функции и их взаимосвязь с истинным распределением изучаемой величины. Так, если известна детальная модель истинного распределения, которой истинные значения следуют в точности (напр., если истинное распределение – кусочно-линейное), и количество свободных параметров в которой позволяет однозначно их идентифицировать, то можно говорить о восстановлении истинного распределения на мелкой сетке. Если же можно предполагать только некоторые общие свойства распределения, выполняемые в среднем (напр. гладкость), то задачу определения значений табличной функции с учетом этих свойств следует интерпретировать как нахождение регуляризированного (сглаженного) приближения к истинному распределению, игнорирующего особенности истинного распределения, определяющие отклонения от среднестатистического свойства.

Наиболее распространенным непараметрическим подходом является требование максимизации гладкости решения, определяемое как минимизация нормы разностной производной какого-либо порядка от искомой функции. Такое требование достаточно естественно при работе с величинами, обладающими плавной зависимостью от рассматриваемого параметра. Например, если параметром является время, то требование гладкости естественно для величин с инерционной динамикой или при рассмотрении временны́х интервалов, много бо́льших характерного временно́го масштаба возмущающих факторов. Вместе с тем, поиск наиболее гладкого приближения на мелкой сетке может приводить к заглаживанию особенностей распределения, имеющих масштаб порядка размера ячейки сетки сгруппированных данных. При изучении долгопериодических эффектов (масштаба ячейки сетки сгруппированных данных и более) такое заглаживание может быть полезно и устранять шум в данных, но при изучении более мелкомасштабных эффектов использование восстановленных заглаженных данных как истинных ведет к непоправимым ошибкам и, по сути, к подмене реальности математической моделью.

ПОСТАНОВКА ЗАДАЧИ

Обозначим искомую величину как x, а параметр, от которого она зависит – как a. Пусть искомое распределение определяется значениями $x_j = x(a_j)$ в точках $a_1, \ldots, a_J$. Пусть определена группировка точек a_j, $j = \overline{1,J}$ в I групп, заданная матрицей C размера $I \times J$, чей элемент c_{ij} равен 1, если a_j принадлежит i-ой группе, и равен 0 в противном случае:

$$C = \begin{pmatrix} 1 & \cdots & 1 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & 1 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & 1 & \cdots & 1 \end{pmatrix}.$$

Пусть заданы сгруппированные значения y_i, $i = \overline{1,I}$, причем величина x обладает свойством аддитивности в том смысле, что ее значение для i-ой группы есть сумма значений x_j для всех a_j, принадлежащих i-ой группе:

$$y = Cx, \qquad y = \begin{bmatrix} y_1 \\ \vdots \\ y_I \end{bmatrix}, \qquad x = \begin{bmatrix} x_1 \\ \vdots \\ x_J \end{bmatrix}.$$

Математическая биология и биоинформатика. 2016. Т. 11. № 2. doi: 10.17537/2016.11.367

Требуется найти табличную функцию $\gamma_j = \gamma(a_j)$, $j = \overline{1, J}$ (служащую приближением к x_j), удовлетворяющую некоторым дополнительным условиям (напр., условию максимизации гладкости) и при этом должно выполняться или приближаться условие соответствия входным данным:

$$y \approx C\gamma. \tag{1}$$

СУЩЕСТВУЮЩИЕ ПОДХОДЫ

Один из вариантов решения данной задачи был рассмотрен в работе Rizzi, Gampe и Eilers [1]. В этой работе предполагается, что наблюдаемые отсчеты y_i являются реализациями пуассоновских случайных величин с математическими ожиданиями μ_i, определяемыми через $\mu = C\gamma$, где $\mu = (\mu_1, \ldots, \mu_I)^T$ и $\gamma = (\gamma_1, \ldots, \gamma_J)^T$. Поэтому условие (1) реализуется за счет максимизации функции правдоподобия для пуассоновских случайных величин. Для поддержания неотрицательности величин γ в работе [1] вводится параметризация $\gamma = e^\beta$, $\beta = (\beta_1, \ldots, \beta_J)^T$. Дополнительно на γ накладывается условие гладкости, задаваемое как минимизация "штрафной" квадратичной формы $P_{Ri} = (D_2\beta)^2 = \beta^T D_2^T D_2 \beta$, в которой матрица D_2 является конечно-разностной матрицей вторых производных (размера $(J-2) \times J$):

$$D_2 = \begin{pmatrix} 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & \cdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & -2 & 1 \end{pmatrix}.$$

В итоге в работе [1] задача формулируется как максимизация логарифмической функции правдоподобия со штрафом, в которую штрафная квадратичная форма входит с весом λ:

$$\beta^* = \underset{\beta}{\operatorname{argmax}} \left(\sum_{i=1}^{I} (y_i \ln \mu_i(\beta) - \mu_i(\beta)) - \frac{\lambda}{2} \beta^T D_2^T D_2 \beta \right). \tag{2}$$

Таким образом, в постановке задачи, использованной Rizzi с соавт., возникает конкуренция (т.н. trade-off) между точностью приближения входных данных и минимизацией негладкости. Параметр λ, управляющий этой конкуренцией, авторы [1] предлагают находить за счет минимизации информационного критерия Акаике (AIC), вычисленного для решений β^*, получаемых для спектра значений λ (т.е. перебором по сетке по λ).

В качестве вычислительного примера в статье [1] использовались данные по количеству смертей от различных причин в Соединенных Штатах Америки (США) в 2009 году, полученные с сайта Центра по контролю заболеваемости (CDC) [2]. Эти данные предоставляются по 1-летним возрастным группам (x), но для вычислительного эксперимента данные были просуммированы по 5-летним возрастным группам от 0 до 85 лет и сформирована общая группа "85 лет и старше" (y). По сгруппированным данным y был произведен расчет оценок γ по 1-летним группам и они графически сравнивались с истинными значениями x. В целом, совпадение γ с x можно охарактеризовать как весьма хорошее, но авторы [1] не проводили детальной количественной оценки качества приближения (как по совпадению γ с x, так и по совпадению μ с y).

МОТИВАЦИЯ АЛЬТЕРНАТИВНОГО ПОДХОДА

При попытке повторить метод Rizzi с соавт. [1] было обнаружено, что в нем использовалось расширенное определение AIC с вычислением эффективной размерности модели, тесно связанным с используемым в [1] численным методом для максимизации функции правдоподобия со штрафом. Недостаточная документированность вычислительного метода и расхождения между представленным в [1] программным кодом на языке R [3] и его текстовым описанием не позволили независимо воспроизвести вычисление эффективной размерности модели. Кроме того, в [1] (как в тексте, так и в вычислительном коде), по-видимому, имеется ошибка в определении AIC[2]. При этом вычисление AIC является строго необходимым в методе Rizzi с соавт., поскольку без него невозможно определение оптимального значения весового параметра λ.

