**КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2016 Т. 8 № 1 С. 19–36**

Ки&М

МАТЕМАТИЧЕСКИЕ ОСНОВЫ И ЧИСЛЕННЫЕ МЕТОДЫ МОДЕЛИРОВАНИЯ

УДК: 51-76, 51-77, 57.042

Алгоритм метода по расчету границ качественных классов для количественных характеристик систем и по установлению взаимосвязей между характеристиками. Часть 1. Расчеты для двух качественных классов

Д. В. Рисник[1,a], А. П. Левич[1], П. В. Фурсова[1], И. А. Гончаров[2]

[1] Московский государственный университет им. М. В. Ломоносова, биологический факультет,
Россия, 119991, ГСП-1, г. Москва, Ленинские горы, д. 1, стр. 12
[2] Московский государственный университет им. М. В. Ломоносова,
механико-математический факультет,
Россия, 119991, ГСП-1, г. Москва, Ленинские горы, д. 1, стр. 1

E-mail: [a] biant3@mail.ru

*Получено 19 мая 2015 г.,
после доработки 22 декабря 2015 г.*

Предложен метод расчета границ качественных классов для количественных характеристик систем любой природы. Метод позволяет установить: связи, не поддающиеся обнаружению при помощи корреляционного и регрессионного анализа; границы для качественных классов индикатора состояния систем и факторов, влияющих на это состояние; вклад факторов в степень «неприемлемости» значений индикатора; достаточность программы наблюдений за факторами для описания причин «неприемлемости» значений индикатора.

Ключевые слова: анализ связи, максимизация силы связи, индикаторы, факторы, границы качественных классов, вклад фактора

Citation: *Computer Research and Modeling*, 2016, vol. 8, no. 1, pp. 19–36 (Russian).

Работа выполнена при частичной поддержке РФФИ (гранты №№ 13-04-01027а, 12-07-00580а, 14-04-01873а, 14-04-00143а, 15-04-02601а, 16-04-01024а).

КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2016 Т. 8 № 1 С. 19–36

Ки&М

МАТЕМАТИЧЕСКИЕ ОСНОВЫ И ЧИСЛЕННЫЕ МЕТОДЫ МОДЕЛИРОВАНИЯ

The algorithm of the method for calculating quality classes' boundaries for quantitative systems' characteristics and for determination of interactions between characteristics. Part 1. Calculation for two quality classes

D. V. Risnik[1], A. P. Levich[1], P. V. Fursova[1], I. A. Goncharov[2]

[1] Lomonosov Moscow State University, Faculty of Biology, 1-12 Leninskie Gory, Moscow, GSP-1, 119991, Russia
[2] Lomonosov Moscow State University, Faculty of Mechanics and Mathematics, 1-1 Leninskie Gory, Moscow, GSP-1, 119991, Russia

Abstract. — A calculation method for boundaries of quality classes for quantitative systems characteristics of any nature is suggested. The method allows to determine interactions which are not detectable using correlation and regression analysis; quality classes' boundaries of systems' condition indicator and boundaries of the factors influencing this condition; contribution of the factors to a degree of «inadmissibility» of indicator values; sufficiency of the program observing the factors to describe the causes of «inadmissibility» of indicator values.

Keywords: analysis of interaction, maximization of interaction power, indicators, factors, boundaries of quality classes, factor contribution

Citation: *Computer Research and Modeling*, 2016, vol. 8, no. 1, pp. 19–36 (Russian).

Работа выполнена при частичной поддержке РФФИ (гранты №№ 13-04-01027а, 12-07-00580а, 14-04-01873а, 14-04-00143а, 15-04-02601а, 16-04-01024а).

Введение

При изучении систем любой природы выделим ее «внутренние» и «внешние» характеристики. Среди «внутренних» существуют те, которые характеризуют состояние системы на шкале «приемлемость–неприемлемость», — «индикаторы» состояния. Среди «внешних» — те, которые влияют на состояние системы, — «факторы». Так, для экосистемы индикаторами могут служить, например, численность видов биоты, а факторами — физико-химические характеристики среды. Для производственной системы индикаторами могут служить объем и ассортимент выпускаемой продукции, а факторами — доступная энергия, рабочая сила, инвестиции, материальные ресурсы и др. На индикаторы состояния антропной экосистемы — рождаемость, смертность, заболеваемость — влияют, в частности, такие факторы, как технологии очистки сточных вод, воздушных выбросов, утилизации бытовых отходов; среднедушевая заработная плата; количество автотранспорта и тип горючего; количество зеленых насаждений, климатические факторы, качество продуктов питания, уровень образования и просвещения и т. п.

Анализ натурных данных, полученных в результате измерения или расчета каких-либо индикаторных характеристик и факторов, сложен с точки зрения статистики. Примером такого анализа может служить поиск связей между биологическими и физико-химическими характеристиками, значения которых получены в наблюдениях за природными экосистемами. Сложность подобного анализа связана с «размытостью» («неоднозначностью») зависимостей между переменными, обусловленной одновременным действием на индикаторную характеристику множества факторов.

Корреляционный анализ «размытых» зависимостей, как правило, характеризует отраженную в них связь как слабую и/или незначимую.

Многомерный регрессионный анализ также осложнен рядом обстоятельств [Налимов, Чернова, 1965]: 1) нельзя достаточно строго проверить гипотезу об адекватности представленных данных выбранной модели, поскольку, рассматривая натурные данные как пассивный многофакторный эксперимент, трудно оценить ошибки эксперимента; 2) невозможно выделить парциальное влияние на индикатор попарно скоррелированных (интеркоррелированных) факторов.

