УДК 51-74:665.71

В.П. КОВАЛЕНКО, Е.И. ВЫБОЙЧЕНКО, канд. техн. наук, **Д.О. СКОБЕЛЕВ,** канд. эконом. наук
(ФГУП «Всероссийский научно-исследовательский институт стандартизации материалов
и технологий», г. Москва)

Влияние **способа формирования выборок на ограничительные нормы нефтепродуктов**

Ключевые слова: ограничительные нормы, качество нефтепродуктов, обучающая выборка, проверочная выборка, генеральная совокупность.

Подробно описывается один из этапов алгоритма формирования ограничительных норм показателей качества нефтепродуктов с использованием метода группового учёта аргументов. Проанализированы четыре различных способа получения обучающей и проверочной выборок. На основании результатов измерений в реальных лабораториях доказана независимость ограничительных норм от способа формирования выборок.

В предыдущей публикации [1] было дано определение понятия «ограничительные нормы» и предложены общие рекомендации, позволяющие получить и оценить регламентированные значения показателей качества. Одним из этапов рассмотренного алгоритма является формирование обучающей (часть "А") и проверочной (часть "В") частей выборки [1] для определения оптимальной модели, описывающей результаты измерений показателя качества. Существует ряд способов, позволяющих из исходной выборки (генеральной совокупности) выделить необходимые для работы части, но выбрать оптимальный для данной ситуации без проведения дополнительного анализа достаточно сложно. Поэтому в настоящей публикации предпринята попытка оценить влияние различных способов формирования рабочих выборок при определении ограничительных норм показателей качества нефтепродуктов. Рассмотрим и проанализируем следующие способы:

1. Разделение исходной выборки на две части (пополам);
2. Точечный выбор случайных значений из генеральной совокупности;
3. Разделение исходной выборки на две части по принципу «чётное-нечётное»;
4. Разделение исходной выборки на две части по принципу «3 через 3».

Подробно рассмотрим влияние каждого из этих способов на ограничительные нормы.

Способ 1. Разделение исходной выборки на две части (пополам)

Исходные данные представляют собой 2223 последовательных измерения температуры вспышки в открытом тигле топочного мазута марки 100 (М-100, ГОСТ 10585–2013), полученных согласно ГОСТ 4333–87 в реальной лаборатории предприятия, выпускающего товарный мазут марки «М-100». Все значения с 1 по 1112 являются значениями обучающей выборки, а значения с 1113 по 2223 являются значениями проверочной выборки. Далее, согласно алгоритму [1, 2], необходимо построить 31 модель для 5 уровней сложности с использованием обучающей выборки и рассчитать критерии регулярности для данных моделей с использованием проверочной выборки. Для получения симметричных критериев проводят аналогичный расчёт, поменяв обучающую и проверочную выборки местами. Затем определяют 10 наиболее оптимальных моделей по критерию регулярности (10 моделей с минимальными критериями регулярности). Эти модели вместе со значениями критериев регулярности приведены ниже в порядке возрастания критериев.

Оптимальные модели по критерию регулярности (K_{reg}) при первом способе формирования выборок

Модель	K_{reg}
1. $k + k_1 \cdot x^2$	309203,1
2. $k + k_1 \cdot x^3$	448219,5
3. $k + k_1 \cdot x$	534293,3
4. k	857505,3
5. $k + k_1 \cdot x + k_2 \cdot x^2$	1123136,9
6. $k + k_1 \cdot x + k_2 \cdot x^3$	2811386,8
7. $k + k_1 \cdot x^4$	2930268,3
8. $k + k_1 \cdot x^2 + k_2 \cdot x^3$	12057258,9
9. $k + k_1 \cdot x + k_2 \cdot x^4$	14138224,1
10. $k \cdot x$	33618485,6

Для получения результирующей (единственной) оптимальной модели рассчитывают критерии минимума смещения (K_{unbias}) для отобранных 10 моделей. Результаты данного расчёта приведены ниже.