Общая схема метода Rizzi с соавт., опирающаяся на двухэтапную оптимизацию, может представлять значительную вычислительную сложность при использовании простых вычислительных методов, требующих программирования только целевой функции. Однако программный код на языке R, представленный в статье [1], реализует весьма эффективный вычислительный метод, опирающийся на выполненную аналитически (т.е. вручную) линеаризацию целевой функции, а потому при использовании именно этой программной реализации метода проблемы вычислительной сложности маловероятны.

Помимо этого, результаты расчетов как при помощи предоставленного в [1] программного кода на языке R, так и при помощи собственного программного кода, использующего оптимальные значения λ, указанные в [1], показали, что для некоторых 5-летних возрастных групп метод Rizzi с соавт. дает весьма значительные отклонения групповых сумм μ от данных y (до $\approx 40\%$).

Все эти обстоятельства привели к выводу о необходимости разработки более простого метода решения задачи восстановления распределений, лишенного недостатков метода Rizzi с соавт.

КП-МЕТОД

В данной работе предлагается альтернативный подход к задаче восстановления распределений, позволяющий избавиться от конкуренции (trade-off) между требованием точности совпадения групповых сумм μ с y и требованием минимизации негладкости получаемой табличной функции. Этот метод опирается на требование точного выполнения равенства (1) и использование его как условия при минимизации меры негладкости решения γ:

$$\begin{cases} P_{QP}(\gamma) = \gamma^T D^T D\gamma \to \min, \\ y = C\gamma, \end{cases} \tag{3}$$

где D есть некоторая матрица размера $K \times J$, задающая критерий гладкости. В качестве матрицы D может выступать матрица D_2. При необходимости система (3) может быть дополнена явным требованием неотрицательности решения, заменяющим нелинейную параметризацию γ через β в методе Rizzi с соавт.:

$$\gamma_j \geq 0, \quad j = \overline{1, J}. \tag{4}$$

[2]В [1] отклонение (deviance) модели определяется как $Dev(y|\mu) = 2\sum_{i=1}^{I} y_i \ln \frac{y_i}{\mu_i}$, тогда как из определения отклонения (deviance) и из вида функции правдоподобия для распределения Пуассона следует $Dev(y|\mu) = 2\sum_{i=1}^{I} \left(y_i \ln \frac{y_i}{\mu_i} - \mu_i \right)$. Т.к. μ_i в данной задаче *почти* фиксированы, разница двух определений может быть невелика. Но нельзя исключать и обратную ситуацию.

Нетрудно видеть, что задача (3) является классической задачей квадратичного программирования с ограничениями вида равенства. Поэтому далее нашу модель мы будем называть КП-метод.

Необходимое условие минимума $P_{QP}(\boldsymbol{\gamma})$ может быть получено по методу множителей Лагранжа. Это условие оказывается системой линейных алгебраических уравнений (СЛАУ), которая имеет следующую блочно-матричную форму (здесь $\boldsymbol{\lambda}$ есть вектор множителей Лагранжа длины I):

$$\begin{pmatrix} D^T D & C^T \\ C & 0 \end{pmatrix} \begin{bmatrix} \boldsymbol{\gamma} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix}. \tag{5}$$

Для того, чтобы решение (5) было также и достаточным условием экстремума $P(\boldsymbol{\gamma})$, требуется, чтобы его решение было единственным. Обычно это гарантируется требованием положительной определенности матрицы квадратичной формы ($D^T D$ в данном случае). Однако при использовании матриц D, являющихся конечно-разностными приближениями производных какой-либо степени, матрица квадратичной формы оказывается лишь неотрицательно определенной. В этом случае решение (5) единственно при дополнительном условии

$$\boldsymbol{\gamma}^T D^T D \boldsymbol{\gamma} > 0 \quad \forall \boldsymbol{\gamma} : C\boldsymbol{\gamma} = 0. \tag{6}$$

Для используемых в данной работе матриц D и C условие (6) выполняется, а потому уравнения (5) дают единственное решение, являющееся решением задачи (3).

Если решение (5) дает отрицательные значения $\boldsymbol{\gamma}$, недопустимые по физическому смыслу задачи, необходимо введение дополнительных условий (4). В этом случае задача (3, 4) является задачей квадратичного программирования с ограничениями смешанного вида (равенства и неравенства). Необходимые условия экстремума в этом случае задаются условиями Каруша-Куна-Таккера [4, 5], но получающаяся система уравнений и неравенств, по-видимому, не имеет такого же простого решения, как СЛАУ (5). Поэтому множеством авторов было разработано большое количество итеративных методов численного решения задачи квадратичного программирования со смешанными ограничениями. В силу учета вида задачи, эти методы, как правило, имеют высокую вычислительную эффективность.

ВЫЧИСЛИТЕЛЬНЫЕ ЭКСПЕРИМЕНТЫ И СРАВНЕНИЕ ПОДХОДОВ

Данные

В вычислительных экспериментах использовались те же наборы данных, что и в статье Rizzi с соавт. [1]: количество смертей от болезней системы кровообращения (коды I00-I99 по Международной классификации болезней 10-го пересмотра (МКБ-X)), от новообразований (C00-D48), от болезней крови, кроветворных органов и отдельных нарушений, вовлекающих иммунный механизм (D50-D89), от некоторых инфекционных и паразитарных заболеваний (A00-B99) в Соединенных Штатах Америки в 2009 году. Данные были загружены с сайта CDC [2] в виде таблиц по 1-летним возрастным группам: “$<$ 1 года”, “1 год”, “2 года”, ..., “99 лет”, “100 лет и старше”.

Сгруппированные данные формировались так же, как и в [1]: использовались 5-летние возрастные группы “0–4 года”, “5–9 лет”, ..., “80–84 года”, широкая группа “85–114 лет” и добавочная виртуальная группа “115 лет и старше”. Широкая группа “85–114 лет” на практике соответствует группе “85 лет и старше”, поскольку

предполагается, что продолжительность жизни не превышает 115 лет. Виртуальная группа “115 лет и старше” с нулевым количеством смертей добавляется чтобы внести в модель предположение о том, что $x(a) \to 0$ при $a \to \infty$. В целях графического представления данных было также предположено, что смерти в группе “100 лет и старше” равномерно распределены между возрастами 100–114 лет (на сгруппированные данные это никак не влияло, т.к. смерти перераспределялись внутри одной группы “85–114 лет”).

Программная реализация

Для проведения вычислительных экспериментов была воспроизведена модель Rizzi с соавт. (2), но для численного поиска экстремума использовалась библиотечная реализация квази-ньютоновского алгоритма Бройдена-Флетчера-Гольдфарба-Шанно (BFGS) [6]. Было проведено сравнение с результатами расчетов при помощи программного кода на языке R, предоставленного Rizzi с соавт.: результаты совпадают с относительной точностью, используемой в численных методах в качестве целевой (10^{-6}).[3]

Использование чистой КП-модели (3) приводило к неестественным отрицательным значениям количества умерших. Поэтому использовалась дополненная КП-модель (3, 4). Для ее настройки на данные использовалась библиотечная функция для решения задачи квадратичного программирования, реализующая метод внутренней точки [7]. Для сравнения с результатами Rizzi с соавт. в модели (3, 4) в качестве матрицы D использовалась матрица D_2, что соответствует минимизации среднеквадратичного значения второй производной от γ по a.