Даже если регрессионный анализ позволяет получить высокий коэффициент множественной регрессии, остается нерешенной исходная задача: поиск «тесноты» связи между биологической переменной и каждым из факторов, поскольку частные коэффициенты корреляции остаются незначительными и незначимыми. Следует также отметить [Рисник, Рыбка, 2011] особенность сбора натурных данных: единичные значения по отдельным показателям в матрице данных наблюдений часто отсутствуют. Наличие подобных пропусков приводит к отсеву из анализируемого массива большинства наблюдений как неудовлетворяющих требованиям регрессионного анализа, поскольку матрица анализируемых данных должна быть заполнена полностью.

Таким образом, для анализа натурных данных необходим метод, менее подверженный перечисленным трудностям и позволяющий выявлять связи, скрытые при рассмотрении парных зависимостей между характеристиками из-за влияния остальных факторов. Один из методов анализа «плохо организованных» данных — переход от количественных переменных к их качественным классам. Такими классами могут быть: «низкие», «средние» и «высокие» значения; «благополучные» и «неблагополучные»; «допустимые» и «недопустимые»; «нормальные» и «нарушенные»; «приемлемые» и «неприемлемые» и т. п. После выделения качественных классов возможен поиск связей уже между качественными классами различных переменных.

Однако и анализ качественных переменных сталкивается по меньшей мере с двумя трудностями. Во-первых, возникает проблема выбора объективного критерия для выделения качественных классов: какие значения считать «высокими» и какие — «низкими», какие — «допустимыми» и какие — «недопустимыми». Обычно границы между качественными классами для природных данных вводят в большой степени субъективно: например, диапазон измерения ха-

рактеристики делят на равные интервалы (в линейной или логарифмической шкале) или экспертным образом назначают «высокие» и «низкие» или другие значения. Субъективность выбора границ ставит под сомнение обоснование всех последующих процедур установления связей. Вторая трудность вызвана упомянутым выше неустранимым *in situ* влиянием на индикаторную характеристику всех внешних факторов и состоит в том, что некоторое значение этой характеристики может быть вызвано не исследуемым фактором, а каким-либо другим фактором, действующим одновременно с исследуемым.

Метод расчета границ качественных классов для количественных характеристик систем (метод ГКК) создан для анализа связей между индикаторными характеристиками и воздействующими на них факторами путем расчета границ качественных классов с учетом необходимости преодоления перечисленных выше трудностей.

Отметим, что метод ГКК позволяет обнаружить, но не позволяет объяснить связи между переменными. Этап обнаружения связей тем не менее с необходимостью предшествует этапу объяснения, который представляет собой самостоятельную задачу в предметной области, к которой принадлежит исследуемая система. Этап объяснения не входит в круг задач, решаемых методом ГКК.

Предложенный метод ГКК разработан на базе имеющихся подходов анализа связей между характеристиками: методов классического детерминационного анализа (коэффициентов Валлиса [Goodman, Kruskal, 1954, 1959, 1963, 1972] и Гуттмана [Guttman, 1941]), метода детерминационного анализа С. В. Чеснокова [Чесноков, 1982], метода локальных экологических норм [Левич, Булгаков, Максимов, 2004; Левич, Булгаков, Рисник, 2010; Левич и др., 2011; Левич и др., 2012; Левич, Булгаков, Максимов, 2013; Левич и др., 2013], метода максимизации связей между качественными классами [Рисник и др., 2013а, Рисник и др., 2013б].

1. Принцип работы алгоритма для расчета границ качественных классов

Массив данных наблюдений представляет собой набор значений для некоторых индикаторных характеристик и соответствующих им значений различных факторов. Рассмотрим взаимосвязь между одной индикаторной характеристикой и одним фактором. Каждую точку на плоскости «индикатор–фактор» можно отнести к некоторому классу качества по индикаторной характеристике (например, к «приемлемым» или «неприемлемым» значениям) и по фактору (например, к «допустимым» или «недопустимым» значениям). Если некоторая характеристика действительно является индикаторной для воздействия фактора, то «приемлемые» значения индикаторной характеристики должны встречаться в наблюдениях только совместно с «допустимыми» значениями фактора, а «недопустимые» значения фактора — только совместно с «неприемлемыми» значениями индикаторной характеристики. Поскольку к «неприемлемым» значениям индикаторной характеристики может приводить «недопустимость» значений хотя бы одного из нескольких влияющих на этот индикатор факторов, то возможна ситуация, когда «неприемлемое» значение индикаторной характеристики встречается совместно с «допустимым» значением рассматриваемого фактора. При этом, если все элементы изучаемой системы (например, продукты производства, экосистемы, популяции организмов, значения экономических индикаторов в разных точках временной шкалы) одинаково чувствительны к действию фактора, «недопустимые» значения фактора никогда не должны приводить к «приемлемым» значениям индикаторной характеристики независимо от действия других факторов. Суть метода расчета границ между классами качества (метода ГКК) состоит в поиске таких границ между «приемлемыми» и «неприемлемыми» значениями индикатора (границ нормы индикатора) и между «допустимыми» и «недопустимыми» значениями фактора (границ нормы фактора), при которых существует пустая область плоскости «индикатор–фактор», соответствующая «приемлемым» значениям индикатора при «недопустимых» значениях фактора (или с учетом погрешности измерений — «достаточно пустая» область). Алгоритм поиска состоит в проведе-

нии границ, при которых количество точек-наблюдений в указанной области минимально в сравнении с другими вариантами положения границ и в сравнении этого количества с критерием «допустимой пустоты».

Для оценки «степени минимальности» количества точек-наблюдений в области «приемлемое значение индикатора при недопустимом значении фактора» в методе ГКК предложен, так называемый *критерий существенности* (подробно он будет описан в разделе 3.2), величина которого тем больше, чем меньше точек в указанной области. Работа алгоритма по расчету границ качественных классов для пары «индикатор–фактор» заключается в переборе всех возможных положений границ, в расчете для каждого положения границ критерия существенности и выборе таких границ, для которых критерий существенности максимален.