Критерии минимума смещения при первом способеформирования выборок

Модель	K_{unbias}
1. $k + k_1 \cdot x^2$	33920,8
2. $k + k_1 \cdot x^3$	164753,5
3. $k + k_1 \cdot x$	5029054,9
4. k	534075,8
5. $k + k_1 \cdot x + k_2 \cdot x^2$	879759,8
6. $k + k_1 \cdot x + k_2 \cdot x^3$	3221853,6
7. $k + k_1 \cdot x^4$	2649807,2
8. $k + k_1 \cdot x^2 + k_2 \cdot x^3$	11916088,1
9. $k + k_1 \cdot x + k_2 \cdot x^4$	13461281,6
10. $k \cdot x$	29680768,6

Для первого случая формирования выборок по данным, предоставленным реальной лабораторией, оптимальная модель имеет вид $y = k + k_1 \cdot x^2$.

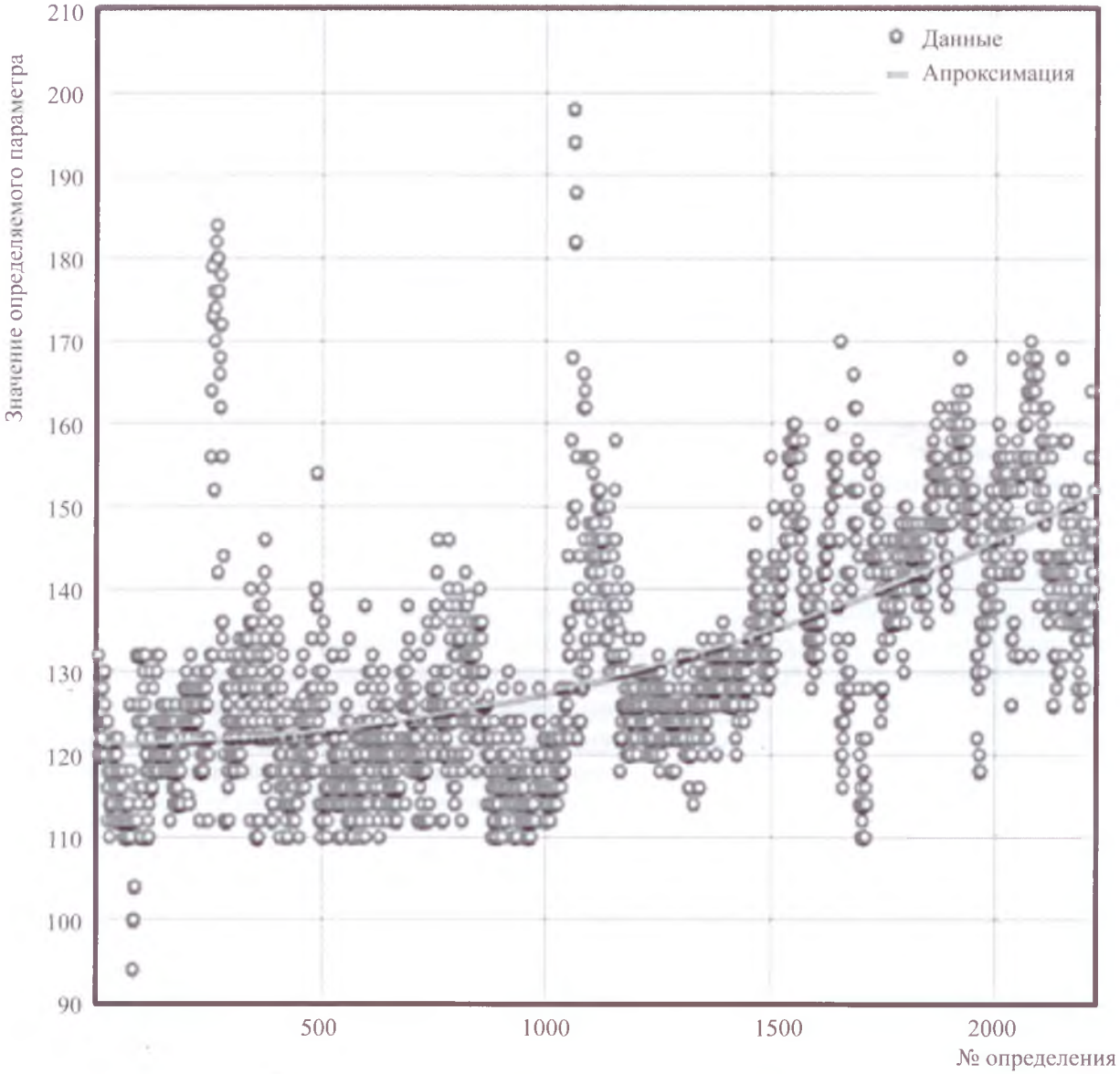
При помощи метода наименьших квадратов (МНК) определим коэффициенты k и k_1 по всей генеральной совокупности