Результаты расчетов

Результаты расчетов по обеим моделям и по всем четырем наборам данных представлены на рисунке 1–4. В качестве значений весового коэффициента λ в модели Rizzi с соавт. использовались оптимальные значения для каждого набора данных, приведенные в [1]. В интервале возрастов от 15–20 и до 80–85 лет обе модели дают практически неотличимые результаты. Определенные различия наблюдаются в областях с малым количеством смертей (0–15 лет) и в области широкой группы “85 лет и более”.

Модель Rizzi с соавт., как отмечалось выше, может давать отклонения от групповых сумм y. На рисунке 5 показана относительная невязка групповых сумм $((y - \mu)/y)$ для модели Rizzi с соавт. Невязка групповых сумм для КП-модели не показана, т.к. отклоняется от нуля в пределах машинной точности вычислений. Следует отметить, что модель Rizzi с соавт. значительно (до $\approx 40\%$) искажает количество смертей в младших возрастных группах, что может быть значимо для исследований, связанных с этими группами.

Для анализа качества восстановления 1-летних данных x по групповым сумммам y были построены графики относительной ошибки $(x - \gamma)/x$ (рис. 6) и вычислены среднеквадратические значения абсолютной и относительной ошибок ($(x - \gamma)$ и $(x - \gamma)/x$ соответственно) в интервале возрастов 0–99 лет (табл. 1, 2). Также были произведены эксперименты с КП-моделью с использованием сглаживающих матриц

[3] Следует отметить, что в силу, по-видимому, небольшой ошибки, допущенной Rizzi с соавт. в их коде, одинаковые результаты расчетов при одинаковом значении коэффициента λ достигаются тогда, когда в воспроизведенной модели весовой множитель в (2) задается как $\sqrt{\lambda}$, а не как $\frac{\lambda}{2}$.

Математическая биология и биоинформатика. 2016. Т. 11. № 2. doi: 10.17537/2016.11.367

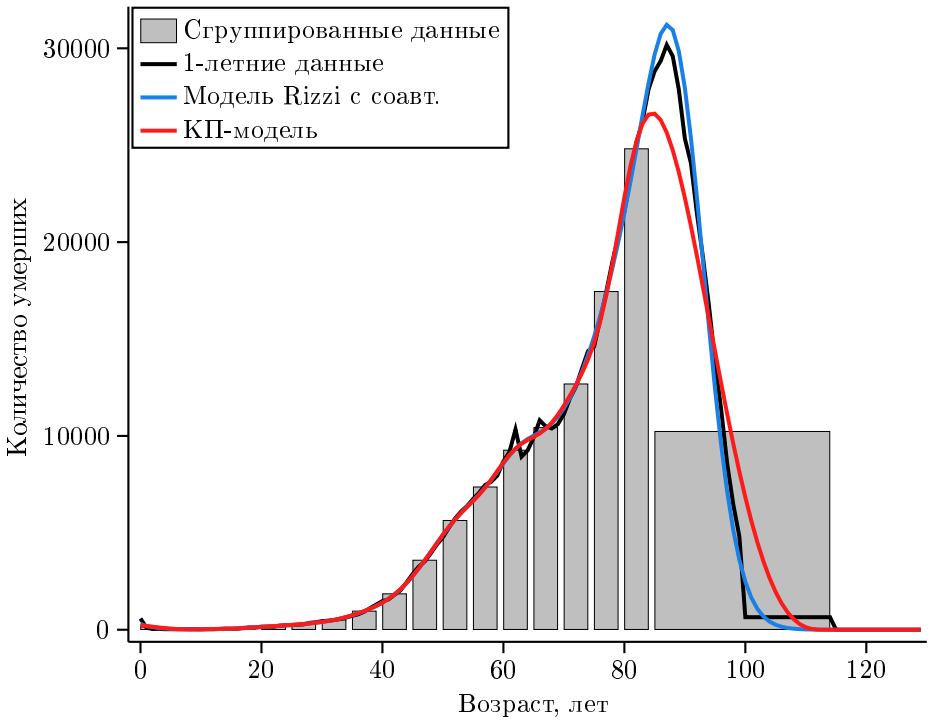

Рис. 1. Результаты восстановления повозрастного распределения количества смертей от болезней системы кровообращения (I00-I99 по МКБ-X) в США в 2009 г.

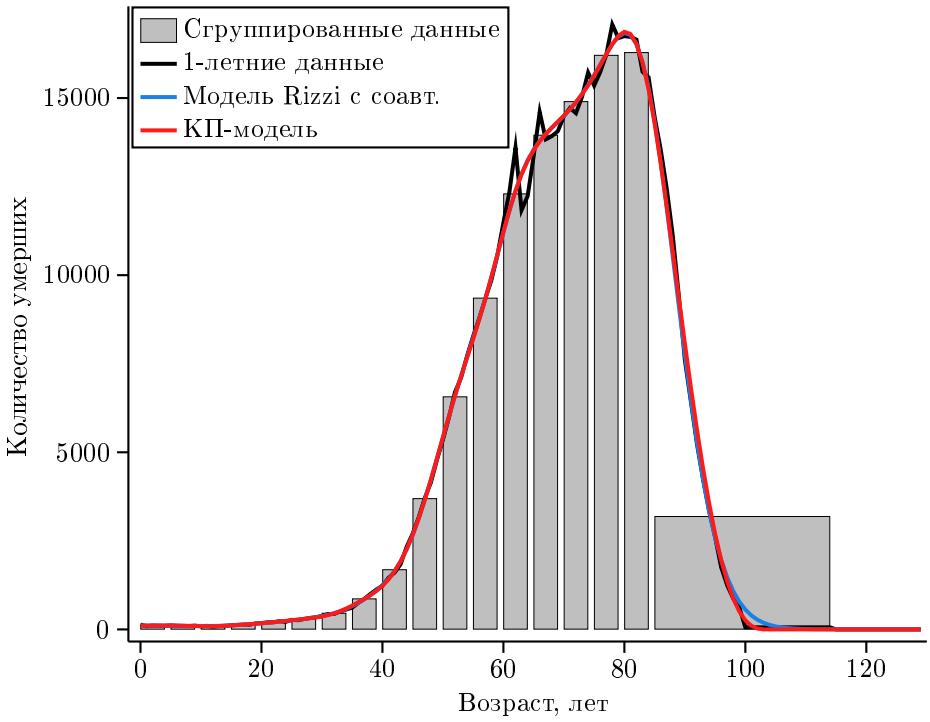

Рис. 2. Результаты восстановления повозрастного распределения количества смертей от новообразований (C00-D48 по МКБ-X) в США в 2009 г.