Однако простого выбора границ, соответствующих максимальному значению критерия существенности, недостаточно. Причиной этого является тот факт, что максимум критерия может быть обусловлен не наличием связи в исследуемой паре «индикатор-фактор», а дополнительными обстоятельствами: недостаточной пустотой минимизируемой области, недостаточным количеством наблюдений для других областей и совокупности совместных наблюдений индикатора и фактора, статистической незначимостью полученных результатов. Поэтому для корректного выбора нужных границ норм необходимо введение *критериев проверки границ* (см. раздел 3.3). Факторы, удовлетворяющие критериям проверки границ, названы потенциально существенными для неприемлемости значений индикатора.

В качестве возможного положения границ исследуют комбинации всех встречавшихся значений индикатора и фактора.

Алгоритм метода предполагает, что существуют единые для большинства факторов границы между приемлемыми и неприемлемыми значениями индикатора, соответствующие наибольшему числу потенциально существенных факторов.

Если граница нормы по индикатору или по фактору известна исследователю, алгоритм позволяет ее зафиксировать и проводить поиск только парной к ней границы нормы по фактору или индикатору соответственно.

Итак, сформулируем последовательность работы алгоритма по поиску границ нормы индикатора и воздействующих на него факторов:

1) для каждого возможного положения границ пары «индикатор–фактор» проверяют выполнение критериев проверки границ;
2) находят границу по индикатору с наибольшим числом потенциально существенных факторов;
3) для найденной границы по индикатору и всех возможных положений границ потенциально существенных факторов проводят расчет критериев существенности и в качестве границ качественных классов по факторам выбирают те, для которых существенность максимальна;
4) если число потенциально существенных факторов для нескольких границ индикатора одинаково, границу нормы индикатора выбирают среди них по максимальному произведению критериев существенности (см. раздел 3.2).

Кроме выявления существенных факторов и расчета непосредственных значений границ классов качества для индикаторной характеристики и фактора, алгоритм позволяет:

– упорядочить все исследуемые факторы по величине их вклада в «неприемлемость» значений индикаторной характеристики;
– оценить полноту программы наблюдения за исследуемым объектом и дать необходимые рекомендации по ее сокращению или расширению;
– указать предпочтения в выборе индикаторной характеристики состояния исследуемого объекта.

Прежде чем подробно описывать используемые критерии поиска границ, сделаем несколько общих замечаний.

Алгоритм предназначен только для обработки совместных наблюдений индикатора и фактора. То есть каждое наблюдение фактора должно соответствовать наблюдению индикатора, отобранному в той же точке, что и значение индикатора, и в то же время (для оценки предположения о запаздывании влияния факторов допускается временной сдвиг, когда все значения фактора отобраны, например, за месяц до измерения значений индикатора в соответствующих точках).

Кроме того, важно отметить, что в результате работы алгоритма границы норм могут и не быть установлены. Это происходит в том случае, когда ни одно положение границ не удовлетворяет набору критериев их проверки.

Отсутствие результатов поиска может означать: 1) что все значения фактора в исследованном массиве были только «допустимыми», и тогда фактор несущественен как причина «неприемлемости» значений индикаторной характеристики; 2) что все значения фактора были «недопустимыми», в силу чего его роль в «неприемлемости» значений существенна; 3) что все значения индикаторной характеристики были только «приемлемыми», т. е. ни один из факторов не оказывал негативного влияния; 4) что все значения индикаторной характеристики были только «неприемлемыми», т. е. в каждом наблюдении хотя бы одна причина приводила к «неприемлемому» состоянию исследуемого объекта; 5) исследуемая индикаторная характеристика не является удачным индикатором влияния исследуемого фактора. Если границы норм не выявлены ни для одного из исследуемых факторов, можно считать, что исследованная индикаторная характеристика не является удачным индикатором влияния исследованных факторов. В случае когда граница норм индикатора для существенных факторов была найдена, алгоритм метода позволяет анализировать некоторые из указанных выше возможностей.

Стандартные методы анализа связей между характеристиками в таблицах сопряженности (коэффициенты Юла, Пирсона, Валлиса, Гуттмана и Чеснокова) оперируют с априори заданными качественными классами переменных. Это ограничивает возможность описания «размытых» связей между качественными классами количественных переменных, не поддающихся корреляционному и регрессионному анализу. Основные отличия метода ГКК от стандартных статистических методов определения силы связи между качественными классами переменных и от метода максимизации силы связи между качественными классами приведены в таблице 1.

Таблица 1. Отличия метода расчета границ качественных классов для количественных характеристик систем от стандартных статистических методов определения силы связи между качественными классами переменных (коэффициенты Юла, Пирсона [Миркин, 1980], Валлиса [Goodman, Kruskal, 1954, 1959, 1963, 1972], Гуттмана [Guttman, 1941], Чеснокова [1982]) и метода максимизации силы связи между качественными классами [Рисник и др., 2013а; Рисник и др., 2013б]

Метод расчета границ качественных классов для количественных характеристик систем (ГКК)	Методы определения силы связи между качественными классами переменных	Метод максимизации силы связей между качественными классами переменных (МКК)
Применим для количественных признаков	Применим для качественных признаков	Применим для количественных признаков
Определяет силу связи для части распределения, обусловленной односторонним влиянием фактора на индикатор	Определяет силу связи	Определяет силу связи
Находит границы качественных классов путем минимизации количества наблюдений в области, соответствующей «приемлемым» значениям индикаторной характеристики при «недопустимых» значениях независимых характеристик	Границы соответствуют априори существующим границам качественных классов	Находит границы качественных классов путем максимизации силы связи между качественными классами признаков
Применим только для зависимой индикаторной характеристики и влияющих на нее независимых характеристик	Применим для двух характеристик любой природы	Применим для двух характеристик любой природы