$k = 121{,}3$; $k_1 = 6{,}1 \cdot 10^{-6}$.

Тогда $y = 6{,}1 \cdot 10^{-6} \cdot x^2 + 121{,}3$ – оптимальная модель, описывающая исходные данные (**рисунок**).

Способ 2. Точечный выбор случайных значений из генеральной совокупности

При таком способе формирования обучающей выборки из генеральной совокупности выделяют случайным образом 1112 значений и располагают их в порядке возрастания номера измерения. Оставшиеся 1111 значений представляют собой проверочную выборку.



Оптимальная модель, описывающая исходные данные при первом способе формирования выборок

Аналогично предыдущему случаю рассчитывают симметричные критерии регулярности и отбирают 10 наиболее оптимальных моделей.

Оптимальные модели по критерию регулярности (K_{reg}) при втором способе формирования выборок

Модель	K_{reg}
1. $k + k_1 \cdot x + k_2 \cdot x^3 + k_3 \cdot x^4$	270126,2
2. $k + k_1 \cdot x^3 + k_2 \cdot x^4$	272996,7
3. $k + k_1 \cdot x + k_2 \cdot x^2 + k_3 \cdot x^4$	274195,6
4. $k + k_1 \cdot x^2$	276609,7
5. $k + k_1 \cdot x^2 + k_2 \cdot x^3$	277137,0
6. $k + k_1 \cdot x^2 + k_2 \cdot x^4$	277351,2
7. $k + k_1 \cdot x + k_2 \cdot x^2$	277511,0
8. $k + k_1 \cdot x + k_2 \cdot x^4$	282027,7
9. $k + k_1 \cdot x^3$	283246,2
10. $k + k_1 \cdot x^2 + k_2 \cdot x^3 + k_3 \cdot x^4$	285012,8

Для получения результирующей (единственной) оптимальной модели, как и в первом случае, рассчитывают критерии минимума смещения (K_{unbias}) для отобранных 10 моделей. Результаты данного расчета приведены ниже.

Критерии минимума смещения при втором способе формирования выборок

Модель	K_{unbias}
1. $k + k_1 \cdot x + k_2 \cdot x^3 + k_3 \cdot x^4$	6366,7
2. $k + k_1 \cdot x^3 + k_2 \cdot x^4$	2777,5
3. $k + k_1 \cdot x + k_2 \cdot x^2 + k_3 \cdot x^4$	3944,4
4. $k + k_1 \cdot x^2$	**1607,4**
5. $k + k_1 \cdot x^2 + k_2 \cdot x^3$	2124,2
6. $k + k_1 \cdot x^2 + k_2 \cdot x^4$	3059,6
7. $k + k_1 \cdot x + k_2 \cdot x^2$	3573,4
8. $k + k_1 \cdot x + k_2 \cdot x^4$	1902,3
9. $k + k_1 \cdot x^3$	2343,8
10. $k + k_1 \cdot x^2 + k_2 \cdot x^3 + k_3 \cdot x^4$	58864,1

Таким образом, как и в первом случае, оптимальная модель имеет вид $y = k + k_1 \cdot x^2$.

Так как оптимальные модели идентичные и исходная выборка для расчётов одна, коэффициенты k и k_1 будут одинаковыми как в первом, так и во втором случае.