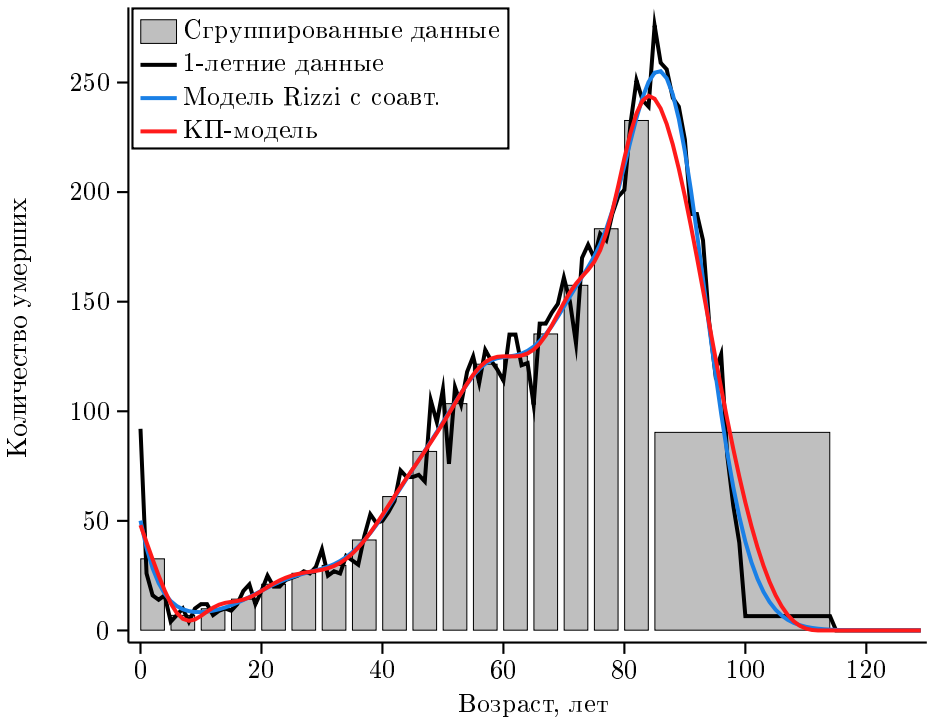

Рис. 3. Результаты восстановления повозрастного распределения количества смертей от болезней крови, кроветворных органов и отдельных нарушений, вовлекающих иммунный механизм (D50-D89 по МКБ-X) в США в 2009 г.

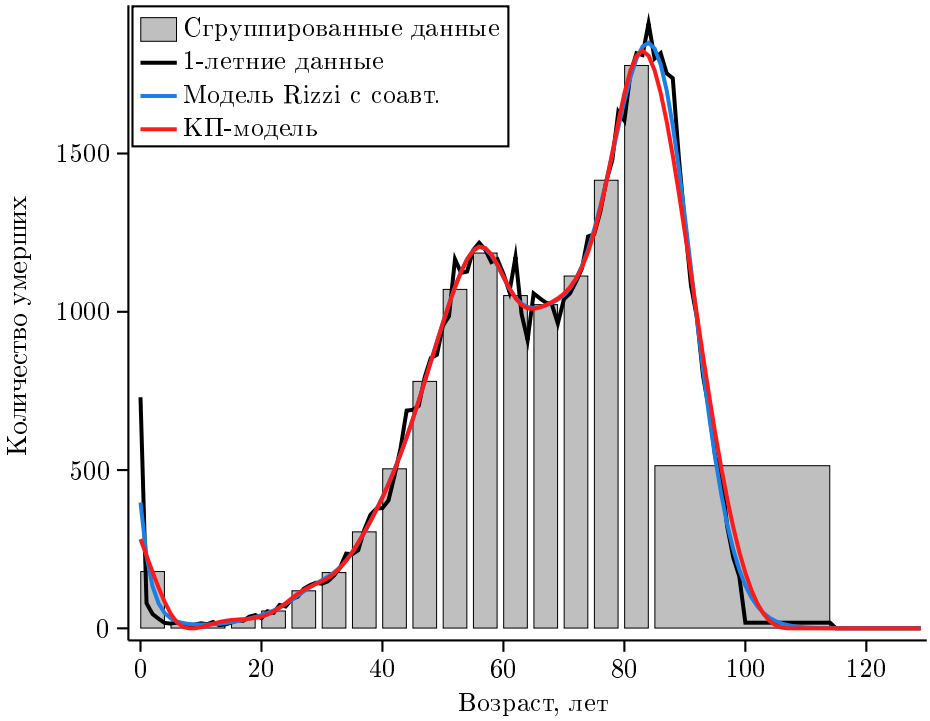

Рис. 4. Результаты восстановления повозрастного распределения количества смертей от некоторых инфекционных и паразитарных заболеваний (A00-B99 по МКБ-X) в США в 2009 г.