Для корректного решения поставленных задач желательно соблюдать репрезентативность исходных данных. В частности, следует учитывать, что достоверность результатов любого статистического анализа обеспечена только при наличии достаточного количества наблюдений. Так, для корреляционного и регрессионного анализа С. Грин [Green, 1991] рекомендует использовать эмпирическое правило $N > 50 + 8g$, где g — количество независимых переменных. В случае менее чем пяти независимых переменных согласно Р. Харрису [Harris, 1985] можно использовать эмпирическое правило $N > 50 + g$. Для проведения факторного анализа [Comrey, Lee, 1992] предложены следующие градации по количеству наблюдений: 50 — слишком мало, 100 — мало, 200 — приемлемо, 300 — хорошо, 500 — очень хорошо, 1000 — отлично. Ограничения, накладываемые на количество совместных наблюдений при использовании метода ГКК, приведены в разделах 2.3 и 3.3 настоящей статьи.

2. Расчет односторонних границ норм для двух классов качества по индикатору и фактору

Перейдем непосредственно к формализации алгоритма метода ГКК. Как было отмечено выше, метод ГКК предполагает выделение нескольких классов качества переменных. Границы норм при этом могут быть односторонними или двусторонними. Термин «односторонние границы нормы» отражает тот факт, что диапазон нормы («приемлемости», «допустимости») ограничен только с одной стороны (сверху или снизу, слева или справа), и означает простейший случай, когда «неприемлемыми» являются только «высокие» или только «низкие» значения индикаторной характеристики, аналогично для фактора — «недопустимы» только высокие или только низкие значения. Термин «двусторонние границы нормы» отражает тот факт, что диапазон нормы ограничен с двух сторон (как сверху, так и снизу; как слева, так и справа). Двусторонние границы нормы фактора необходимо искать, когда к «неприемлемым» значениям индикаторной характеристики приводят как «низкие», так и «высокие» значения фактора, «допустимыми» же являются «средние» значения фактора. Одновременно с двумя границами нормы по фактору при необходимости могут быть найдены две границы нормы для индикатора, когда к «неприемлемым» относятся и «низкие», и «высокие» значения индикатора (например, температура тела теплокровных животных).

В работе рассмотрены случаи двух, трех и произвольного количества классов как с односторонними, так и с двусторонними границами классов качества. Для каждого из случаев приведены формулы для критерия существенности и для критериев проверки границ. Кроме того, приведены формулы для расчета *полнот* (см. раздел 3.3) существенных факторов для найденных границ — полноты факторов, совместные полноты факторов и достаточность программы мониторинга. Полноты позволяют ранжировать факторы и оценивать достаточность программы наблюдений за состоянием объекта исследования.

2.1. Термины и обозначения

На рис. 1 представлена диаграмма зависимости значений индикатора от фактора, на которой приведены обозначения областей и количества наблюдений в этих областях для различных сочетаний двух классов качества по индикатору и фактору. Рассмотрен случай, когда для индикаторной характеристики «приемлемы» высокие значения, а для фактора «допустимы» — низкие.

В приведенном на рис. 1 простейшем случае метод включает одновременный поиск двух границ: 1) границы, разделяющей «приемлемые» и «неприемлемые» значения индикатора, — границы нормы индикатора (ГНИ); 2) границы, разделяющей «допустимые» и «недопустимые» значения фактора, — границы нормы фактора (ГНФ).

Согласно предлагаемому методу необходимо провести поиск такой взаимосвязи между индикаторами и факторами, которая соответствует допустимой «пустоте» единственной области «*b*».

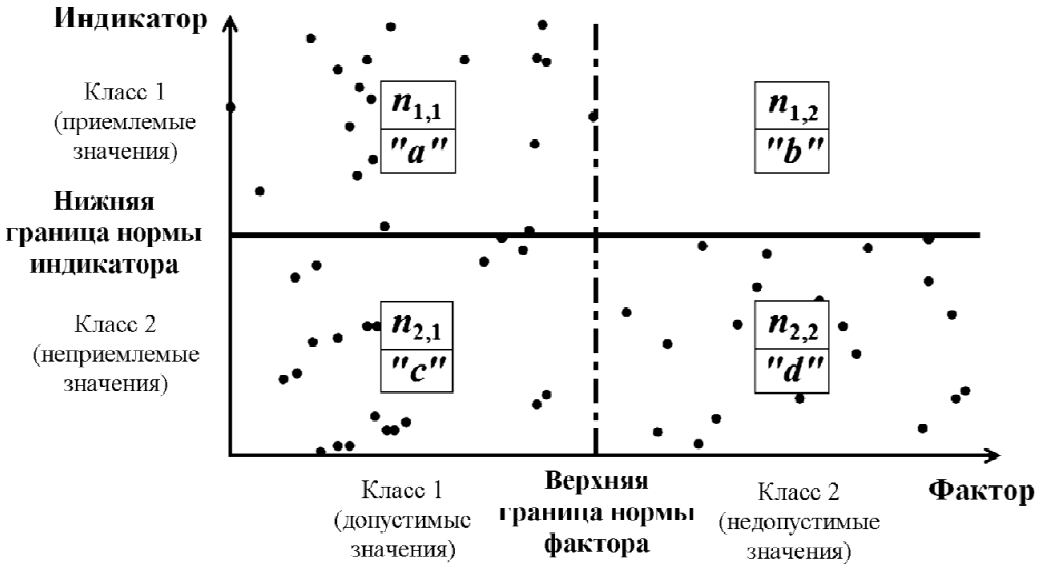

Рис. 1. Классы качества значений индикатора и фактора в случае совокупного влияния на индикатор множества факторов среды. Обозначения: «*a*», «*b*», «*c*», «*d*» — обозначения областей; $n_{m,p}$ — количество наблюдений в области класса «*m*» по индикатору и класса «*p*» по фактору

