О ВОССТАНОВЛЕНИИ ПЛОТНОСТИ ВЕРОЯТНОСТИ МЕТОДОМ ГИСТОГРАММ В ПОЧВОВЕДЕНИИ И ЭКОЛОГИИ

Глаголев М.В., Сабреков А.Ф.
m_glagolev@mail.ru

В работе исследуются различные методы решения задачи восстановления плотности вероятности распределения некоторого параметра. Показано, что наиболее часто используемый в почвоведении и экологии для решения этой задачи метод "гистограмм с равными интервалами группировки" является совершенно неприемлемым. Из числа простейших методов рекомендуется метод "гистограмм с интервалами равной вероятности".

Введение

Экспериментальные данные представляют собой реализации случайных величин или случайных процессов. Из теории вероятностей известно, что исчерпывающее описание случайной величины дается ее законом распределения вероятностей – правилом, позволяющим определять вероятность попадания этой величины в любую заданную область ее значений. Закон распределения случайной величины (распределение) может быть задан с помощью любой из двух взаимно однозначно связанных между собой функций: функции распределения и плотности вероятности [*Костылев с соавт.*, 1991: с. 56-57].

Введём необходимую для дальнейших рассуждений терминологию. Пусть в эксперименте исследуется случайная величина y, причем измеренные значения оказались равными $y_1, \ldots, y_n$; измерения независимы между собой. Подмножество из n элементов $y_1, \ldots, y_n$ из генеральной совокупности называется выборкой объема n. Для получения эмпирического распределения случайной величины y имеющийся набор значений $y_1, \ldots, y_n$ (выборку объема n) представляют в виде гистограмм: для непрерывной случайной величины y производится разбиение отрезка значений этой величины $[C_1\, C_{k+1}]$ на k интервалов $[C_1\, C_2]$, $[C_2\, C_3]$, …, $[C_k\, C_{k+1}]$; здесь C_1 равно наименьшему значению в выборке, а C_{k+1} – наибольшему, т.е.

$$C_1 = \min_{i=1,\ldots,n} (y_i), \quad C_{k+1} = \max_{i=1,\ldots,n} (y_i)$$

(отрезок $[C_1\, C_{k+1}]$ называется «интервалом выборки», а отрезки $[C_1\, C_2]$, …, $[C_k\, C_{k+1}]$ – «интервалами группировки» [*Костылев с соавт.*, 1991: с. 78]). Для простоты некоторых дальнейших обозначений мы будем считать, что выборка $y_1, \ldots, y_n$ представлена в виде вариационного ряда, т.е. упорядочена по возрастанию элементов.

Тогда

$$C_1 = y_1, \quad C_{k+1} = y_n.$$

Подсчитывается число элементов исходного массива $y_1, \ldots, y_n$, лежащих в каждом интервале группировки. Обозначим эти числа f_i ($i = 1, \ldots, k$). При этом предполагается, что каждый элемент (y_j), попавший в i-ый интервал, совпадает с «усредненным» значением интервала, то есть равняется этому значению:

$$\check{C}_i = (C_i + C_{i+1})/2.$$

Затем определяются статистические частоты p_i путем деления чисел f_i на объем выборки:

$$p_i = f_i/n.$$

Совокупность значений $\check{C}_i$ и соответствующих им частот p_i называют гистограммой [*Живописцев и Иванов*, 1993: с. 49-50] (иногда вместо p_i по оси ординат откладывают просто количества элементов f_i). Типичная гистограмма наблюдений ($\check{C}_i$, f_i) представлена на рис. 1.

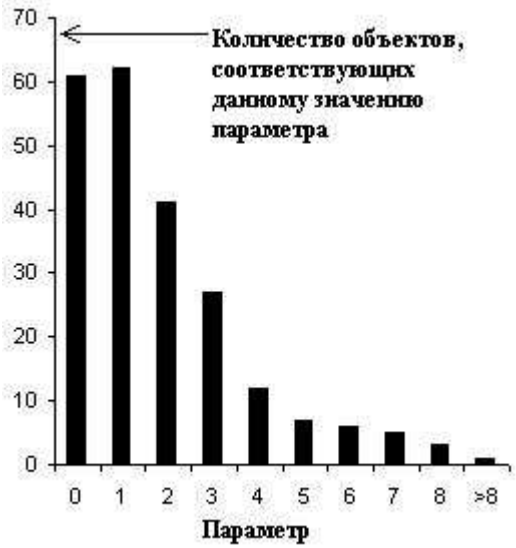

Рис. 1. *Типичная гистограмма наблюдений.*

При подготовке группированных данных большую значимость приобретает вопрос об обоснованном разбиении интервала выборки на интервалы группировки. В литературе имеется множество рекомендаций по выбору числа интервалов группировки (см., например, [*Костылев с соавт.*, 1991: с. 78-79] и ссылки там, а также [*Косарев*, 2003; 2008: с. 31-32]).

Однако при этом оказывается, что для одной и той же выборки разные методы выбора числа интервалов могут приводить к качественно разным распределениям.

В данной работе мы преследовали следующие цели:

1) проанализировать те особенности, которые имеет построение плотности распределения в почвоведении и экологии;

2) предложить метод восстановления (по экспериментальным данным) плотности распределения, лишенный неоднозначности (свойственной методам построения гистограмм), связанной с выбором интервалов группировки.

Теория

Классические способы разбиения выборки на равные интервалы группировки

Здесь мы упомянем лишь некоторые подходы, которые часто рекомендуются для практического использования [*Костылев с соавт.*, 1991: с. 78; *Живописцев и Иванов*, 1993: с. 50; *Косарев*, 2008: с. 32]):

1. Формула Старджеса: $k = 1 + \log_2(n)$.
2. Формула Брукса-Каррузера: $k = 5 \cdot \lg(n)$.
3. Формула Хайнхольда-Гаеде: $k = n^{½}$.
4. 1-ая формула Костылева с соавт.: $k = 4 \cdot \lg(n)$.
5. 2-ая формула Костылева с соавт.: $k = 5 \cdot \lg(n/10)$.
6. Формула Живописцева-Иванова: $k = 10 \cdot \ln(n)$.
7. Формула Косарева: $k = n^{1/3}$.

Заметим, что уже на этом этапе возникает неоднозначность. Даже не говоря о том, что разные формулы дают различные результаты, то, что по своему смыслу количество интервалов группировки должно быть целым (а это условие не обеспечивает ни одна из вышеприведенных формул), ставит нас перед выбором процедуры округления до целого. Как правильней провести эту процедуру?

Очевидно, что бóльшая однозначность достигается при округлении либо всегда в сторону меньшего целого числа, либо всегда в сторону большего. Однако при округлении в сторону большего целого мы будем получать неудовлетворительные результаты при малых k, поэтому остается рекомендовать округление в сторону меньшего целого. К сожалению, это не сильно улучшает ситуацию.

В табл. 1 для разных n приведены k, рассчитанные по различным формулам (с округлением в сторону меньшего целого). Заметим, что сейчас мы не рассматриваем формулу Живописцева-Иванова в силу ее кажущейся очевидной абсурдности: действительно, например, выборку,

содержащую лишь 3 элемента, согласно этой формуле следует разделить на… $k = 10 \cdot \ln(4) = 10.986 \approx 10 \div 11$ интервалов (однако ниже мы увидим, что в некоторых ситуациях именно эта формула дает лучшие результаты).

Таблица 1

Количество интервалов группировки (k)
для различных объемов выборок (n)

k	*Расчет по формуле*									
	Старджеса		*Брукса-Каррузера*		*Хайнхольда-Гаеде*		*Костылева с соавт.*			
							1-ой		*2-ой*	
	БО	*СО*	*БО*	*СО*	*БО*	*СО*	*БО*	*СО*	*БО*	*СО*
1	1	1	0	1	1	1	0	1	-5	1
2	2	1	1	1	1	1	1	1	-3	1
4	3	1	3	1	2	1	2	1	-1	1
8	4	2	4	2	2	2	3	2	0	2
16	5	3	6	3	4	3	4	3	1	3
32	6	5	7	5	5	5	6	5	2	3
64	7	5	9	5	8	5	7	5	4	3
128	8	7	10	7	11	7	8	7	5	5
256	9	9	12	11	16	11	9	9	7	7
512	10	9	13	13	22	15	10	9	8	7
1024	11	11	15	15	32	19	12	11	10	9

Примечание: БО – без ограничений; СО – с ограничением (подробности см. далее в тексте).

Из табл. 1 видно, что с увеличением объема выборки разрыв между оптимальными значениями k, рекомендуемыми различными формулами, увеличивается (заметим, что пока мы обсуждаем формулы «в чистом виде», без каких-либо ограничений, т.е. мы обсуждаем столбцы «БО» соответствующих формул в табл. 1). В частности, k может быть от 10 до 32 для $n \sim 10^3$. Впрочем, даже для выборок малых объемов вариации количества интервалов также существенны. Например, согласно 2-ой формуле Костылева с соавт. выборка объема $n = 16$ еще слишком мала, чтобы ее можно было представлять в виде гистограммы. Действительно, при $n = 16$ по 2-ой формуле Костылева с соавт. имеем $k = 1$, т.е. должен быть лишь один интервал группировки, и смысл построения гистограммы теряется. А по формуле Брукса-Каррузера получается, что гистограмму не только можно строить, но, более того, нужно рассортировать выборку по достаточно большому числу интервалов – по шести! Кстати, обратим внимание на поведение формул при малых объемах выборок. Достаточно очевидно, что довольно бессмысленно строить гистограмму если выборка состоит лишь из двух элементов, однако, например, формула Старджеса (которая, пожалуй, применяется на практике чаще всего), все-таки, рекомендует это делать, разбив интервал выборки на 2 интервала

группировки. Не лучше и абсурдная рекомендация разбить выборку из 4 элементов по трем интервалам группировки, даваемая формулой Брукса-Каррузера. Лишь 2-ая формула Костылева с соавт. не рекомендует строить гистограммы при малых объемах выборки, но она в этом «перегибает палку». Представляется достаточно очевидным, что хоть на два-то интервала группировки можно было бы разбить выборку, содержащую 16 элементов.