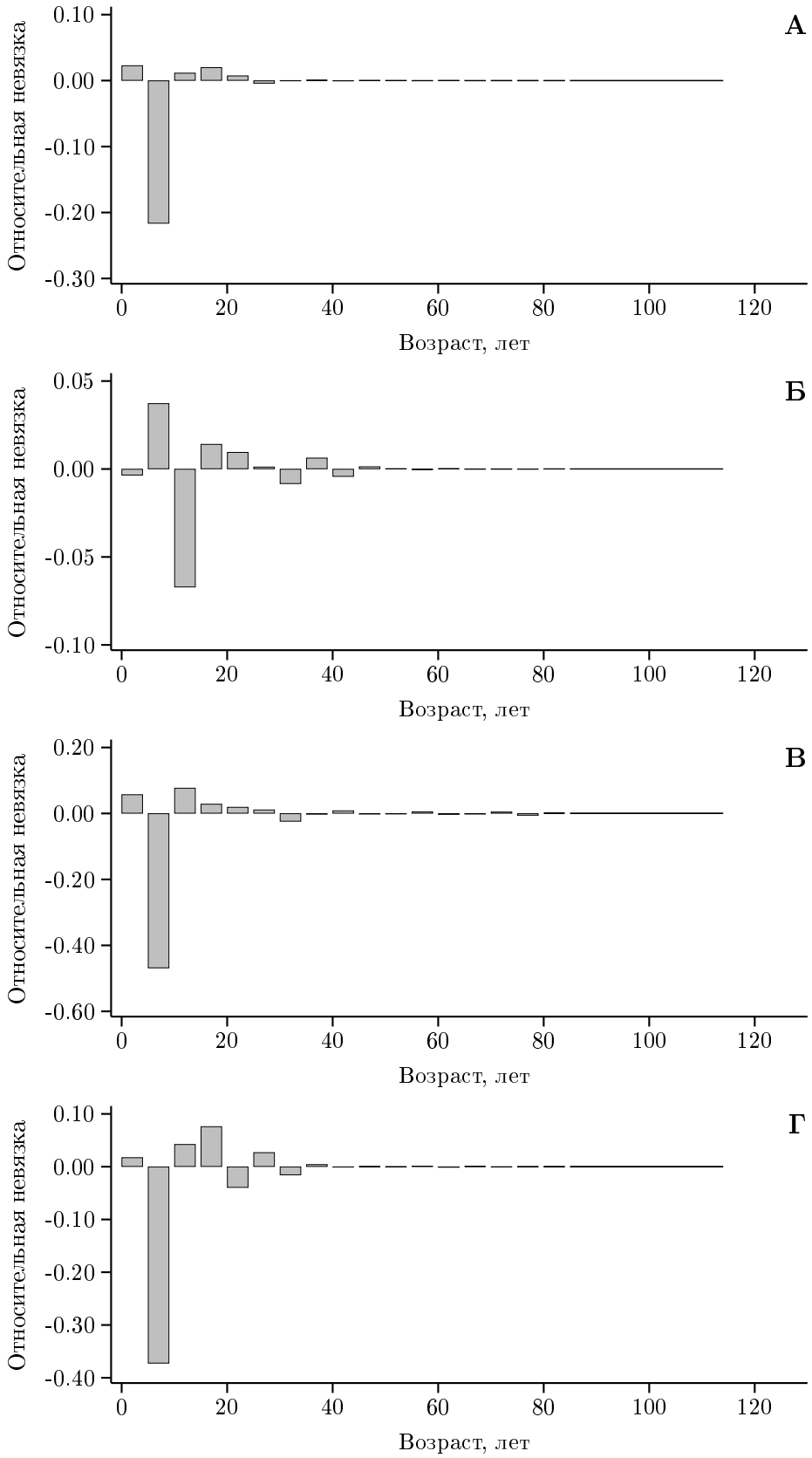

Рис. 5. Относительная невязка воспроизведения групповых сумм моделью Rizzi с соавт. $((y-\mu)/y)$. Гистограмма А соответствует результатам на рис. 1, Б – результатам на рис. 2, В – результатам на рис. 3, Г – результатам на рис. 4.

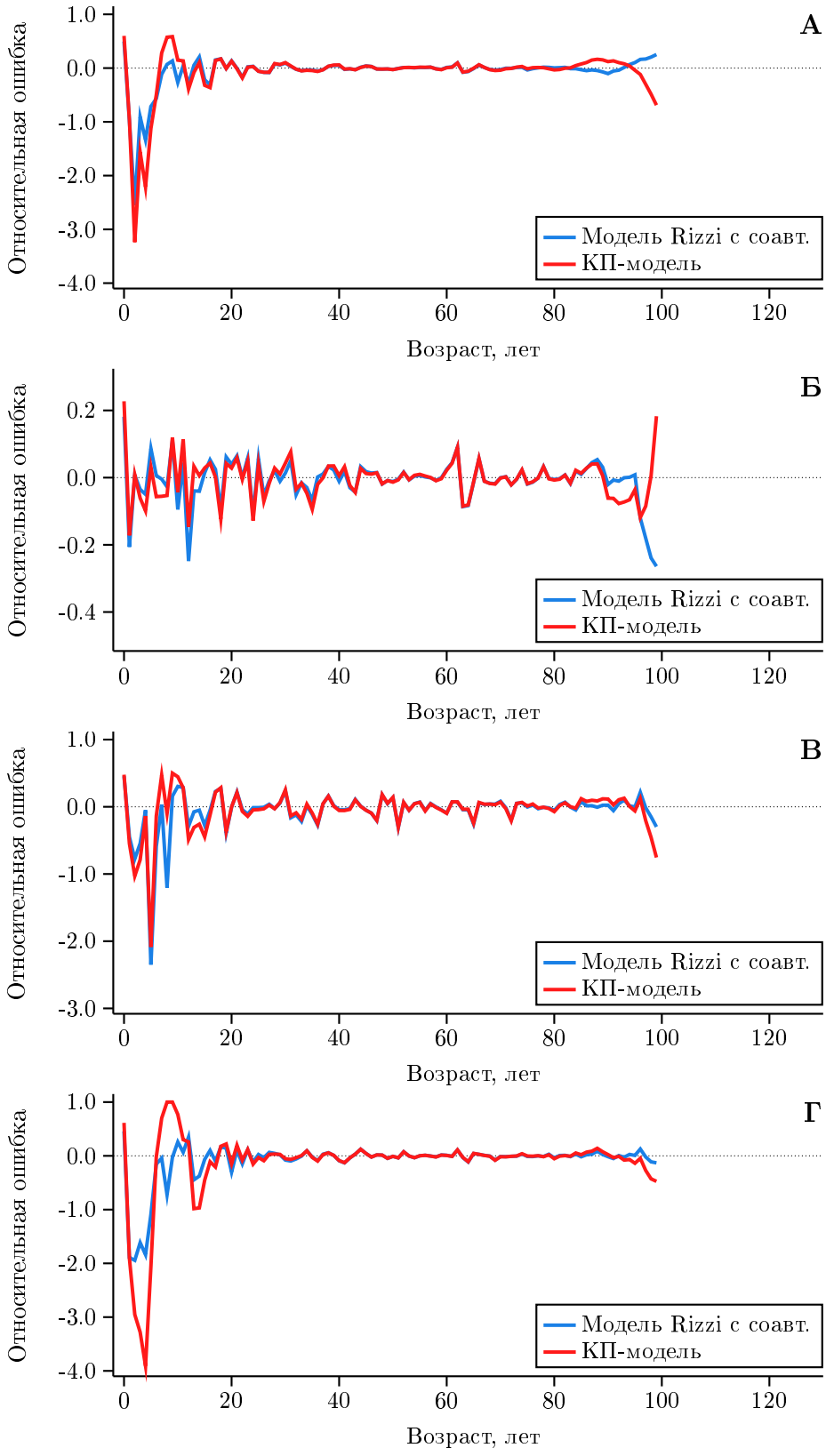

Рис. 6. Относительная ошибка воспроизведения 1-летних данных $((x - \gamma)/x)$ на интервале 0-99 лет. График А соответствует результатам на рис. 1, Б – результатам на рис. 2, В – результатам на рис. 3, Г – результатам на рис. 4.