Распределение значений и положение границ на рис. 1 приведено для примера. Метод также применим к трем другим случаям: 1) когда «приемлемы» высокие значения индикаторной характеристики и «допустимы» высокие значения фактора (метод находит нижнюю ГНИ и нижнюю ГНФ); 2) когда «приемлемы» низкие значения индикаторной характеристики и «допустимы» высокие значения фактора (метод находит верхнюю ГНИ и нижнюю ГНФ); 3) когда «приемлемы» низкие значения индикаторной характеристики и «допустимы» низкие значения фактора (метод находит верхнюю ГНИ и верхнюю ГНФ). Во всех этих случаях меняется положение областей на рисунке так, чтобы область «*a*» соответствовала «приемлемым» значениям индикатора при «допустимых» значениях фактора, область «*b*» — «приемлемым» значениям индикатора при «недопустимых» значениях фактора, область «*c*» — «неприемлемым» значениям индикатора при «допустимых» значениях фактора, область «*d*» — «неприемлемым» значениям индикатора при «недопустимых» значениях фактора. Формулы расчета критериев для всех перечисленных случаев остаются неизменны.

2.2. Критерий расчета границ

Для оценки степени пустоты области «*b*» в методе ГКК использован *критерий существенности*:

$$\mathrm{C} = \frac{n_{1,1} + n_{2,2}}{n_{1,1} + n_{2,2} + n_{1,2}} - \frac{n_{1,*}n_{*,1} + n_{2,*}n_{*,2}}{n_{1,*}n_{*,1} + n_{2,*}n_{*,2} + n_{1,*}n_{*,2}}, \qquad (1)$$

где $n_{1,*} = n_{1,1} + n_{1,2}$, $n_{2,*} = n_{2,1} + n_{2,2}$ — количества наблюдений в классах 1 и 2 по индикатору; $n_{*,1} = n_{1,1} + n_{2,1}$, $n_{*,2} = n_{1,2} + n_{2,2}$ — количества наблюдений в классах 1 и 2 по фактору, $n_{*,p}$ — для экономии места в работе использованы обозначения сумм наблюдений в областях согласно [Миркин, 1980]:: $n_{*,p} = \sum_{m=1}^{2} n_{m,p}$; $n_{m,*} = \sum_{p=1}^{2} n_{m,p}$. Существенность характеризует приращение доли правильных предсказаний одной характеристики, полученное за счет использования информации о значении другой [Миркин, 1980].

Первая часть формулы существенности — критерий точности Чеснокова [1982], отвечает за «субъективную» пустоту области «*b*» относительно близлежащих областей «*a*» и «*d*»:

$$\mathrm{T} = \frac{n_{1,1} + n_{2,2}}{n_{1,1} + n_{2,2} + n_{1,2}}. \qquad (2)$$

Вторая часть формулы,

$$\frac{n_{1,*}n_{*,1} + n_{2,*}n_{*,2}}{n_{1,*}n_{*,1} + n_{2,*}n_{*,2} + n_{1,*}n_{*,2}},$$

отражает пустоту области «*b*» относительно областей «*a*» и «*d*», обусловленную соотношением частот обеих характеристик, т. е. отвечает за «объективную» пустоту области «*b*» относительно близлежащих областей «*a*» и «*d*», обусловленную асимметрией частот сопоставляемых характеристик. Здесь использованы расчетные количества наблюдений в областях, полученные путем умножения количества наблюдений в классе по индикатору на количество наблюдений в классе по фактору и деления на общее количество наблюдений (общее количество наблюдений в числителе и знаменателе сокращено).

Отметим, что в ряде предшествующих работ [Левич, Булгаков, Максимов, 2004; Левич, Булгаков, Рисник, 2010; Левич и др., 2011; Левич и др., 2009; Левич и др., 2010] для анализа взаимосвязи между индикаторами и факторами в рамках метода ЛЭН был использован именно критерий точности Чеснокова. Однако с целью учета вклада в установленную связь соотношения частот обеих характеристик он заменен на критерий существенности.

Приведем иллюстрацию необходимости использования критерия существенности, а не критерия точности для нахождения связи. Для простоты возьмем случай с двумя качественными классами и заведомо известными границами между этими классами. Проанализируем связь между отношением детей к фруктам («любят», «не любят») и полом ребенка («мальчик», «девочка»), исходя из данных, представленных в таблице 2. Из таблицы на основании критерия точности

$$\mathrm{T} = \frac{n_{1,1} + n_{2,2}}{n_{1,1} + n_{2,2} + n_{1,2}} = \frac{40 + 40}{40 + 40 + 10} = 0,89$$

можно сделать вывод, что если ребенок «не любит фрукты» (первая строка таблицы), то он, скорее всего, женского пола, в то же время если ребенок мужского пола (второй столбец таблицы), то он, скорее всего, «любит фрукты». Первое утверждение несостоятельно, так как оно не учитывает соотношения распределения опрошенных детей по полам, т. е. не учитывает, что было опрошено 200 девочек и всего 50 мальчиков.

Таблица 2. Гипотетический пример. Связь пола ребенка с отношением детей к фруктам

	Девочки	Мальчики	Количество наблюдений в классах отношения детей к фруктам
Не любят фрукты	40	10	50
Любят фрукты	160	40	200
Количество наблюдений в классах детей по полу	200	50	250

Рассчитав критерий существенности, учитывающий влияние соотношения частот характеристик, для этого случая получим

$$\mathrm{C} = \frac{40 + 40}{40 + 40 + 10} - \frac{50 \cdot 200 + 200 \cdot 50}{50 \cdot 200 + 200 \cdot 50 + 50 \cdot 50} = 0,89 - 0,89 = 0,$$

т. е. связи между полом ребенка и его отношением к фруктам в данном случае нет.