Способ 3. Разделение исходной выборки на две части по принципу «чётное-нечётное»

Чтобы получить обучающую выборку из генеральной совокупности вычленяют только значения с нечётными порядковыми номерами измерений. Проверочная же выборка будет представлять собой совокупность, состоящую из значений с чётными порядковыми номерами измерений. Далее, как и в первых двух случаях, рассчитывают критерии регулярности и определяют 10 наиболее оптимальных моделей по полученным критериям.

Оптимальные модели по критерию регулярности (K_{reg}) при третьем способе формирования выборок

Модель	K_{reg}
1. $k + k_1 \cdot x + k_2 \cdot x^2 + k_3 \cdot x^3 + k_4 \cdot x^4$	206068,4
2. $k + k_1 \cdot x + k_2 \cdot x^3 + k_3 \cdot x^4$	268431,9
3. $k + k_1 \cdot x^3 + k_2 \cdot x^4$	271133,8
4. $k + k_1 \cdot x + k_2 \cdot x^2 + k_3 \cdot x^4$	271581,5
5. $k + k_1 \cdot x + k_2 \cdot x^2 + k_3 \cdot x^3$	274056,6
6. $k + k_1 \cdot x^2 + k_2 \cdot x^4$	275348,9
7. $k + k_1 \cdot x + k_2 \cdot x^2$	275621,2
8. $k + k_1 \cdot x^2$	275692,0
9. $k + k_1 \cdot x^2 + k_2 \cdot x^3$	275741,9
10. $k + k_1 \cdot x + k_2 \cdot x^3$	278109,8

Аналогично предыдущим двум случаям для получения результирующей оптимальной модели рассчитывают критерии минимума смещения (K_{unbias}) для отобранных 10 моделей.

Критерии минимума смещения в третьем способе формирования выборок

Модель	K_{unbias}
1. $k + k_1 \cdot x + k_2 \cdot x^2 + k_3 \cdot x^3 + k_4 \cdot x^4$	1439372,2
2. $k + k_1 \cdot x + k_2 \cdot x^3 + k_3 \cdot x^4$	1005,7
3. $k + k_1 \cdot x^3 + k_2 \cdot x^4$	293,9
4. $k + k_1 \cdot x + k_2 \cdot x^2 + k_3 \cdot x^4$	489,3
5. $k + k_1 \cdot x + k_2 \cdot x^2 + k_3 \cdot x^3$	902,5
6. $k + k_1 \cdot x^2 + k_2 \cdot x^4$	223,5
7. $k + k_1 \cdot x + k_2 \cdot x^2$	524,1
8. $k + k_1 \cdot x^2$	**171,2**
9. $k + k_1 \cdot x^2 + k_2 \cdot x^3$	283,4
10. $k + k_1 \cdot x + k_2 \cdot x^3$	1313,2

Получаем, что, как и в первых двух случаях, оптимальная модель имеет вид $y = k + k_1 \cdot x^2$.

Так как оптимальные модели идентичны и исходная выборка для расчётов одна, а отличаются только способы её дробления, коэффициенты k и k_1 будут одинаковыми во всех трёх случаях.

Способ 4. Разделение исходной выборки на две части по принципу «3 через 3»

Для получения проверочной и обучающей выборок из генеральной совокупности вычленяют первые три элемента по порядку, затем пропускают три элемента и так до конца всей выборки. Вычлененные элементы будут составлять обучающую выборку, а оставшиеся – проверочную. Далее, согласно алгоритму, рассчитывают критерии регулярности и определяют 10 наиболее оптимальных моделей по полученным критериям.

Оптимальные модели по критерию регулярности (K_{reg}) при четвёртом способе формирования выборок

Модель	K_{reg}
1. $k + k_1 \cdot x^2 + k_2 \cdot x^3 + k_3 \cdot x^4$	268384,6
2. $k + k_1 \cdot x + k_2 \cdot x^3 + k_3 \cdot x^4$	269932,4
3. $k + k_1 \cdot x^3 + k_2 \cdot x^4$	271436,7
4. $k + k_1 \cdot x + k_2 \cdot x^2 + k_3 \cdot x^4$	272609,5
5. $k + k_1 \cdot x + k_2 \cdot x^2 + k_3 \cdot x^3$	273056,8
6. $k + k_1 \cdot x^2 + k_2 \cdot x^4$	275535,0
7. $k + k_1 \cdot x^2$	275713,7
8. $k + k_1 \cdot x + k_2 \cdot x^2$	275829,6
9. $k + k_1 \cdot x^2 + k_2 \cdot x^3$	275878,4
10. $k + k_1 \cdot x + k_2 \cdot x^3$	278165,9

Аналогично предыдущим случаям для получения результирующей (единственной) оптимальной модели рассчитывают критерии минимума смещения (K_{unbias}) для отобранных 10 моделей.