Указанные трудности были давно осознаны, в результате чего на практике стали использовать различные ограничения. Так, согласно [*Костылев с соавт.*, 1991: с. 79], часто полагают, что, во-первых,

$$0.55 \cdot n^{0.4} \le k \le 1.25 \cdot n^{0.4} \qquad (1)$$

и, во-вторых, k рекомендуют выбирать нечетным. Сразу заметим, что эти два правила иногда противоречат друг другу. Так, например, при $n = 8$ $0.55 \cdot n^{0.4} \approx 1.3$, а $1.25 \cdot n^{0.4} \approx 2.9$, т.е. первому правилу удовлетворяет единственное целое значение $k = 2$, но оно не удовлетворяет второму правилу. Такое же противоречие мы имеем и при $n = 5$, 6 или 7. Учитывая, что второе правило носит лишь рекомендательный характер, при возникновении подобного противоречия мы будем следовать только первому правилу. Заметим, что ограничения, предлагаемые различными авторами, очень сильно различаются. Так, *Ф.А. Живописцев и В.А. Иванов* [1993: с. 50] предлагают примитивное ограничение, не зависящее от объема выборки (ср. с №1):

$$5 \le k \le 30 \qquad (1a)$$

В столбцах «СО» табл. 1 приведено число интервалов группировки, рассчитанное по каждой формуле с учетом (№1) и требования нечетности. Из табл. видно, что, действительно, теперь достигнута гораздо большая однозначность, хотя и не полная. Например, при $n = 32 \div 64$, все формулы кроме последней дают оптимальное значение $k = 5$, а последняя формула рекомендует $k = 3$. На это можно было бы не обращать внимание, однако при относительно больших n разнобой возникает и между другими формулами (см. последние три строки в табл. 1). Причем для $n \sim 10^3$ оптимальные количества интервалов, рекомендуемые формулами Старджеса и Хайнхольда-Гаеде различаются почти в два раза (мы специально сравниваем здесь только эти две хорошо известные и часто применяемые формулы, если же обратить внимание на более экзотическую 2-ую формулу Костылева с соавт., то мы увидим, что различие может достигать более двух раз).

Пока мы занимались только простейшим классом формул разбиения интервала выборки на интервалы группировки: все рассмотренные выше формулы подразумевали, что k – это количество ***равных*** интервалов

группировки, т.е. если k было найдено по одной из вышеприведенных формул, то каждый интервал $[C_j\ C_{j+1}]$ строился таким образом, что

$$C_{j+1} = C_j + \Delta C,$$

где $\Delta C = (C_{k+1} - C_1)/k = \text{const}$. Не говоря уже о том, что существенная неоднозначность разбиения появится тогда, когда мы перейдем к методам ***неравных интервалов*** группировки, даже сейчас, все еще оставаясь в классе формул равных интервалов, мы вскроем дополнительный источник неоднозначности разбиения.

Способы разбиения, зависящие от типа распределения

Понятно, что представление в виде гистограммы (функции плотности вероятности на системе интервалов группировки) с формально-математической точки зрения является кусочно-постоянной аппроксимацией (т.е. аппроксимацией кусочными полиномами 0-го порядка) этой функции на сетке с узлами C_1, C_2, …, C_{k+1} (кстати, заметим, что почти столь же часто используемое, как и гистограмма, представление плотности вероятности в виде «полигона частот» – см., например, [*Дмитриев*, 1995: с. 45] – представляет собой кусочно-линейную аппроксимацию, т.е. аппроксимацию кусочными полиномами 1-го порядка). Из курса вычислительной математики (см., например, [*Петров и Лобанов*, 2006: с. 138]) известно, что на равномерной сетке, заданной на отрезке [a, b], абсолютная погрешность интерполяции функции $f(C)$ одной переменной C (под этой погрешностью мы будем понимать абсолютное значение остаточного члена и обозначать ее R_N) удовлетворяет оценке:

$$R_N \sim \Delta C^{N+1} \cdot \max_{C \in [a,\,b]} |f^{(N+1)}(C)|/(N+1), \qquad (2)$$

где N – степень аппроксимирующего полинома (т.е. $N = 0$ для гистограммы и 1 для полигона частот), $f^{(N+1)}(C)$ – производная $(N+1)$-го порядка от f. Пусть мы хотим обеспечить некоторую наперед заданную точность R_N. Как-либо изменять $f^{(N+1)}(C)$ мы не можем, поскольку это - производная реально существующего (хотя пока и не известного нам) искомого распределения. Единственное, за счет чего мы можем достигнуть надлежащей точности – это величина интервала ΔC которую следует выбрать тем меньше, чем больше $f^{(N+1)}(C)$, согласно оценке

$$\Delta C \sim [(N+1) \cdot R_N / \max_{C \in [a,\,b]} |f^{(N+1)}(C)|]^{1/(N+1)}, \qquad (2a)$$

Отсюда видно, что разные распределения могут потребовать разного количества интервалов группировки: чтобы наглядно представить такие

распределения, у которых функция плотности меняется очень резко (т.е. максимальное абсолютное значение производной велико), понадобится большое количество относительно коротких интервалов ΔC, а чтобы представить распределения с плавно изменяющейся функцией плотности, можно ограничиться небольшим количеством относительно длинных интервалов (в предельном случае максимально гладкого распределения, которым является равномерное распределение, достаточно одного единственного столбца, занимающего весь интервал выборки от C_1 до $C_{k+1} = C_2$). Итак, мы видим, что на самом деле ***количество интервалов группировки должно определяться не только имеющимся в нашем распоряжении объемом выборки, но и свойствами распределения***. При этом может оказаться, что мы имеем выборку гораздо меньшего объема, чем необходимо для восстановления данного распределения (тогда несмотря на выполнение рекомендаций табл. 1 мы его восстановить не сможем). Или, напротив, может оказаться, что наша выборка гораздо больше чем необходимо для надежного восстановления некоторого относительно гладкого распределения (тогда, даже если мы сильно нарушаем рекомендации табл. 1, распределение все равно может быть успешно восстановлено).

Однако формул, позволяющих рассчитать потребное количество интервалов группировки в зависимости от свойств распределения также известно несколько и они опять порождают неоднозначность. Приведем лишь два примера совершенно различных формул из [*Костылев с соавт.*, 1991: с. 78]:

$$k = (4/æ)\cdot \lg(n/10) \qquad (3a)$$

и

$$k = (\mathrm{E}\{y\} + 1.5)\cdot n^{0.4}/6, \qquad (3b)$$

где $æ = (\mathrm{E}\{y\})^{-½}$ - контрэксцесс распределения.

Трудность использования двух последних выражений состоит в том, что число интервалов группировки нужно выбрать до того, как будут определены числовые характеристики распределения, в том числе и эксцесс или контрэксцесс. Обычно полагают, что для эмпирических распределений

$$1.8 \le \mathrm{E}\{y\} \le 6, \qquad (4)$$

и получают соответствующие границы числа интервалов [*Костылев с соавт.*, 1991: с. 79]. Именно из формул (№3b) и (№4) были получены ограничения (№1). Очевидно, что если для получения ограничений использовать формулу (№3a) вместо (№3b), то и ограничения были бы совершенно другими, а тогда более или менее стройная система табл. 1 (которая дает однозначность хотя бы для не очень больших n) нарушилась

бы! Более того, ограничения эти (даже без всяких изменений - в том виде, в котором они заданы сейчас) выписаны в предположении справедливости условия (№4), а для распределений, у которых, например, E{y} = 70, рекомендации табл. 1 (по ограничению количества интервалов группировки) будут просто неправильными.

Но у способов разбиения на равные интервалы есть и более существенный недостаток. Формула (№2а) требует везде использовать малые интервалы ΔC, определяемые наибольшим значением производной на всем интервале выборки $[C_1\ C_{k+1}]$. Однако на практике мы имеем выборку ограниченного объема и в соответствие с табл. 1 не можем разбивать интервал выборки на слишком большое количество интервалов группировки. А ведь плотность распределения может на каком-то участке изменяться так быстро, что потребное количество интервалов группировки превысит имеющееся в нашем распоряжении количество измерений n.

Разбиения выборки на неравные интервалы группировки

Предположим, что мы хотим построить гистограмму (т.е. аппроксимировать функцию плотности распределения) не на всем интервале выборки, а лишь на его левой половине (положим для определенности, что здесь функция меняется достаточно резко). При этом по условию (№2а) будет получено некоторое значение ΔC_L (относительно малое, поскольку резкое изменение функции количественно выражается в том, что $\max|f'(C)|$ будет достаточно большой величиной).

Предположим теперь, что мы хотим построить гистограмму лишь на правой половине интервале выборки (положим для определенности, что здесь функция меняется очень плавно). При этом по условию (№2а) будет получено некоторое значение ΔC_R (относительно большое, поскольку плавное изменение функции количественно выражается в том, что $\max|f'(C)|$ будет достаточно малой величиной).