Аналогичная ситуация возникает при расчете границ норм, когда область «*b*» сравнительно пуста в сравнении с областями «*a*» и «*d*» за счет влияния соотношения частот индикатора и фактора при отсутствии связи между ними, чем и обусловлено применение критерия существенности в качестве критерия расчета границ.

2.3. Критерии проверки границ

В рамках рассматриваемого простейшего случая проиллюстрируем необходимость введения критериев проверки границ. Как уже было отмечено, критерий существенности может быть максимален за счет нескольких причин, в частности связанных с алгоритмом поиска границ норм, а не с наличием связи. Так, например, пустота области «*b*» для рассматриваемого на рис. 1 случая может быть обусловлена проведением границы нормы индикатора вблизи максимальных значений индикатора и/или проведением границы нормы фактора вблизи максимальных значений фактора. Тогда при незначительном числе наблюдений в областях «*a*» и «*d*» и пустоте области «*b*» критерий точности достигает максимальных значений. За счет этого и критерий существенности может достигать значений бóльших, чем при значительном числе наблюдений в областях «*a*» и «*d*» и незначительном числе наблюдений в области «*b*». Таким образом, для корректного проведения расчета границ норм и получения значимых результатов необходимо введение *критериев проверки границ*.

1. *Область «b» должна быть достаточно пуста в сравнении с областями «a» и «d».* Эту пустоту, характеризует критерий точности (см. формулу (2)): чем более пуста область «*b*», тем больше значение критерия. Чтобы утверждать, что область «*b*» достаточно пуста в сравнении с областями «*a*» и «*d*», критерий точности должен быть выше *параметра минимальной точности* ($\mathrm{T}_{\text{мин}}$), заданного исследователем: $\mathrm{T} > \mathrm{T}_{\text{мин}}$. В расчетах, как правило, используют значения параметра $\mathrm{T}_{\text{мин}}$ от 0,85 до 0,95.
2. *Каждая из областей «a» и «d» должна содержать представительное количество точек.* Выбор границы нормы индикатора вблизи максимальных значений индикатора и/или границы нормы фактора вблизи максимальных значений фактора приведет к тому, что область «*b*» окажется пустой, однако тогда и области «*a*» и/или «*d*» также окажутся пустыми. Чтобы удостовериться в наличии точек в областях «*a*» и «*d*», вводится *параметр минимальной представительности* ($\mathrm{ПР}_{\text{мин}}$) областей «*a*» и «*d*», выше которого относительное количество наблюдений в областях «*a*» и «*d*» таково, что эти области можно считать непустыми:

$$n_{1,1}/N > \mathrm{ПР}_{\text{мин}},\ n_{2,2}/N > \mathrm{ПР}_{\text{мин}},$$

   где N — общее количество совместных наблюдений индикатора и фактора.
   В расчетах, как правило, используют значения параметра минимальной представительности от 0,1 до 0,3.
3. *Необходимо наличие достаточного для анализа количества совместных наблюдений индикатора и фактора.* Алгоритм принципиально реализуем при числе совместных наблюдений выше экспертного минимума:

$$N > N_{\text{мин}},$$

   где $N_{\text{мин}}$ — *параметр минимального количества совместных наблюдений*.
   В расчетах, как правило, используют значения параметра $N_{\text{мин}}$ от 30 до 80.
4. *Результаты расчетов должны быть значимы в статистическом смысле. Значимость* (*доверительную вероятность*) полученных *результатов* оценивают как вероятность того, что при независимости распределений двух характеристик, между которыми проводится поиск связи, не будут найдены границы норм при заданных параметрах минимальной точности и представительности. Анализ значимости проводят для каждой пары «индикатор–фактор»,

исходя из результирующих (соответствующих итоговому положению границ нормы) значений критерия точности ($\mathrm{T}_{\text{рез}} = \dfrac{n_{1,1} + n_{2,2}}{n_{1,1} + n_{2,2} + n_{1,2}}$), минимального из критериев представительности $\mathrm{ПР}_{\text{рез}} = \min(n_{1,1}/N; n_{2,2}/N)$ и количества совместных наблюдений ($N_{\text{рез}}$). Для анализируемой пары задание в качестве параметров поиска $\mathrm{T}_{\text{мин}} = \mathrm{T}_{\text{рез}}$, $\mathrm{ПР}_{\text{мин}} = \mathrm{ПР}_{\text{рез}}$, $N_{\text{мин}} = N_{\text{рез}}$ не приведет к изменению результата анализа (т. е. будут найдены те же границы классов качества при той же величине критерия существенности). С учетом этого для расчета доверительной вероятности генерируют 2000 наборов равномерно распределенных случайных чисел, в каждом из наборов $N_{\text{рез}}$ значений. Из этих наборов составляют 1000 заведомо независимых пар «индикатор–фактор». Для полученных 1000 пар проводят поиск границ качественных классов, задавая следующие критерии поиска: $\mathrm{T}_{\text{мин}} = \mathrm{T}_{\text{рез}}$ и $\mathrm{ПР}_{\text{мин}} = \mathrm{ПР}_{\text{рез}}$. Доверительную вероятность (ДВ) определяют по формуле $\mathrm{ДВ} = \left(1 - \dfrac{M_{\text{найд}}}{1000}\right) \cdot 100\,\%$, где $M_{\text{найд}}$ — число пар «индикатор–фактор» с найденными границами классов качества, удовлетворяющими заданным критериям минимальной точности и представительности. Критерий существенности исключен из расчета доверительной вероятности ввиду громоздкости учета в величине ДВ зависимости от четырех показателей (параметр минимальной точности, параметр минимальной представительности, количество совместных наблюдений, критерий существенности).