Критерии минимума смещения при четвёртом способе формирования выборок

Модель	K_{unbias}
1. $k + k_1 \cdot x^2 + k_2 \cdot x^3 + k_3 \cdot x^4$	4087,7
2. $k + k_1 \cdot x + k_2 \cdot x^3 + k_3 \cdot x^4$	2181,9
3. $k + k_1 \cdot x^3 + k_2 \cdot x^4$	1015,9
4. $k + k_1 \cdot x + k_2 \cdot x^2 + k_3 \cdot x^4$	3634,7
5. $k + k_1 \cdot x + k_2 \cdot x^2 + k_3 \cdot x^3$	670,0
6. $k + k_1 \cdot x^2 + k_2 \cdot x^4$	475,4
7. $k + k_1 \cdot x^2$	**204,2**
8. $k + k_1 \cdot x + k_2 \cdot x^2$	665,6
9. $k + k_1 \cdot x^2 + k_2 \cdot x^3$	461,0
10. $k + k_1 \cdot x + k_2 \cdot x^3$	893,1

Получаем, что, как и в предыдущих случаях, оптимальная модель, описывающая исходный набор данных, имеет вид $y = k + k_1 \cdot x^2$, а коэффициенты k и k_1 будут одинаковыми во всех случаях ввиду причин, описанных выше.

В результате представленного анализа можно сделать заключение о том, что общий вид оптимальной модели не зависит от способа формирования выборок. Поэтому значение ограничительных норм также не зависят от способа формирования обучающей и проверочной выборок. Эти выводы были подтверждены данными и других реальных лабораторий.

Список литературы

1. Скобелев Д.О., Коваленко В.П., Выбойченко Е.И. Алгоритм формирования ограничительных норм показателей качества нефтепродуктов с использованием метода группового учёта аргументов // Мир нефтепродуктов. Вестник нефтяных компаний. – 2015. – № 1 –С. 45–49.

2. Ивахненко А.Г., Степашко В.С. Помехоустойчивость моделирования – Киев: Наукова Думка, 1985.

Kovalenko V.P., Vyboychenko E.I., Skobelev D.O.
(Federal State Unitary Enterprise «Russian Research Institute on Standardization of Materials and Technologies», Moscow)

The influence of the method of sampling for restrictive standard of petroleum products

Keywords: restrictive standard; quality; petroleum products; learning sample, validation sample, general totality.

Abstracts

Article «The influence of the method of sampling for restrictive standard of petroleum products» is a continuation of the publication [1]. It considers in detail the procedure of formation of the training and validation samples, which is an integral part of the algorithm [1]. In the main part of the article outlines four possible ways to solve the problem. It also presents calculations with graphs, comparative analysis of these methods and the obtained regularities. Analysis of real data allowed to conclude that the general form of the optimal model does not depend from the method of sampling. From this, it follows that the restrictive standards does not depend on the method of formation of the trainimg and validation samples.

References

1. Skobelev D.O., Kovalenko V.P., Vyboychenko E.I. The algorithm of formation restrictive standard of petroleum products quality usimg group method of data handling [Algoritm formirovaniya ogranichitel'nih norm pokazateley kachestva nefteproductov s ispol'zovaniem metoda gruppovogo ucheta argumentov]. *Mir nefteproduktov. Vestnik neftyanih kompaniy – World of oil products. The Oil Companies' Bulletin*, 2015, no. 1, pp. 45–49.

2. Ivakhnenko A.G., Stepashko V.S. *Pomekhoustoychivost' modelirovaniya* [Modeling noise stability]. Kiev, Naukova dumka Publ., 1985.