Но ничто не мешало нам рассмотреть не левую и правую части интервала выборки, а разделить его, скажем на три части. Или на четыре… Если довести эту идею до логического завершения, то мы приходим к очевидному требованию: длину каждого интервала группировки следует выбирать в соответствии с поведением функции плотности распределения - на участках быстрого изменения этой функции (т.е. на участках с большим абсолютным значением производной) следует использовать малые интервалы ΔC, а на участках медленного изменения функции (т.е. на участках с малым абсолютным значением производной) можно использовать большие интервалы ΔC. Иначе говоря, мы пришли к необходимости использования неравномерной системы интервалов. Внешне это будет выглядеть так, что гистограмма состоит из относительно узких столбиков на участках резкого убывания или возрастания функции распределения, а на участках плавного изменения ширина столбиков

увеличится. Строить систему неравномерных интервалов можно различными способами и среди них есть весьма простые.

Например, ***используют не интервалы равной длины, а интервалы с равной вероятностью***. Для оценки числа таких интервалов K часто используется формула Уильямса, которая при уровне значимости 0.1 имеет вид [*Костылев с соавт.*, 1991: с. 79]:

$$K = 1.9 \cdot n^{0.4}. \qquad (5)$$

(естественно, K каким-либо образом округляется до целого числа, например, до ближайшего меньшего целого).

Алгоритм построения гистограммы на неравных интервалах заключается в следующем. По вышеприведенной формуле определяется оптимальное количество таких интервалов. Выборка упорядочивается по возрастанию и определяется количество элементов (N) выборки, которые должны попасть в каждый интервал группировки:

$$N = n/K$$

(конечно, n может не делиться на K нацело, но мы сейчас для простоты будем считать, что получено именно целое значение N; если же деление нацело невозможно, то следует воспользоваться каким-либо простейшим чисто техническим приемом, например, в качестве N взять округленное до ближайшего целого значение n/K, и в каждый интервал группировки поместить N значений выборки, а в последний интервал - N_1 оставшихся значений выборки; также можно использовать и более равномерное распределение элементов по интервалам, например, следя за тем, чтобы количество элементов выборки в соседних интервалах группировки не отличалось более чем на 1).

Итак, мы имеем K интервалов группировки, в каждом из которых находится $f_i = N$ ($i = 1, \ldots, K$) элементов выборки. Пусть ΔC_i – ширина i-го интервала группировки. Значения ΔC_i выбираются именно таким образом, чтобы в каждом i-м интервале оказалось N элементов выборки. Для этого в упорядоченной по возрастанию выборке, отсчитывается N элементов и они помещаются в первый интервал, затем отсчитывается еще N элементов и они помещаются во второй интервал и т.д. При этом в первом интервале окажутся элементы от y_1 до y_N включительно, во втором интервале – от y_{N+1} до $y_{2 \cdot N}$ включительно, в третьем – от $y_{2 \cdot N+1}$ до $y_{3 \cdot N}$ и т.д. Тогда в качестве границ интервалов группировки можно принять

$$C_1 = y_1, \; C_2 = (y_N + y_{N+1})/2, \; C_3 = (y_{2 \cdot N} + y_{2 \cdot N+1})/2, \ldots, C_{K+1} = y_n. \qquad (6)$$

Отсюда

$$\Delta C_1 = C_2 - C_1, \;\; \Delta C_2 = C_3 - C_2, \ldots, \Delta C_i = C_{i+1} - C_i.$$

Поскольку площадь всей гистограммы должна быть равна полной вероятности, т.е. 1 (или 100%), то для определения высот столбиков гистограммы ($h_1, h_2, \ldots, h_K$) мы имеем систему уравнений:

$$K$$

$$\Delta C_1 \cdot h_1 \cdot \alpha = f_1, \quad \Delta C_2 \cdot h_2 \cdot \alpha = f_2, \ldots, \quad \Delta C_K \cdot h_K \cdot \alpha = f_K, \quad \text{где } \alpha = \sum_{i=1} f_i = n. \quad (№7)$$

Отсюда мы вычисляем высоты столбиков $h_1, h_2, \ldots, h_K$ и строим гистограмму.

К сожалению, разбиение выборки на неравные интервалы группировки также неоднозначно. В частности, кроме упомянутой выше формулы Уильямса, используется еще и формула Манна-Вальда [*Костылев с соавт.*, 1991: с. 79].

Главный вопрос:
А восстанавливается ли плотность распределения?

Однако, даже если мы, используя какие-то совершенные ограничения, сможем добиться однозначности выбора интервалов, это не гарантирует нам нахождения истинной плотности распределения. Сложность состоит в том, что правильность восстановления плотности распределения никак не связана с однозначностью количества интервалов.

При слишком малом числе интервалов k эмпирическая плотность вероятности (гистограмма) утратит детальность и связь с особенностями теоретического распределения станет малоинформативной. Если же число интервалов будет слишком большим, то некоторые из них окажутся пустыми или слабо заполненными, т.е. гистограмма окажется изрезанной [*Костылев с соавт.*, 1991: с. 78], иначе говоря, эмпирическая плотность вероятности обретет множество экстремумов, которых, возможно, не было в теоретическом распределении.

Но если истинная плотность распределения нам не известна, то каким же образом можно решить вопрос о том, восстанавливается ли она при том или ином разбиении интервала выборки? Сделать это можно (и весьма просто) при помощи метода математического моделирования: необходимо задаться некоторым известным распределением, сгенерировать на его основе выборку, а после этого, исходя из данной выборки, попытаться восстановить исходную плотность распределения.

Уточнение вопроса:
Какова погрешность плотности распределения?

Анализ конкретизированного в предыдущем разделе вопроса нельзя проводить, не сделав терминологическое уточнение: ***а что, вообще говоря, значит «восстановить плотность распределения»***? Рассмотрим пример. Пусть какая-то величина характеризуется нормальным распределением с параметрами $m = 0$ и $\sigma = 1$. Если мы, взяв некоторую выборку значений этой величины, установим, что распределение

действительно нормальное и его параметры действительно $m = 0$ и $\sigma = 1$, то следует признать, что мы можем в точности восстановить плотность распределения. Между тем, совершенно ясно, что, имея выборку ***конечного*** объема, мы сможем абсолютно точно вычислить m и σ лишь случайно, а в подавляющем большинстве случаев полученные значения m и σ будут слегка отличаться от 0 и 1, да и про нормальность распределения мы сможем говорить лишь с той или иной долей уверенности, но полной гарантии у нас никогда не будет. То есть, формально, истинная плотность распределения не может быть восстановлена. Она может быть восстановлена лишь с некоторой погрешностью.

Таким образом качественный вопрос «можем ли мы при помощи однозначного разбиения найти истинную плотность распределения» переходит в более глубокий количественный вопрос: ***с какой погрешностью мы можем найти истинную плотность распределения?*** Теперь становится ясно, что вскрытая выше проблема неоднозначности восстановления плотности распределения при использовании гистограмм, построенных по различным интервалам группировки, сводится именно к проблеме погрешности. Погрешность построения гистограммы весьма велика и в пределах этой погрешности равнозначны все возможные варианты гистограмм.

Оценим точность метода гистограмм. Пусть ΔC – ширина интервала гистограммы. Если ширина интервала много меньше интервала выборки, а полное количество наблюдений $n >> 1$, то количество наблюдений в данном интервале имеет пуассоновское распределение и относительная статистическая погрешность числа наблюдений в данном интервале (с центром C) равна[18]

$$\delta_s = [n \cdot p(C) \cdot D]^{-1/2},$$

где $p(x)$ – плотность вероятность. Кроме статистической погрешности существует еще погрешность аппроксимации [*Косарев*, 2008: с. 31]. Мы уже приводили формулу для абсолютной погрешности аппроксимации – см. выше формулу (№2). Для относительной погрешности аппроксимации (δ_a) в частном случае кусочно-постоянной аппроксимации (т.е. для $N = 0$):

$$\delta_a = R_N / p(C) = A \cdot \Delta C^{N+1} \cdot \max| p^{(N+1)}(C)| / (N+1) / p(C) = A \cdot \Delta C \cdot |p'(C)| / p(C),$$

здесь мы заменили знак пропорциональности (использованный в №2) на знак равенства, но, естественно, ввели константу пропорциональности A

[18] В [*Косарев*, 2008: с. 31] принимается, что $D = \Delta C$.

(согласно [*Косарев*, 2008: с. 31], $A = 0.5$), а также учли, что на небольшом интервале ΔC можно принять производную p' постоянной.

Таким образом, суммарная относительная погрешность (δ) равна:

$$\delta = (\delta_s^{\,2} + \delta_a^{\,2})^{1/2} = \{[n \cdot p(C) \cdot D]^{-1} + 0.25 \cdot [\Delta C \cdot p'(C)/p(C)]^2\}^{1/2}.$$

Оценка погрешности метода гистограмм была впервые получена Н.В. Смирновым в 1950 г. При оптимальном интервале группировки

$$D_{opt} = [2 \cdot p(C) \cdot n^{-1} \cdot p'(C)^{-2}]^{1/3}$$

достигается минимальная относительная погрешность

$$\delta_{min} = A_1 \cdot [n^{-1} \cdot p(C)^{-2} \cdot p'(C)]^{1/3},$$

где $A_1 = 3^{1/2} \cdot 4^{-1/3} \approx 1.1$ [*Косарев*, 2008: с. 31-32]. Абсолютная погрешность

$$\Delta \cdot p = \delta \cdot p(C) = \{p(C)/(n \cdot D) + 0.25 \cdot [\Delta C \cdot p'(C)]^2\}^{1/2} \qquad (8)$$

Вообще говоря, надо понимать, что полученная формула – это лишь грубая оценка погрешности. Поэтому неудивительно, если окажется, что $p - \Delta \cdot p < 0$. Конечно, отрицательная плотность вероятность невозможна, поэтому в таких случаях следует исправить $\Delta \cdot p$, приписав ей значение p.