2.4. Полноты факторов для найденных границ

Полученные границы норм индикатора и фактора и соответствующие этим границам количества наблюдений в областях диаграммы на рис. 1 позволяют получить дополнительную информацию, характеризующую влияние факторов на индикатор как в отношении пар «индикатор–фактор», так и в отношении индикатора в целом. А именно, информацию о *полноте вклада факторов в степень «неприемлемости» значений индикаторной характеристики* и о *достаточности программы наблюдений за факторами* для отражения причин «неприемлемости» значений исследуемой индикаторной характеристики. Эта информация в алгоритме метода ГКК охарактеризована полнотами факторов для найденных границ.

1. *Полнота фактора* характеризует вклад каждого из исследуемых факторов в степень «неприемлемости» значений индикаторной характеристики как отношение количества наблюдений, «недопустимых» по фактору и «неприемлемых» по индикатору, к общему количеству «неприемлемых» значений по индикатору (при любых значениях всех факторов):

$$\Pi = n_{2,2} \big/ N^{-},$$

где N^{-} — количество наблюдений, «неприемлемых» по индикатору, при любых значениях всех факторов ($N^{-} \geq n_{2,1} + n_{2,2}$, так как достаточно часто в анализируемой предыстории существуют «неприемлемые» наблюдения индикатора, при которых не измеряли значения фактора).
2. *Совместные полноты факторов* позволяют учесть «неприемлемость» значений индикатора, обусловленную «недопустимостью» нескольких факторов среды.
   Совместные полноты факторов отражают отношение количества «неприемлемых» наблюдений по индикатору, обусловленных одновременной «недопустимостью» значений по группе факторов, к общему количеству «неприемлемых» наблюдений по индикатору (при любых значениях всех факторов). Дальнейшую роль фактора из группы влияющих совместно факторов сможет определить исследователь из содержательных, а не формальных соображений.
3. При анализе массива данных наблюдений может возникнуть ситуация, когда значения индикатора «неприемлемы», а значения всех факторов при этом принимают «допустимые» значе-

ния. Это означает, что какой-то важный фактор, оказывающий негативное влияние на исследуемую индикаторную характеристику, не учтен, т. е. программа наблюдений за факторами, влияющими на индикаторную характеристику, недостаточна. Рассчитанные границы норм индикатора и фактора позволяют определить полноту всех факторов (*достаточность программы наблюдений за факторами*) для исследуемого индикатора. Она отражает долю среди всех «неприемлемых» значений индикатора (при любых значениях всех факторов) таких значений, «неприемлемость» которых обусловлена «недопустимостью» значений хотя бы одного из исследуемых в программе наблюдений факторов.

3. Одновременный поиск одной границы нормы индикатора вместе с верхней и нижней границами нормы факторов для двух классов качества

3.1. Термины и обозначения

Рассмотрим случай поиска двух границ нормы фактора, когда «неприемлемы» низкие значения индикаторной характеристики и к «неприемлемым» значениям приводят как «низкие», так и «высокие» значения фактора, допустимыми же являются «средние» значения фактора. Случай рассмотрен на примере влияния содержания ионов кальция в водных объектах бассейна Нижней Волги на значения индекса Шеннона (рис. 2).