Таблица 1. Среднеквадратические абсолютные ошибки восстановления 1-летних данных по смертности в США в 2009 г. по четырем классам причин в интервале возрастов 0–99 лет. Абсолютная ошибка определялась как $(x - \gamma)$. Для КП-метода представлены результаты расчетов для различных сглаживающих матриц D

	Абс.ош., Rizzi	Абсолютная ошибка, КП-метод			
	D_2	D_1	D_2	D_3	D_4
А) Б-ни сист. кровообращения	58.74	227.28	118.11	71.62	48.13
Б) Новообразования	26.16	42.67	26.58	25.65	26.06
В) Б-ни крови, кроветворн. органов и отд. нарушения, вовлекающие иммунный мех-м	0.97	1.97	1.22	0.98	0.92
Г) Некоторые инфекционные и паразитарные заболевания	5.13	11.13	7.05	6.16	5.94

Таблица 2. Среднеквадратические относительные ошибки восстановления 1-летних данных по смертности в США в 2009 г. по четырем классам причин в интервале возрастов 0–99 лет. Относительная ошибка определялась как $(x - \gamma)/x$. Для КП-метода представлены результаты расчетов для различных сглаживающих матриц D

	Отн.ош., Rizzi	Относительная ошибка, КП-метод			
	D_2	D_1	D_2	D_3	D_4
А) Б-ни сист. кровообращения	0.0345	0.0579	0.0479	0.0418	0.0379
Б) Новообразования	0.0067	0.0163	0.0059	0.0072	0.0093
В) Б-ни крови, кроветворн. органов и отд. нарушения, вовлекающие иммунный мех-м	0.0319	0.0389	0.0314	0.0261	0.0234
Г) Некоторые инфекционные и паразитарные заболевания	0.0403	0.0883	0.0702	0.0640	0.0603

$D_1, \ldots, D_4$, соответствующих разностной аппроксимации 1-ой,..., 4-ой производной (см. приложение 1).

При использовании сглаживающей матрицы D_2 КП-модель по абсолютной ошибке проигрывает модели Rizzi с соавт., но по относительной ошибке КП-модель на двух наборах данных из четырех опережает модель Rizzi с соавт. С ростом порядка дифференцирования в матрице D в большинстве случаев ошибки КП-модели значительно снижаются и при использовании матриц D_3, D_4 в половине случаев оказываются лучше ошибок модели Rizzi с соавт., использующей матрицу D_2.

ОБСУЖДЕНИЕ

Различие результатов расчетов по двум моделям при использовании одной и той же сглаживающей матрицы D_2 можно объяснить не только разными подходами к оптимизации целевой функции, но и структурными различиями целевых функций. Модель Rizzi с соавт. требует минимизации меры негладкости для коэффициентов β ($P_{Ri} = \beta^T D_2^T D_2 \beta$), связанных с восстанавливаемыми значениями γ соотношением $\gamma = e^{\beta}$. КП-модель же требует минимизации негладкости самих восстанавливаемых значений γ ($P_{QP}(\gamma) = \gamma^T D^T D \gamma$). В результате модель Rizzi с соавт. стремится к приближению данных экспоненциальными функциями от возраста, а КП-модель с матрицей D_2 – линейными. Для данных по смертности в развитых странах естественно ожидать близость профиля смертности к распределению Гомпертца, имеющему именно экспоненциальный характер. Поэтому подход Rizzi с соавт. более успешен в восстановлении хвостов распределений количества смертей по возрастам.

Кроме того, обе рассмотренные модели плохо воспроизводят повышенную младенческую смертность (т.е. резкий рост смертности в возрастах 0–1 года), заменяя ее более плавным подъемом в возрастах 0–5 лет, а иногда и старше.

Эти два обстоятельства вкупе с тем, что КП-модель значительно улучшает ошибку восстановления с ростом порядка матрицы D для большинства, но не для всех наборов данных, подчеркивают необходимость более глубокого анализа того, какие модели более эффективны для восстановления распределений каждого типа, а также того, как эффективно вносить дополнительные априорные знания (например, о факте существования повышенной младенческой смертности) в вычислительную модель. В работе [1] и вслед за ней в данной работе применялся простейший прием для внесения априорной информации – создание виртуальной возрастной группы “115–130 лет” с нулевым количеством смертей, фактически эквивалентное краевому условию $\gamma(115) = 0$. Более сложные условия могут существенным образом изменять структуру модели, что особенно важно для КП-модели, т.к. она может потерять “квадратичность” целевой функции.

В случае восстановления количества смертей более естественным было бы предполагать максимальную гладкость не кривых количества умерших по возрастам, а повозрастных удельных коэффициентов смертности. Однако коэффициенты смертности не обладают свойством аддитивности внутри возрастных групп. В работе Rizzi с соавт. [1] предложен следующий подход: сначала следует получить или восстановить количество населения по 1-летним возрастным группам (обладающее свойством аддитивности), а затем внести эти количества в качестве суммирующих коэффициентов в матрицу C и настраивать модель на сгруппированные данные по количеству умерших – при этом переменные γ будут соответствовать повозрастным коэффициентам смертности.

ДОПОЛНИТЕЛЬНЫЕ ТЕСТЫ

Для проверки эффективности КП-модели был проведен ряд вычислительных экспериментов на данных по онкологическим заболеваниям и численности населения.

В первом тесте были использованы данные по численности постоянного женского населения Российской Федерации с разбивкой по 1-летним возрастным группам (от "$<$ 1 года" и до "100 лет и старше") на 1 января 2014 г. [8] Аналогично анализу данных по количеству умерших, данные были сгруппированы в 5-летние группы и открытую группу "85 лет и старше". Была добавлена виртуальная группа "115–130 лет" с нулевым значением. Затем был применен КП-метод с $D = D_2$. Результаты показаны на рисунке 7. В целом восстановление численности населения можно охарактеризовать как неплохое, хотя метод заглаживает резкие колебания численности населения в районе 65–75 лет (связанные с демографической динамикой во время Великой Отечественной войны и после нее).

Во втором тесте рассматривались данные по количеству случаев заболевания злокачественными новообразованиями молочной железы (ЗНО МЖ) у женщин, постоянно проживавших в Российской Федерации в 2014 г. [9] Эти данные собираются в рамках Формы № 7 Государственного статистического наблюдения, а потому отражают возрастную структуру только по 5-летним группам (от "0–4 года" и до "85 лет и старше"). Результаты применения КП-метода с $D = D_2$ к этим данным (с добавлением "зануляющей" виртуальной группы "115–130 лет") показаны на рисунке 8. В старших возрастных группах наблюдаются осцилляции количества заболевших, существование которых в реальности маловероятно. Применение сглаживающих матриц D_3 и D_4 дает похожие результаты (не показаны). Применение матрицы D_1 (не показано) приводит к монотонному решению в области старших возрастов, но решение начинает иметь форму, близкую к кусочно-линейной.

Для оценки удельных (на душу насления) частот заболевания ЗНО МЖ было применено три подхода (рис. 9):

1. Наивный подход: восстановленные 1-летние количества случаев ЗНО МЖ были поделены на известные количества женщин по 1-летним группам.