Подчеркнем также, что полученные формулы дают только точность *гистограммы, т.е. кусочно-постоянной аппроксимации*. Для точности полигона частот (т.е. кусочно-линейной аппроксимации) формулы будут уже другие (поскольку тогда в №2 $N = 1$, следовательно $\delta_a \sim \Delta C^2$). В [*Косарев*, 2008: с. 32] указано, что для полигона частот

$$\delta_{min} \sim n^{-2/5}.$$

Таким образом, мы видим, что метод гистограмм не самый точный – убывание погрешности с ростом n происходит медленнее, чем для метода полигона частот [*Косарев*, 2008: с. 32].

В заключение этого раздела укажем, что существует и иной подход к вычислению погрешности области определения – построение доверительного интервала плотности распределения. Задача доверительного оценивания восстановленной плотности распределения $p_n(C)$ как случайной функции состоит в построении вокруг нее некоторой “трубки”, размер которой определяется вероятностными характеристиками нормированного отклонения $p_n(C)$ от истинной плотности распределения $p(C)$, объемом исходной статической выборки и заданным коэффициентом доверия β [*Лапко с соавт.*, 1996: с. 52]. Для простейшего случая гистограммы с равными интервалами группировки доверительная область

может быть построена при помощи теоремы Мании, как это изложено, например, в [*Лапко с соавт.*, 1996: с. 52-53].

Особенность задачи восстановления распределений в почвоведении

Прежде чем поставить описанный выше вычислительный эксперимент по восстановлению некоторой известной нам плотности распределения, вычленим особенность задачи восстановления, характерную для почвоведения и экологии. Обычно, когда в литературе рассматривается вопрос о построении эмпирической плотности распределения (в виде гистограммы или полигона частот) для некоторой выборки $y_1, \ldots, y_n$, то предполагается, что каждый ее элемент точно известен и может быть поэтому отнесен в какой-то конкретный интервал группировки (см., например, [*Ллойд и Ледерман*, 1989: с. 96-98; *Белов с соавт.*, 1990: с. 221-225; *Дмитриев*, 1995: с. 54, 55, 172, 173]). Но, с другой стороны, хорошо известно, что данные эксперимента всегда зашумлены некоторой погрешностью, причем на это указывают в том числе и те авторы, (см. [*Ллойд и Ледерман*, 1989: с. 10; *Белов с соавт.*, 1990: с. 18-23; *Дмитриев*, 1995: с. 7]), которые при построении гистограмм рассматривают лишь данные без погрешностей. Как видим, эти две взаимоисключающие точки зрения встречаются в одних и тех же литературных источниках. При общем введении в статистику постулируется, что экспериментальные данные всегда заданы с некоторой погрешностью, но при построении гистограммы плотности распределения этих экспериментальных данных об их погрешностях уже не вспоминают.

А может быть такой подход имеет право на существование? Действительно, зачем лишний раз вспоминать о погрешности данных, если мы собираемся восстановить полную характеристику случайной величины – ее закон распределения. Ведь «погрешность» - это одна из очень упрощенных характеристик, которыми можно пытаться приближенно охарактеризовать распределение…

С этим можно согласиться в том случае, когда распределение измеряемой случайной величины обусловлено только погрешностями измерений. Но при работе с такой сложной средой как почва гораздо чаще встречается иная ситуация. Поясним ее на примере.

Рассмотрим эмиссию (из почвы в атмосферу) метана – важного парникового газа. Во многих работах, посвященных математическому моделированию этого процесса (см., например, [*Cao et al.*, 1995; *Walter et al.*, 1996; *Glagolev*, 1998; *Глаголев*, 2006]) не без оснований принимается, что интенсивность эмиссии пропорциональна произведению нескольких множителей, каждый из которых отражает влияние одного независимого фактора (логические основания этого вскрыты в [*Глаголев с соавт.*, 2007: с. 203-204], а экспериментальные доказательства независимости действия гидротермических факторов на процесс разложения органики – одну из стадий метаногенеза, приведены в [*Van der Linden et al.*, 1987]). Случайные величины, формирующиеся под воздействием большого числа независимых случайных факторов,

взаимодействующих мультипликативно и имеющих один порядок значений, описываются логнормальным распределением [*Костылев с соавт*., 1991: с. 62]. Действительно, *Н.С. Паников* [1995: с. 20] опубликовал результаты измерения сотрудниками его лаборатории эмиссии метана на территории Бакчарского болота (Северо-Восточный отрог Большого Васюганского Болота, Бакчарский район Томской обл.), которые удовлетворяли именно логнормальному распределению (подчеркнем, что мы сейчас рассматриваем всего лишь один конкретный пример и вовсе не хотим сказать, что эмиссия удовлетворяет логнормальному распределению всегда, более того, в [*Шнырев и Глаголев*, 2007] приведены примеры иных распределений эмиссии). Следует отметить, что пока мы не вводили в рассмотрение какие-либо погрешности. Логнормальное распределение потока метана возникает не из-за неточности измерений, а из физической картины влияния факторов среды на образование, транспорт и окисления метана в почве. Таким образом, даже если бы мы могли произвести все измерения абсолютно точно, все равно мы бы не получали всегда какое-то одно значение величины эмиссии, но выборку различных значений $y_1, \ldots, y_n$, удовлетворяющих, например, логнормальному распределению. Однако не следует забывать, что мы не можем произвести анализ газа абсолютно точно. Поэтому каждый элемент выборки (y_i) будет получен с некоторой абсолютной ошибкой (Δy_i).

В почвоведении распределение ошибок анализов обычно неплохо аппроксимируется нормальным законом [*Дмитриев*, 1995: с. 73]. Вообще говоря, предположение о том, что закон распределения ошибок измерения близок к нормальному, принимается достаточно часто и в других предметных областях (см., например, [*Белов с соавт*., 1990: с. 20; *Живописцев и Иванов*, 1993: с. 8-9]). Таким образом, для почвоведения и экологии будет характерна задача построения эмпирической плотности распределения некоторой выборки $y_1, \ldots, y_n$, при том, что элементы выборки заданы не точно, а имеют погрешности (соответственно, $\Delta y_1, \ldots, \Delta y_n$), распределенные по некоторому закону, чаще всего – по закону Гаусса.

Вычислительные эксперименты и их обсуждение

Восстановление распределения потока CH_4

Итак, допустим, что для какого-то объекта (например, для конкретного болота) на самом деле эмиссия метана имеет логнормальное распределение с плотностью вероятности

$$W(y) = \begin{cases} (2\cdot\pi\cdot\sigma^2\cdot y^2)^{-½}\cdot \exp\{-0.5\cdot[\ln(y) - \ln(m)]^2/\sigma^2\} & \text{при } y \geq 0 \\ 0 \text{ при } y < 0 \end{cases}$$

с параметрами $m = 5$ и $\sigma = 1.9$. Задача определения статистических характеристик распределения потока метана для различных объектов в разных природных зонах Западной Сибири впервые решалась в [*Шнырев и Глаголев*, 2007]. На основании этой работы можно принять, что средний объем выборки для одного объекта составляет порядка 50 значений. Создадим выборку такого объема при помощи генератора случайных чисел (мы использовали функцию `lognrnd` из «Statistics Toolbox» системы MATLAB 7; результат работы данного генератора приведен в табл. 2, где мы уже упорядочили полученную выборку по возрастанию элементов).

Таблица 2

Выборка, полученная при помощи генератора логнормального распределения: `lognrnd`(1.6,1.9,50,1)

i	y_i		*i*	y_i		*i*	y_i		*i*	y_i		*i*	y_i
1	0.08		11	1.03		21	4.47		31	8.75		41	25.52
2	0.21		12	1.08		22	4.65		32	9.31		42	37.95
3	0.24		13	1.34		23	4.82		33	13.65		43	47.89
4	0.24		14	1.63		24	5.60		34	14.80		44	48.04
5	0.32		15	2.20		25	6.21		35	17.81		45	48.05
6	0.39		16	2.34		26	6.34		36	18.55		46	54.16
7	0.51		17	3.51		27	6.97		37	19.34		47	58.03
8	0.57		18	3.71		28	7.58		38	19.43		48	73.56
9	0.67		19	3.86		29	8.15		39	19.85		49	109.31
10	0.87		20	4.17		30	8.64		40	23.55		50	316.54

Теперь на каждое измерение нам нужно наложить «погрешности измерений», распределенные в соответствие с нормальным законом – именно результирующая выборка будет представлять собой модель реальных измерений потока метана в природе. Но пока мы этого делать не будем, а зададимся вопросом – насколько точно смогут восстановить исходную плотность распределения (если даже элементы выборки не «зашумлены» никакими дополнительными погрешностями) стандартные методы построения гистограмм. Правда, для большей наглядности (чтобы несоответствие между теоретическим исходным и восстановленным распределениями не «скрадывалось» за счет ширины столбиков гистограммы) мы будем использовать графическое представление не в виде гистограммы, а в виде сглаженного полигона частот, однако подчеркнем, что это – только способ графического представления данных, количество же интервалов группировки, их расположение и ширина совершенно одинаковы что при построении гистограммы (кусочно-постоянной аппроксимации плотности распределения), что при

построении обычного полигона частот (кусочно-линейной аппроксимации), что при использовании сглаженного полигона частот (аппроксимации кусочными полиномами порядка выше первого, в частности – при широко используемой в настоящее время аппроксимации сплайнами).

Прежде всего сосредоточимся на формуле Старджеса, как наиболее широко применяемой в практической деятельности. Сама по себе эта формула для $n = 50$ дает $k = 1 + \log_2(50) \approx 6.6$ интервала. Поскольку нецелое количество интервалов невозможно, то рассмотрим несколько вариантов: округление до ближайшего целого, в данном случае совпадающее с округлением до ближайшего большего целого ($k_R = 7$), округление до ближайшего меньшего целого ($k_L = 6$). Оба эти значения удовлетворяют ограничениям Живописцева-Иванова (№1а). Но они не удовлетворяют верхнему ограничению Костылева с соавт. (№1), действительно $1.25 \cdot 50^{0.4} \approx 5.98$. Следовательно, нам придется проверить еще $k_{LO} = 5$ интервалов – это количество интервалов группировки удовлетворяет всем ограничениям, в том числе и «рекомендации нечетности». Также обратим внимание на формулу Живописцева-Иванова с обязательным ограничением (№1а), поскольку среди всех формул она дает максимальное количество интервалов – 30.

Из табл. 1 видно, что $y_1 = 0.08$, $y_{50} = 316.54$. Следовательно при $k_R = 7$ ширина интервала группировки будет составлять $\Delta C = 45.2086$. При $k_L = 6$ $\Delta C = 52.7433$, а при $k_{LO} = 5$ $\Delta C = 63.2920$. Однако ни одно из этих значений не является удовлетворительным – из рис. 2 видно, что эмпирические плотности распределений, построенные с использованием указанных интервалов группировки не имеют никакого отношения к исходному теоретическому распределению. На первый взгляд удивительно то, что выделение интервалов группировки по формуле Живописцева-Иванова дает наилучший результат. Однако на самом деле ничего удивительного здесь нет. Мы специально выбрали очень «неудобное» распределение с огромным эксцессом (около двух миллионов!!!). Для распределения с таким эксцессом ограничение (№1), выведенное в предположении того, что эксцесс не превышает 6 (!!!), безусловно, не действует. Например, формула (№3а), учитывающая величину эксцесса, рекомендует разбить интервал выборки почти на 4000 равномерных интервалов группировки! Конечно, по сравнению с этим огромным числом кажутся одинаково неправильными разбиения что на 5-7 (по формуле Старджеса), что на 30 интервалов (по формуле Живописцева-Иванова), но, все-таки, 30 интервалов ближе к 4000, чем 5 интервалов!

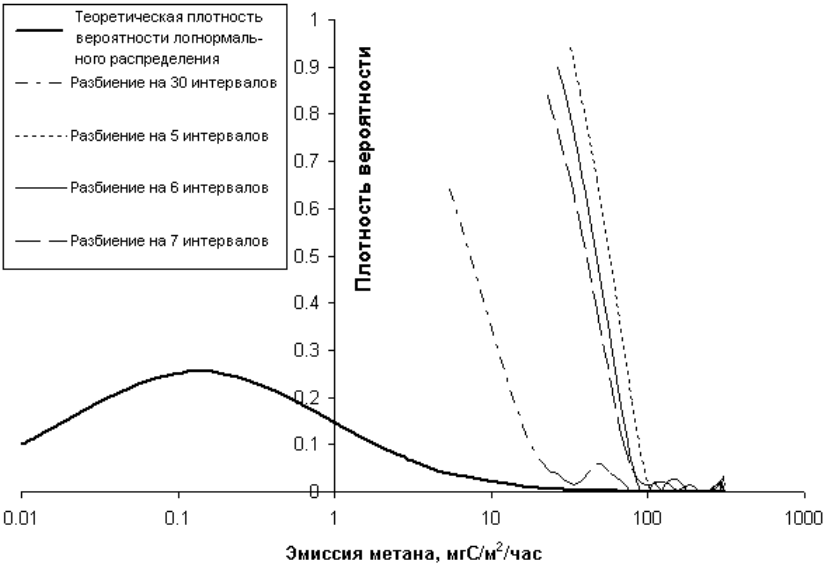

Рис. 2. *Восстановление плотности распределения при помощи интервалов группировки равной длины.*

Неужели действительно необходимо разбивать выборку на такое огромное количество интервалов группировки? Да это и не будет иметь какого-либо смысла для объема выборки $n = 50$! Сейчас мы покажем, что такой элементарный прием, как использование интервалов неравной длины (но равной вероятности) намного улучшает ситуацию даже при необычайно высоком эксцессе.

По формуле (№5) находим, что в случае интервалов равной вероятности их необходимо лишь $K = 1.9 \cdot 50^{0.4} \approx 9.1$. Поскольку дробное количество интервалов невозможно, то опять рассмотрим два варианта: округление до ближайшего целого, в данном случае совпадающее с округлением до ближайшего меньшего целого ($K_L = 9$) и округление до ближайшего большего целого ($K_R = 10$). Последний вариант очень прост – интервалы должны быть таковы, чтобы в каждый из них попало по 5 подряд идущих выборочных значений. Из табл. 2 по формуле (№6) получаем:

[0.08, 0.355); [0.355, 0.975); [0.975, 2.27); [2.27, 4.32); [4.32, 6.275); [6.275, 8.695); [8.695, 18.18); [18.18, 24.535); [24.535, 51.105); [51.105, 316.54]

Тогда

$\Delta C_1 = 0.275$, $\Delta C_2 = 0.620$, $\Delta C_3 = 1.295$, $\Delta C_4 = 2.050$, $\Delta C_5 = 1.955$, $\Delta C_6 = 2.420$, $\Delta C_7 = 9.485$, $\Delta C_8 = 6.355$, $\Delta C_9 = 26.570$, $\Delta C_{10} = 265.435$.

Учитывая, что для нашего примера f_i/α = 5/50 = 0.1, получаем по формуле (7):

$h_1 = f_1/(\alpha \cdot \Delta C_1) = 0.1/0.275 \approx 0.364$, $h_2 \approx 0.161$, $h_3 \approx 0.077$, $h_4 \approx 0.049$,
$h_5 \approx 0.051$, $h_6 \approx 0.041$, $h_7 \approx 0.011$, $h_8 \approx 0.016$, $h_9 \approx 0.004$, $h_{10} \approx 0.000$.

Для $K_L = 9$ уже не удастся выбрать такие интервалы, чтобы в каждый из них попало равное количество элементов выборки. Как же быть в этом случае? Тут нет однозначного решения. Понятно, что разница между количествами элементов, попавших в соседние интервалы должна быть минимальна, т.е. эти количества могут различаться на 1. Если в 4 интервала попадет по 5 элементов в каждый, а в 5 интервалов попадет по 6 элементов в каждый, то нам как раз удастся разместить все 50 элементов выборки по 9 интервалам группировки. Но как должны быть расположены «пятиричные» и «шестиричные» интервалы один относительно другого? Чередоваться? Или сначала подряд должны идти интервалы одного вида, а потом другого? Здесь опять появляется источник неоднозначности. Впрочем, на практике оказывается, что эта неоднозначность не порождает больших расхождений в результате – независимо от расположения интервалов восстановленная плотность вероятности имеет примерно одинаковый вид.

Представляется разумным интервалы с большим количеством элементов размещать на длинных «хвостах» распределений. Раз элементов больше, то, следовательно, больше степень сглаживания (усреднения). Поскольку обычно на «хвостах» значения плотности вероятности достаточно малы, на фоне этих малых значений проявляются «паразитные» высокочастотные колебания, которых в исходном теоретическом распределении не было. Большая степень сглаживания позволяет избавиться от таких осцилляций.

Иногда поступают совершенно формально. Если, например, получается, что в каждый интервал группировки должно попасть по 5.5 элемента (что, казалось бы, лишено смысла), то отсчитывают 5 элементов, помещают их в один класс, а 6-ой элемент помещают как в этот класс, так и в следующий. При этом граница между классами проходит именно по этому 6-му элементу.

На рис. 3 приведены плотности распределения, восстановленные с использованием 9 и 10 неравномерных интервалов. Прежде всего, следует отметить, что теоретическое распределение восстановлено, вообще говоря, неплохо (обращаем внимание читателя, что по оси абсцисс выбран логарифмический масштаб, поэтому кажущееся несоответствие между исходной теоретической и восстановленной по «экспериментальным данным» плотностями распределения в левой части рисунка на самом деле имеет место для небольшого интервала очень малых эмиссий метана. Надо

понимать, что для ограниченной выборки достигнуть полного соответствия теоретическому распределению принципиально невозможно. В той конкретной выборке, которую мы имеем, минимальное значение оказалось равным 0.08 (см. табл. 2). Окажись оно, скажем, 0.04, соответствие было бы лучше.

Во-вторых, отметим, что, как и следовало из общетеоретических соображений, при использовании меньшего количества интервалов удалось избежать осцилляций на правом «хвосте» функции. К сожалению, масштаб рис. №3 не позволяет это увидеть (кривые, соответствующие 9 и 10 интервалам, практически сливаются), поэтому мы использовали врезку иного масштаба. Конечно, непосредственно из рис. №3 может показаться, что это не очень существенно. Однако не будем забывать о логарифмическом масштабе!

В-третьих, мы видим, что абсолютная погрешность, рассчитанная по формуле №8, действительно очень реалистично отражает возможное отклонение восстановленной нами плотности распределения от истинной. Естественно, рассчитывая абсолютную погрешность, мы не использовали какой-либо информации об истинном распределении, а оценивали значения $p(C)$ и $p'(C)$ по восстановленной плотности.