2. Подход, предложенный Rizzi с соавт.: количества людей, находящихся под риском, использовались как коэффициенты в матрице C, в результате чего КП-модель формулировалась относительно рисков:

$$C = C_N = \begin{pmatrix} N_0 & \cdots & N_4 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & 0 & N_5 & \cdots & N_9 & 0 & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & N_{85} & \cdots & N_{99} \end{pmatrix},$$

   где N_i – количество людей, находящихся под риском, возраста i лет, а $\boldsymbol{\gamma}$, оцениваемые в модели, являются удельными рисками.

3. Подход Rizzi с дополнительным требованием постоянства удельных частот в интервале 87–99 лет ($\gamma_{87} = \gamma_{88} = \ldots = \gamma_{99}$). В модели это требование было

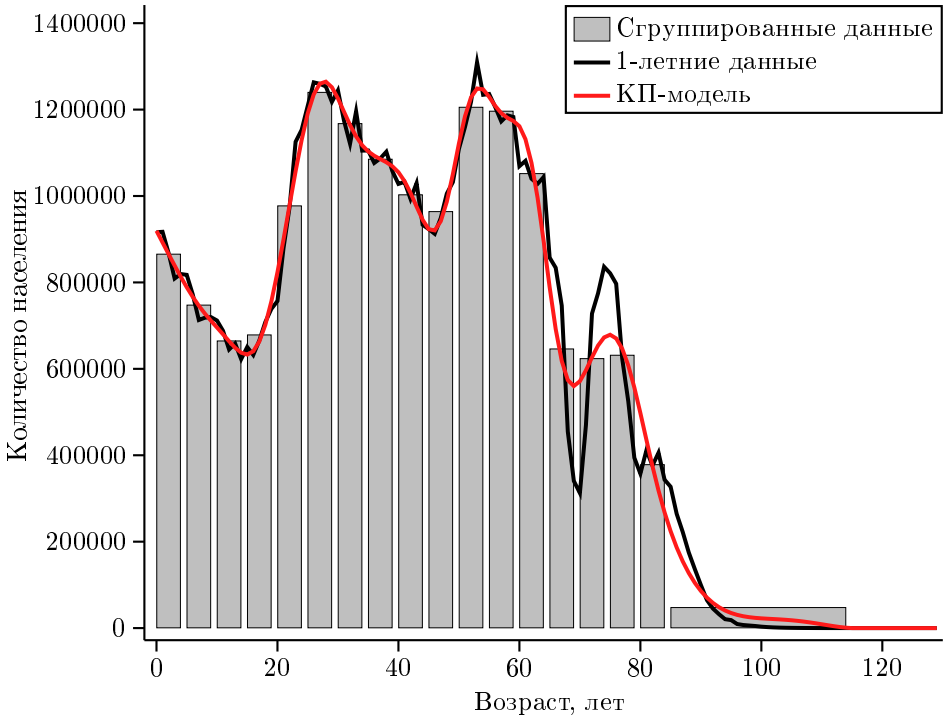

Рис. 7. Результаты восстановления возрастного распределения женского населения Российской Федерации в 2014 г. Использовался КП-метод с $D = D_2$.

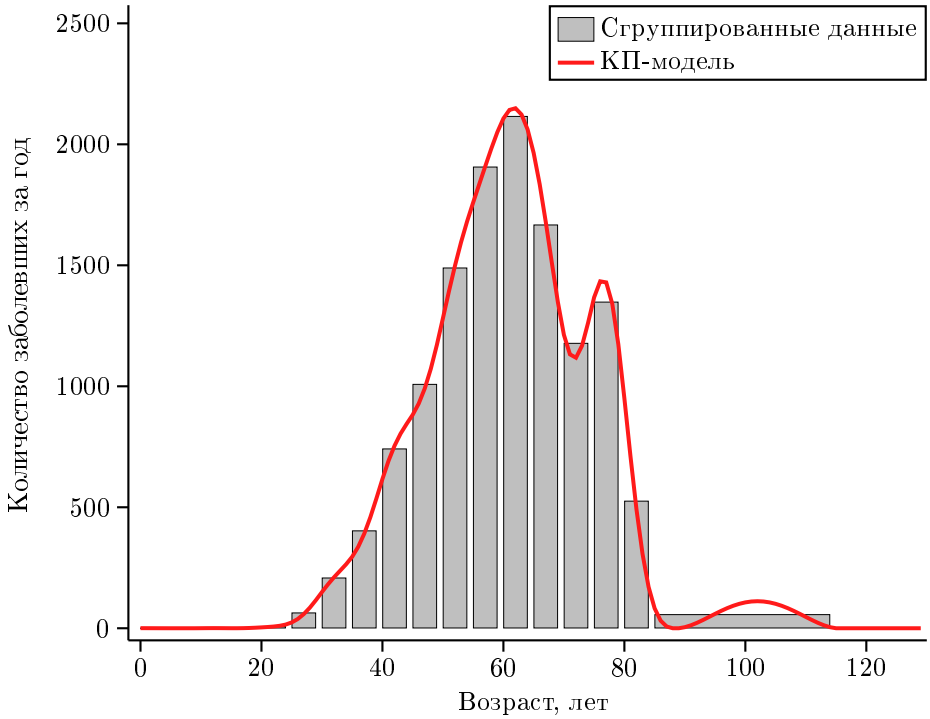

Рис. 8. Результаты восстановления повозрастного распределения количества заболевших злокачественными новообразованиями молочной железы в Российской Федерации в 2014 г. Использовался КП-метод с $D = D_2$.

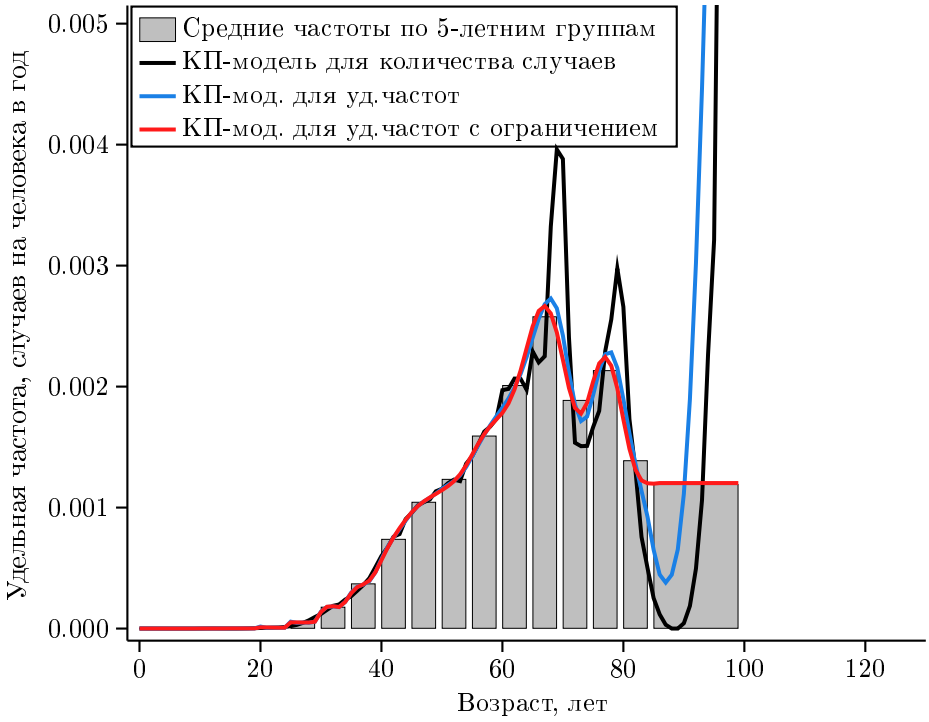

Рис. 9. Результаты восстановления повозрастного распределения удельных частот заболевания злокачественными новообразованиями молочной железы среди женского населения Российской Федерации в 2014 г. Использовался КП-метод с $D = D_2$. Высокие значения рисков в области 95–100 лет не показаны.