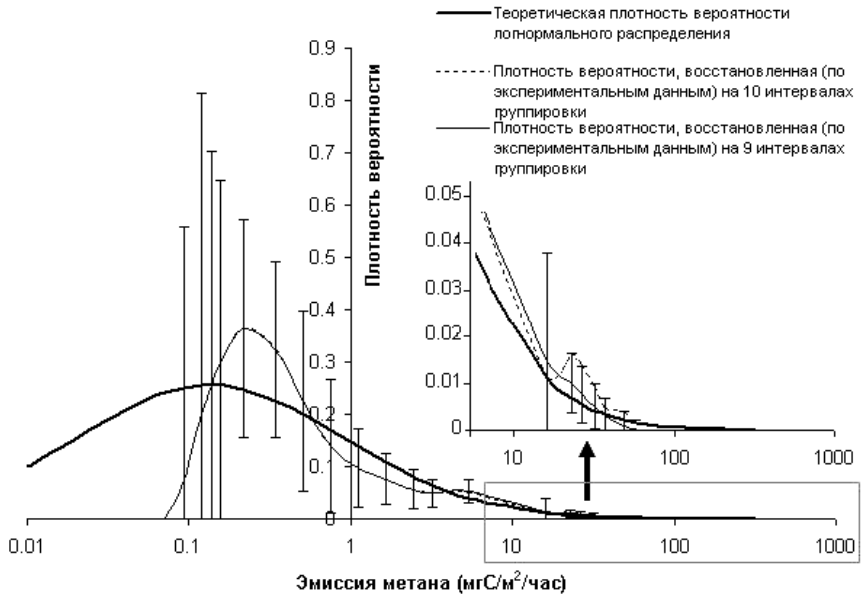

Рис. 3. *Восстановление плотности распределения при помощи интервалов группировки равной вероятности (но неравной длины).*

Завершая данный раздел, заметим, что ***использование интервалов неравной длины позволило достаточно точно восстановить по «экспериментальным данным» плотность распределения, имеющего чрезвычайно высокий эксцесс***. Причем сделано это было, так сказать, совершенно «автоматически» - нам не понадобилась какая-либо информация о свойствах восстанавливаемого распределения, в частности, о величине эксцесса. Обычные методы, основанные на использовании интервалов равной длины оказались совершенно неприменимы – построенные с их помощью распределения чрезвычайно сильно отличались от исходного логнормального распределения. Поскольку априори свойства восстанавливаемого распределения нам не известны, то ***пользоваться равными интервалами группировки никогда не следует***.

Некорректность задачи восстановления плотности распределения

С математической точки зрения постановка любой задачи включает в себя задание множества допустимых входных данных Φ и множества возможных решений Π. Цель вычислительной задачи состоит в нахождении решения $p \in \Pi$ по заданным входным данным $F \in \Phi$. Вычислительная задача называется корректной (по Адамару-Петровскому), если выполнены следующие требования (в том случае когда хотя бы одно из нижеперечисленных требований не выполнено, задача называется *некорректной*) [*Амосов с соавт*., 1994: с. 44]:

1) решение задачи существует при любых входных данных $F \in \Pi$;

2) это решение единственно;

3) решение устойчиво по отношению к малым возмущениям входных данных (отсутствие устойчивости означает, что малым погрешностям входных величин F могут соответствовать сколь угодно большие возмущения идентифицируемого элемента p [*Мацевитый и Лушпенко*, 1990]).

Согласно известному определению,

$$p(C) = dF(C)/dC,$$

где $F(z)$ – функция распределения [*Лапко с соавт*., 2000: с. 32]. Если вычислять плотность распределения согласно этому определению, то необходимо будет дифференцировать эмпирическую функцию распределения. Существует и другой подход к задаче. Интегрируя вышеприведенное определение можно получить интегральное уравнение, которому должны удовлетворять функции распределения и плотности распределения.

Таким образом, функция $p(C)$ является решением интегрального уравнения Фредгольма 1-го рода

$$\int \theta(z-C)\cdot p(C)\mathrm{d}C = F(z),$$

где $\theta(z\text{-}C)$ – функция Хевисайда [*Карандеев и Эйсымонт*, 1998].

Некорректность задач численного дифференцирования и решения интегральных уравнений 1-го рода хорошо известна и подробно разъяснялась в литературе (см., например, [*Калиткин*, 1978: с. 82, 462-464; *Тихонов и Арсенин*, 1979: с. 9, 11-13, 18]). Для решения некорректных задач применяют специальные методы, использующие так называемую регуляризацию решения. В следующем разделе мы перечислим целый ряд методов восстановления плотности распределения, каждый из которых основан на том или ином способе регуляризации.

Какие еще существуют методы восстановления плотности распределения?

Нельзя сказать, что задача восстановления плотности распределения относится к числу совсем неизученных, однако по каким-то причинам почти все, достигнутое в этой области, осталось вне поля зрения большинства почвоведов и экологов, сталкивающихся с этой задачей в своей профессиональной деятельности. Не претендуя на полное освещение методов решения указанной задачи, мы, тем не менее, перечислим ряд из них и приведем ссылки, которые, как мы надеемся, позволят заинтересованному читателю подробно познакомиться с каждым методом (рассмотренный выше метод гистограмм мы вторично упоминать не будем):

1. Дескриптивное приближение сплайнами [*Воскобойников с соавт.*, 1984];
2. Интегральная оценка [*Лапко с соавт.*, 1996: с. 25-29].
3. Метод Карандеева-Эйсымонта [*Карандеев и Эйсымонт*, 1998];
4. Метод корневой оценки [*Крянцев и Лукин*, 2006: с. 71-73];
5. Метод Розенблатта-Парзена [*Лапко с соавт.*, 1996: с. 23-25; *Крянцев и Лукин*, 2006: с. 63-65; *Лагутин*, 2007: с. 388-392], называемый также методом Е. Парзена (1962 г.) или ядерным методом [*Косарев*, 2008: с. 32-33];
6. Метод стохастической регуляризации [*Лапко с соавт.*, 1996: с. 34-36];
7. Метод структурной минимизации риска [*Вапник с соавт.*, 1984: с. 688-706];
8. Оценки максимума правдоподобия (в частности, метод решета) [*Лапко с соавт.*, 1996: с. 31-32];
9. Проекционные методы [*Крянцев и Лукин*, 2006: с. 66-69], в частности метод Н.Н. Ченцова (1962 г.) [*Косарев*, 2008: с. 32-33];
10. Регуляризованный метод гистограмм [*Крянцев и Лукин*, 2006: с. 69-71];
11. Рекуррентные ядерные оценки [*Лапко с соавт.*, 1996: с. 30-31];
12. Ядерная оценка с переменным параметром сглаживания [*Лапко с соавт.*, 1996: с. 29-30];

13. Ядерные оценки с пониженным смещением [*Лапко с соавт*., 1996: с. 32-34].

В [*Лапко с соавт.*, 1996: с. 29] проводилось сравнение методов Розенблатта-Парзена и интегральной оценки при решении двух тестовых задач восстановления плотности на основе выборок объемом n = 50, 100, 200: (а) нормального распределения, (б) распределения $p_Э(x)=0.5/[\pi\cdot(x^2+1)]$. Оказалось, что более точное приближение дает интегральная оценка.

В [*Карандеев и Эйсымонт*, 1998] проводилось сравнение методов Розенблатта-Парзена (с экспоненциальным ядром) и Карандеева-Эйсымонта при решении тестовых задач восстановления плотности: (а) смеси трех нормальных распределений на основе выборки объемом n = 40, (б) распределения Коши по выборке объемом n = 20 и (в) гамма-распределения. Оказалось, что для выборок малого объема более точное приближение дает метод Карандеева-Эйсымонта. Кроме того, для этого метода оценка является более устойчивой относительно выбора константы регуляризации.

При оптимальном выборе ширины σ гауссова ядра погрешность метода Парзена в зависимости от числа точек $\delta_{min} \sim n^{-2/5}$, т.е. такая же как и в методе полигона частот [*Косарев*, 2008: с. 32].

Описание «метода структурной минимизации риска» в [*Вапник с соавт*., 1984] снабжены текстом программы DENSIT на языке FORTRAN и отладочным примером с подробной распечаткой результатов.

В 1978 г. было показано, что ***не существует никаких других методов оценки плотности, имеющих большую скорость убывания (по n) погрешности оценки плотности, чем метод Ченцова*** (в этом методе $\delta_{min} \sim M/n^{½}$, где M – число используемых базисных функций) [*Косарев*, 2008: с. 33].

Преимущество рекуррентных процедур оценивания плотности вероятности заключается в относительной простоте настройки их параметров при получении новых данных. [*Лапко с соавт*., 1996: с. 30].

Исследование ядерных оценок с переменным параметром сглаживания, представляющих собой сложную сумму зависимых случайных величин, к сожалению, находится пока еще на начальном этапе [*Лапко с соавт.*, 1996: с. 29-30].

В заключение это раздела заметим, что вышеприведенный список (непараметрических методов восстановления плотности распределения) достаточно условен – он не должен рассматриваться как строгая классификация. Например, наиболее популярная непараметрическая оценка – метод Розенблатта-Парзена, которую мы, следуя традиции, поддерживаемой практически всеми исследователями (см., например, [*Лапко с соавт.*, 1996: с. 23-25; *Крянцев и Лукин*, 2006: с. 63-65; *Лагутин*, 2007: с. 388-392]), выделяем отдельным пунктом, может рассматриваться как частный случай других методов. Например, в [*Лапко с соавт.*, 1996: с. 26] показано, что при некоторых конкретных значениях параметров

интегральная оценка становится тождественной оценке Парзена. Совершенно очевидно также, что оценка Розенблатта-Парзена (получаемая, как известно, при постоянном значении параметра сглаживания) – это частный случай более общей ядерной оценки с переменным параметром сглаживания. Абсолютно то же самое можно сказать и про взаимоотношение оценки Розенблатта-Парзена с рекуррентными ядерными оценками. Наконец, в [*Лапко с соавт.*, 1996: с. 35] показано, как получить одну конкретную оценку Розенблатта-Парзена методом стохастической регуляризации. Вообще для многих конкретных реализации того или иного метода можно обнаружить их принадлежность одновременно к одной и другой группе непараметрических методов из вышеприведенного списка. Так, сравнивая статистику Бримэна, принадлежащую к классу ядерных оценок с переменным параметром сглаживания, и оценку Валветэна-Вагнера, относящуюся к классу рекуррентных ядерных оценок (см., например, [*Лапко с соавт.*, 1996: с. 29-30]), видим их формальную тождественность.

Некоторое программное обеспечение

Из перечисленных выше методов непараметрического восстановления плотности распределения на сегодняшний день, пожалуй, наиболее популярным является ядерный метод (Розенблатта-Парзена). В частности, именно этот метод реализован в широко распространенной системе MATLAB, где ему соответствует функция `ksdensity`. Но хороши ли результаты, получаемые с помощью данного метода?

Прежде всего заметим, что источниками неоднозначности при использовании функции `ksdensity` являются:

1) тип ядра (иными словами – вид «окна сглаживания»);
2) ширина «окна сглаживания»;
3) границы отрезка, на котором восстанавливается плотность распределения.

Однако при практическом оценивании плотности по заданной реализации выборки важен не столько вид окна, сколько правильное определение ширины «окна сглаживания» [*Лагутин*, 2007: с. 392].

Если границы отрезка нельзя указать из каких-то теоретических предпосылок, то можно использовать некоторое эмпирическое правило, например, формулу Стефанюка.

Эта формула, в частности, используется в программе DENSIT (реализующей «метод структурной минимизации риска»):

$$v = 5 \cdot (y_{\max} - y_{\min})/(n - 1), \quad A = y_{\min} - v, \quad B = y_{\max} + v,$$

где $y_{\max}$, $y_{\min}$ – соответственно, максимальное и минимальное значения элементов выборки; A, B - соответственно, координаты левой и правой

границ интервала восстановления плотности вероятности [*Вапник с соавт*., 1984: с. 692].

В [*Лагутин*, 2007: с. 390] показано, что главная часть квадратичного риска оценки плотности[19] определяется формулой

$$J(C) = 0.25\cdot\beta^2\cdot p''(C)\cdot h_n^4 + \alpha\cdot p(C)/(n\cdot h_n), \qquad (9)$$

где h_n – ширина «окна сглаживания»; α, β могут быть выражены через ядро q:

$$\alpha = \int q^2(y)dy, \quad \beta = \int y^2\cdot q(y)dy.$$

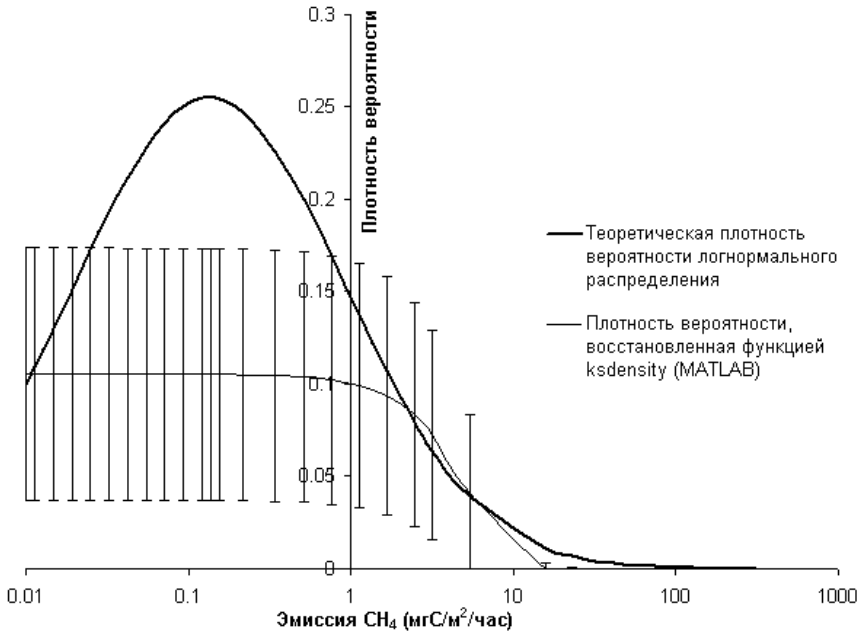

Рис. 4. *Восстановление плотности методом Розенблатта-Парзена.*

Для ядра Гаусса

$$q(y) = (2\cdot\pi)^{-½}\cdot exp(-y^2/2)$$

имеем [*Бронштейн и Семендяев*, 1980: с. 122]:

$$\alpha = \int_{-\infty}^{+\infty} q^2(y)dy = (2\cdot\pi)^{-1}\cdot\int_{-\infty}^{+\infty} exp(-y^2)dy = 2\cdot(2\cdot\pi)^{-1}\cdot\int_{0}^{+\infty} exp(-y^2)dy = \pi^{-1}\cdot\pi^{½}/2 \approx 0.2821.$$

[19] являющаяся аналогом квадрата абсолютной погрешности (№8), аналогом же самой абсолютной погрешности будет, естественно, $J(C)^{½}$.

$$\beta = \int_{-\infty}^{+\infty} y^2 \cdot q(y)dy = 2 \cdot \pi^{-1/2} \cdot \int_{-\infty}^{+\infty} (y/2^{1/2})^2 \cdot exp[-(y/2^{1/2})^2]d(y/2^{1/2}) = 4 \cdot \pi^{-1/2} \cdot \int_{0}^{+\infty} x^2 \cdot exp(-x^2)dx = 1.$$

Следовательно, можно так подобрать ширину «окна сглаживания» (h_n), чтобы минимизировать $\int J(C)^{1/2} dC$, т.е. минимизировать площадь полосы доверительного интервала вокруг восстанавливаемой плотности распределения. Естественно, при этом, изменяя h_n, необходимо следить (по какому-либо статистическому критерию, например, по критерию Колмогорова-Смирнова), чтобы получающаяся плотность распределения соответствовала исходной выборке. Впрочем, функция `ksdensity` может автоматически определять величину h_n. Но восстанавливается ли при этом истинная плотность распределения?

Вновь проведем вычислительный эксперимент. На рис. №4 приведены результаты восстановления плотности распределения по выборке табл. 2 MATLAB-функцией `ksdensity`. Сделаем несколько замечаний к этому рисунку.

1. Поскольку границы отрезка, на котором должна быть восстановлена плотность распределения, нельзя указать из каких-то теоретических предпосылок, то использовалась формула Стефанюка. Ширина «окна сглаживания» выбиралась функцией `ksdensity` в автоматическом режиме. Веса элементов выборки задавались обратно пропорциональными квадратам их абсолютных значений (если измерения не являются равноточными, то в статистике обычно вес индивидуального измерения принимается обратно пропорциональным дисперсии этого измерения, т.е. обратно пропорциональным квадрату средней квадратической ошибки данного измерения [*Румшиский*, 1971: с. 26-27]; в реальности оказывается, что потоки метана из почвы удается измерить с некоторой более или менее постоянной *относительной* ошибкой; таким образом, абсолютная ошибка или пропорциональная ей средняя квадратическая ошибка оказываются пропорциональны абсолютному значению величины потока, следовательно, дисперсия, равная квадрату средней квадратической ошибки, будет пропорциональна квадрату величины потока, а вес, обратно пропорциональный дисперсии, будет обратно пропорционален квадрату величины потока).
2. На первый взгляд кажется, что плотность распределения восстановлена не очень хорошо, однако обратим внимание на логарифмический масштаб по оси абсцисс. «Не очень хорошо» она восстановилась лишь в небольшом диапазоне значений потока метана от 0.01 до 1.
3. Отложенная на рис. №4 (в виде вертикальных «усов») оценка доверительного интервала при величинах потока порядка сотых-десятых не «дотягивается» до истинной функции плотности распределения. Т.е.

мы, вроде бы, недооцениваем погрешность[20] восстановления плотности распределения. Но следует понимать, что $J(C)^{½}$ - это аналог лишь средней квадратической ошибки, и как индивидуальные измерения вполне могут отклоняться от среднего значения более чем на величину средней квадратической ошибки, так и восстановленная функция местами может отклоняться от истинных значений плотности распределения более чем на $J(C)^{½}$.

Наконец, поставим еще один вычислительный эксперимент, более всего приближенный к реальности. Пока мы пытались восстановить плотность распределения по выборке, каждый элемент которой был точным[21] результатом работы генератора случайных чисел. Но при реальных измерениях точные результаты не могут быть нам известны. Индивидуальные измерения потока метана, имеющего в природе некоторое распределение (пусть, например, логарифмически-нормальное), получаются нами с некоторыми погрешностями. Можем ли мы восстановить исходное распределение, если измерения зашумлены погрешностями?

Чтобы смоделировать эту ситуацию, «испортим» каждое значение выборки табл. 2 погрешностью, распределенной по нормальному закону. Для реальных измерений потока погрешность в среднем составляет десятки процентов (см., например, [*Panikov et al.,* 1997; *Kotsyurbenko et al.*, 2004; *Глаголев с соавт*., 2007; *Шнырев и Глаголев*, 2007]). На рис. №5 приведены результаты восстановления плотности распределения функцией `ksdensity` по некоторым использованным в вычислительном эксперименте «испорченным» шумом выборкам. На этом рис. мы не приводим доверительную область, поскольку она практически такая же, что и на рис. №4. Из рис. №5 очевидно, что качество восстановления по зашумленным данным практически такое же, что и для «точных» данных. На первый взгляд совершенно удивительной, почти мистической, кажется возможность восстановления плотности распределения по данным, зашумленным на 160%. Достаточно сказать, что, например, 50-й элемент выборки превратился из «истинного» значения 316.54 в 50.82! Как же можно по настолько сильно «испорченной» выборке восстановить исходную плотность распределения?! Конечно, по значению 50.82 уже никак не восстановишь 316.54! Но не будем забывать о погрешностях и весах!!! Значение 50.82 имеет очень большую погрешность (265.72), а, значит, при построении плотности распределения это значение будет

[20] Кстати, заметим, что MATLAB-функция ksdensity вообще не дает какой-либо оценки погрешности; «усы» рассчитаны с использованием формулы (№9) и представляют собой $\pm J(C)^{½}$.