реализовано за счет расширения условия $C\gamma = y$:

$$\begin{pmatrix} N_0 & \cdots & N_4 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & 0 & N_5 & \cdots & N_9 & 0 & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & N_{85} & \cdots & N_{99} \\ 0 & \cdots & 0 & 1 & 0 & \cdots & \cdots & \cdots & 0 & -1 \\ 0 & \cdots & \cdots & 0 & 1 & 0 & \cdots & \cdots & 0 & -1 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & 1 & -1 \end{pmatrix} \gamma = \begin{bmatrix} y_1 \\ \vdots \\ \vdots \\ y_{18} \\ 0 \\ \vdots \\ \vdots \\ 0 \end{bmatrix}.$$

Результаты расчетов по данным ЗНО МЖ (рис. 8, 9) показывают, что критерий минимизации негладкости решения является недостаточным для корректного восстановления хвоста распределения в широкой возрастной группе без фиксированного значения на правой ее границе. Третий использованный подход иллюстрирует простейший способ регуляризации решения за счет внесения априорного требования его постоянства в определенной области.

По-видимому, полностью непараметрический подход не позволит решать проблемы такого рода. Это еще раз подчеркивает важность разработки комбинированных методов, сочетающих в себе общий непараметрический подход с добавлением априорных элементов, влияющих не на всё решение в целом, а на отдельные его участки или особенности.

ЗАКЛЮЧЕНИЕ

В данной работе был предложен непараметрический метод, опирающийся на задачу квадратичного программирования (КП-метод), позволяющий восстанавливать гладкие распределения по сгруппированным данным. Отличием КП-метода от метода, ранее созданного Rizzi с соавт. [1], являются алгоритмическая и вычислительная простота, точное сохранение групповых сумм и возможность легко вносить дополнительные ограничения вида равенств и линейных неравенств за счет модификации условий задачи квадратичного программирования. Точность работы КП-метода на рассмотренных наборах данных по смертности сопоставима с методом Rizzi с соавт., а при модификации КП-метода за счет использования сглаживающих матриц более высокого порядка точность КП-метода в ряде случаев превосходит точность метода Rizzi.

Восстановление распределений частот заболеваемости раком проиллюстрировало недостаточность чисто непараметрического подхода к задачам без априорного граничного условия и с неинерционным поведением решения у границ рабочей области.

Тем не менее, КП-метод можно рекомендовать для практического применения в биомедицинских исследованиях, работающих со сгруппированными данными, при условии экспертного контроля адекватности получаемых решений и внесения необходимых коррекций в модель. Также следует учитывать, что КП-метод дает сглаженное приближение истинного распределения, а не полностью восстанавливает истинное распределение. Но с ростом гладкости истинного распределения результаты КП-метода будут приближаться к нему.

Одно из возможных направлений развития как КП-метода, так и других схожих методов – это уточнение регуляризующей подмодели, т.е. требований, накладываемых на восстанавливаемое распределение (в данной работе это было требование минимизации нормы разностной производной γ по a). Для это необходимо изучение того, какие локальные статистические связи наблюдаются в различных классах реальных дискретных данных. Обнаруженные связи можно будет включать в модели, применяемые к сгруппированным данным того же класса.

ПРИЛОЖЕНИЕ 1.

Матрицы D_1, D_2, D_3, D_4

Сглаживающие матрицы D, соответствующих конечно-разностной аппроксимации первой (7), второй (8), третьей (9) и четвертой (10) производных:

$$D_1 = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & 0 & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & -1 \end{pmatrix} \tag{7}$$

$$D_2 = \begin{pmatrix} 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & \cdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & -2 & 1 \end{pmatrix} \tag{8}$$

$$D_3 = \begin{pmatrix} 1 & -3 & 3 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -3 & 3 & 1 & \cdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & -3 & 3 & 1 \end{pmatrix} \tag{9}$$

$$D_4 = \begin{pmatrix} 1 & -4 & 6 & -4 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -4 & 6 & -4 & 1 & \cdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & -4 & 6 & -4 & 1 \end{pmatrix} \tag{10}$$

Математическая биология и биоинформатика. 2016. Т. 11. № 2. doi: 10.17537/2016.11.367

СПИСОК ЛИТЕРАТУРЫ

1. Rizzi S., Gampe J., Eilers P.H.C. Efficient Estimation of Smooth Distributions From Coarsely Grouped Data. *American Journal of Epidemiology.* 2015. V. 182. № 2. P. 138–147.
2. National Center for Health Statistics, Centers for Disease Control and Prevention. Underlying cause of death 1999–2010 on CDC WONDER Online Database. URL: http://wonder.cdc.gov/ucd-icd10.html (дата обращения 06.11.2016).
3. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing, 2013.
4. Kuhn H.W., Tucker A.W. Nonlinear programming. In: *Proceedings of 2nd Berkeley Symposium.* Berkeley: University of California Press. 1951. P. 481–492.
5. Karush W. *Minima of Functions of Several Variables with Inequalities as Side Constraints.* Chicago: Illinois, 1939.
6. Nocedal J., Wright S.J. *Numerical Optimization (2nd ed.).* Berlin, New York: Springer-Verlag, 2006. ISBN 978-0-387-30303-1.
7. Gould N., Toint P.L. Preprocessing for quadratic programming. *Math. Programming, Series B.* 2004. V. 100. P. 95—132.
8. Единая межведомственная информационно-статистическая система (ЕМИСС). Численность постоянного населения - женщин по возрасту на 1 января. URL: https://fedstat.ru/indicator/33459 (дата обращения 11.11.2016)
9. *Злокачественные новообразования в России в 2014 году (заболеваемость и смертность).* Под ред. А.Д. Каприна, В.В. Старинского, Г.В. Петровой. М.: МНИОИ им. П.А. Герцена - филиал ФГБУ «НМИРЦ» Минздрава России, 2016.

Материал поступил в редакцию 22.11.2016, опубликован 08.12.2016.