[21] Точнее говоря, «относительно точным». Из табл. 2 видно, что мы везде осуществили округление до 2-го десятичного знака, следовательно погрешность задания наименьшего элемента выборки составляет около 6% ($0.005/0.08 \cdot 100\% = 6.25\%$). Для наибольшего элемента выборки эта погрешность совершенно незначительна: $0.005/316.54 \cdot 100\% = 0.0016\%$.

учитываться с пренебрежимо малым весом ($1/265.72^2 \approx 1.4 \cdot 10^{-5}$), т.е., фактически, 50-й элемент выборки, вообще не будет учитываться при восстановлении плотности распределения. Точно так же автоматически не будут учитываться и другие «совсем испорченные» элементы. Можно сказать, что чем больше шум, тем меньше элементов выборки используется для восстановления плотности распределения.

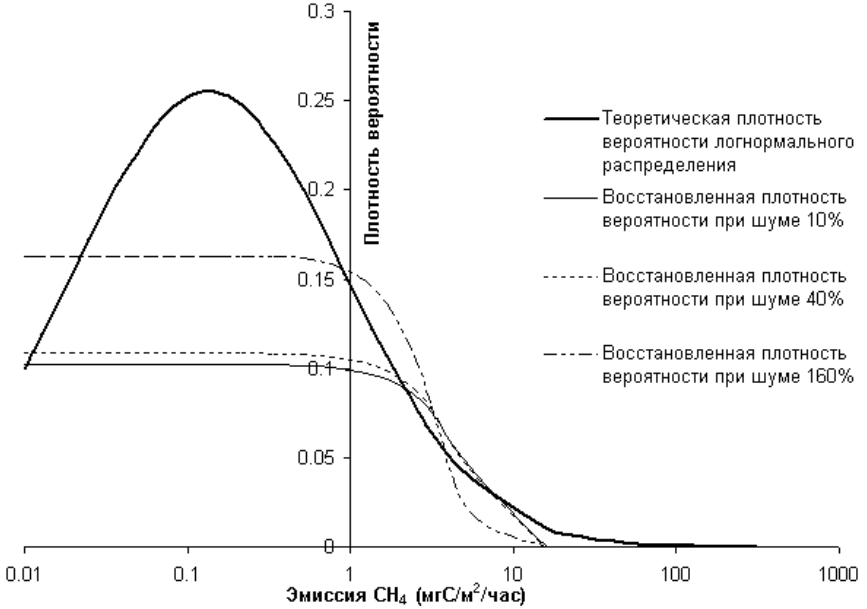

***Рис. 5**. Восстановление плотности распределения по зашумленным данным методом Розенблатта-Парзена.*

Список литературы

Амосов А.А., Дубинский Ю.А., Копченова Н.В. 1994. Вычислительные методы для инженеров. – М.: Высшая школа. – 544 с.

Белов Ю.А., Егоров Б.М., Козлов Н.Н., Ляшко И.И., Макаров В.Л. 1990. Математическое обеспечение сложного эксперимента. – Т. 5: Проблемы построения математического и программного обеспечения измерительно-вычислительных комплексов. – Киев: Наук. думка. – 368 с.

Бронштейн И.Н., Семендяев К.А. 1980. Справочник по математике для инженеров и учащихся втузов. – М.: Наука. – 544 с.

Вапник В.Н., Глазкова Т.Г., Кощеев В.А., Михальский А.И., Червоненкис А.Я. 1984. Алгоритмы и программы восстановления зависимостей. – М.: Наука.

Воскобойников Ю.Е., Преображенский Н.Г., Седельников А.И. 1984. Математическая обработка эксперимента в молекулярной газодинамике. – Новосибирск: Наука. СО.

Глаголев М.В. 2006. Математическое моделирование метанокисления в почве // Труды Института микробиологии имени С.Н. Виноградского РАН. Вып. XIII: К 100-летию открытия метанотрофии / Отв. ред. В.Ф. Гальченко. – М.: Наука. – С. 315-341.

Глаголев М.В., Головацкая Е.А., Шнырев Н.А. 2007. Эмиссия парниковых газов на территории Западной Сибири // *Сибирский экологический журнал*, **14**(2), 197-210.

Дмитриев Е.А. 1995. Математическая статистика в почвоведении. – М.: Изд-во МГУ. – 320 с.

Живописцев Ф.А., Иванов В.А. 1993. Статистический анализ в экспериментальной ядерной физике. – М.: Энергоатомиздат. – 208 с.

Калиткин Н.Н. 1978. Численные методы. - М.: Наука. – 512 с.

Карандеев Д.А., Эйсымонт И.М. 1998. Проблема оценивания плотности вероятности по эмпирическим данным / *Управление большими системами*, Вып. 1. - М.: ИПУ РАН. - С.48-57.

Косарев Е.Л. 2003. Методы обработки экспериментальных данных. – М.: МФТИ. – 256 с.

Косарев Е.Л. 2008. Методы обработки экспериментальных данных. – М.: ФИЗМАТЛИТ. – 208 с.

Костылев А.А., Миляев П.В., Дорский Ю.Д., Левченко В.К., Чикулаева Г.А. 1991. Статистическая обработка результатов экспериментов на микро-ЭВМ и программируемых калькуляторах. – Л.: Энергоатомиздат. ЛО. – 304 с.

Крянцев А.В., Лукин Г.В. 2006. Математические методы обработки неопределенных данных. – М. ФИЗМАТЛИТ. – 216 с.

Лагутин М.Б. 2007. Наглядная математическая статистика. – М.: БИНОМ. Лаборатория знаний. – 472 с.

Лапко А.В., Лапко В.А., Соколов М.И., Ченцов С.В. 2000. Непараметрические системы классификации. – Новосибирск: Наука. – 240 с.

Лапко А.В., Ченцов С.В., Крохов С.И., Фельдман Л.А. 1996. Обучающиеся системы обработки информации и принятия решений. – Новосибирск: Наука. – 296 с.

Ллойд Э., Ледерман У. (ред.) 1989. Справочник по прикладной статистике. Т. 1. – М.: Финансы и статистика. – 510 с.

Мацевитый Ю.М., Лушпенко С.Ф. 1990. Идентификация теплофизических свойста твердых тел. – Киев: Наук. думка. – 216 с.

Паников Н.С. 1995. Таежные болота – глобальный источник атмосферного метана? // *Природа*. №6. С. 14-25.

Петров И.Б., Лобанов А.И. 2006. Лекции по вычислительной математике. – М.: Интернет-Университет Информационных Технологий; БИНОМ. Лаборатория знаний. – 523 с.

Румшиский Л.З. 1971. Математическая обработка результатов эксперимента. – М.: Наука. – 192 с.

Тихонов А.Н., Арсенин В.Я. 1979. Методы решения некорректных задач. – М.: Наука.

Шнырев Н.А., Глаголев М.В. 2007. Характерные значения потоков метана из болот Западной Сибири // Торфяники Западной Сибири и цикл углерода: Прошлое и настоящее: Материалы Второго Международного полевого симпозиума (Ханты-Мансийск, 24 августа – 2 сентября 2007 г.) / Под ред. акад. С.Э. Вомперского. – Томск: Изд-во НТЛ. – с. 144-146.

Cao M., Dent J.B., Heal O.W. 1995. Modeling methane emissions from rice paddies // *Global Biogeochemical Cycles*, **9**, 183-195.

Glagolev M.V. 1998. Modeling of Production, Oxidation and Transportation Processes of Methane // Global Environment Research Fund: Eco-Frontier Fellowship (EFF) in 1997. - Tokyo: Environment Agency. Global Environment Department. Research & Information Office. – p. 79-111.

Kotsyurbenko O.R., Chin K.-J., Glagolev M.V., Stubner S., Simankova M.V., Nozhevnikova A.N., Conrad R. 2004. Acetoclastic and hydrogenotrophic methane production and methanogenic populations in an acidic West-Siberian peat bog // *Environmental Microbiology*, **6**(11), 1159-1173.

Panikov N.S., Glagolev M.V., Kravchenko I.K., Mastepanov M.A., Kosych N.P., Mironycheva-Tokareva N.P., Naumov A.V., Inoue G., Maxutov S. 1997. Variability of methane emission from west-siberian wetlands as related to vegetation type // *J. Ecol. Chem.*, **6**(1), 59-67.

Van der Linden A.M.A., Van Veen J.A., Frissel M.J. 1987. Modeling soil organic matter levels after long-term applications of crop residues, and farmyard and green manures // *Plant and Soil*, **101**, 21-28.

Walter B.P., Heimann M., Shannon R.D., White J.R. 1996. A process-based model to derive methane emissions from natural wetlands // *Geophysical Research Letters*, **23**, 3731-3734.

RECONSTRUCTION OF PROBABILITY DENSITY DISTRIBUTION BY HISTOGRAM METHOD IN SOIL SCIENCE AND ECOLOGY

Glagolev M.V., Sabrekov A.F.

Different methods of probability density reconstruction are investigated in that article. We shown that "histogram method with equal intervals of groupment" which is the most frequently using in soil science and ecology for the solving of that task, is really ill-suited. The "histogram method with intervals of equal probability" is recommended among the simplest methods for solving of that task.