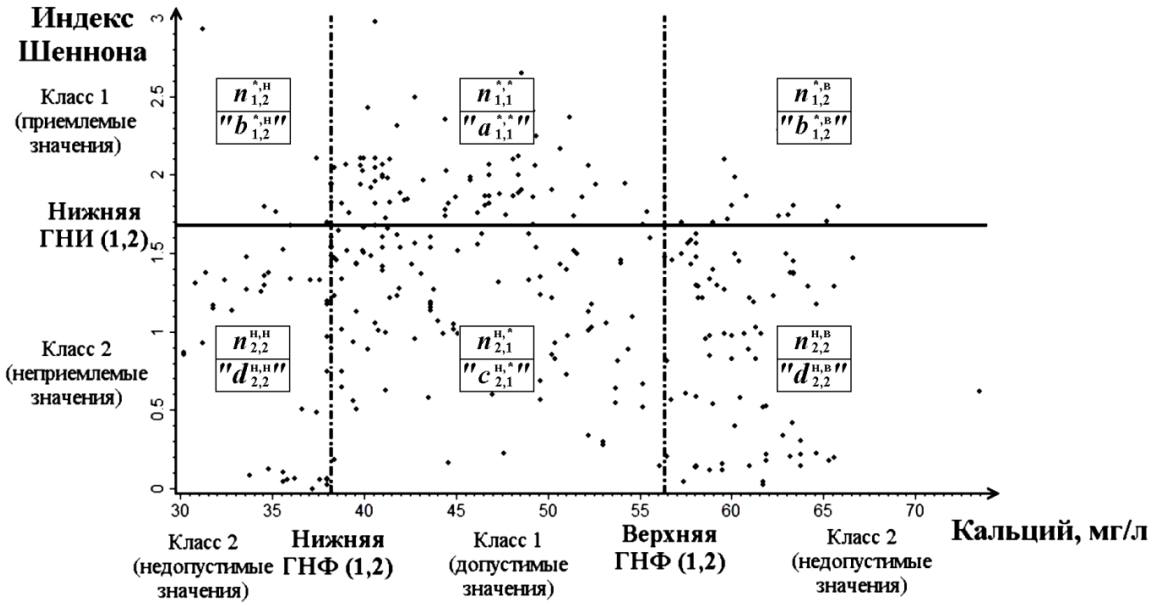

Рис. 2. Зависимость значений индекса Шеннона от концентрации кальция в водных объектах бассейна Нижней Волги (нижняя граница нормы индикатора вместе с верхней и нижней границами фактора для двух классов качества). Обозначения: ГНИ — граница нормы индикатора, ГНФ — граница нормы фактора, в скобках указано, между какими классами качества проведены границы, «*a*», «*b*», «*c*», «*d*» — обозначения областей диаграммы; $n_{m,p}^{k,l}$ — количество наблюдений в области, относящейся к классу m по индикатору, к классу p по фактору, находящейся выше области приемлемых значений по индикатору при k = «в», находящейся ниже области приемлемых значений по индикатору при k = «н», находящейся выше области допустимых значений по фактору при l = «в», находящейся ниже области допустимых значений по фактору при l = «н»; $n_{m,p}^{*,l}$ — для экономии места в работе использованы обозначения сумм наблюдений в областях согласно [Миркин, 1980]: $n_{m,p}^{*,l} = \sum_{k\in\{\text{н,в}\}} n_{m,p}^{k,l}$, $n_{m,p}^{k,*} = \sum_{l\in\{\text{н,в}\}} n_{m,p}^{k,l}$, $n_{*,p}^{k,l} = \sum_{m=1}^{w} n_{m,p}^{k,l}$, $n_{m,*}^{k,l} = \sum_{p=1}^{w} n_{m,p}^{k,l}$.

3.2. Критерий расчета границ

В формулах критериев точности и существенности, для поиска одной границы нормы индикатора и одной границы нормы фактора (формулы (1) и (2)), происходит замена количества наблюдений в областях «*b*» ($n_{1,2}$) и «*d*» ($n_{2,2}$) на суммы чисел наблюдений в однотипных областях (т. е. областях, относящихся к одному классу качества по индикатору и одному классу качества по фактору, например «$b_{1,2}^{*,\text{н}}$» и «$b_{1,2}^{*,\text{в}}$», а также «$d_{2,2}^{\text{н,н}}$» и «$d_{2,2}^{\text{н,в}}$»). Преобразования критериев расчета и проверки границ исходят из того, что классы «недопустимых» значений фактора расположены симметрично относительно класса его «допустимых» значений, причем на одну область «*a*» приходится две однотипные области «*b*», две однотипные области «*d*» и одна область «*c*». Изменение требований к областям связано с тем, что областей каждого типа («*a*», «*b*», «*c*» и «*d*») при поиске верхней и нижней границ становится две (см. таблицу 3), но ввиду соседства некоторых однотипных областей эти области невозможно разделить, они сливаются в более крупную область, требования к которой возрастают пропорционально количеству неразделимых однотипных областей, из которых она состоит.

Из таблицы видно, что неразделимы две области типа «*a*» («$a_{1,1}^{*,*} = a_{1,1}^{\text{н,н}} + a_{1,1}^{\text{н,в}}$»), и две области «*c*» («$c_{2,1}^{\text{н},*} = c_{2,1}^{\text{н,н}} + c_{2,1}^{\text{н,в}}$»).

Таблица 3. Обозначения количества наблюдений для различных сочетаний двух классов качества, пунктиром разделены соседние области одного типа (т. е. относящиеся к одному классу качества по индикатору и одному классу качества по фактору). Значения чисел наблюдений приведены для зависимости индекса Шеннона от концентрации кальция в водных объектах бассейна Нижней Волги

<table>
<tr><td rowspan="3">Класс качества по индикатору</td><td colspan="4">Класс качества по фактору</td><td rowspan="3">Количество наблюдений в классах по индикатору</td></tr>
<tr><td>2 (недопустимость)</td><td>1 (допустимость)</td><td>1 (допустимость)</td><td>2 (недопустимость)</td></tr>
<tr><td colspan="4">(в кавычках обозначение области)</td></tr>
<tr><td>1
(приемлемость)</td><td>$n_{1,2}^{\text{н,н}} = 10$
«$b_{1,2}^{\text{н,н}}$»</td><td>$n_{1,1}^{\text{н,н}} = 37$
«$a_{1,1}^{\text{н,н}}$»
$n_{1,1}^{\text{н},*} = 74$</td><td>$n_{1,1}^{\text{н,в}} = 37$
«$a_{1,1}^{\text{н,в}}$»</td><td>$n_{1,2}^{\text{н,в}} = 16$
«$b_{1,2}^{\text{н,в}}$»</td><td>$n_{1,*}^{\text{н},*} = 100$</td></tr>
<tr><td>2
(неприемлемость)</td><td>$n_{2,2}^{\text{н,н}} = 54$
«$d_{2,2}^{\text{н,н}}$»</td><td>$n_{2,1}^{\text{н,н}} = 52$
«$c_{2,1}^{\text{н,н}}$»
$n_{2,1}^{\text{н},*} = 104$</td><td>$n_{2,1}^{\text{н,в}} = 52$
«$c_{2,1}^{\text{н,в}}$»</td><td>$n_{2,2}^{\text{н,в}} = 80$
«$d_{2,2}^{\text{н,в}}$»</td><td>$n_{2,*}^{\text{н},*} = 238$</td></tr>
<tr><td>Количество наблюдений в классах по фактору</td><td>$n_{*,2}^{\text{н,н}} = 64$</td><td>$n_{*,1}^{\text{н,н}} = 89$
$n_{*,1}^{\text{н},*} = 178$</td><td>$n_{*,1}^{\text{н,в}} = 89$</td><td>$n_{*,2}^{\text{н,в}} = 96$</td><td>$n_{*,1}^{\text{н,н}} + n_{*,1}^{\text{н,в}} + n_{*,2}^{\text{н,н}} + n_{*,2}^{\text{н,в}}$
$= n_{1,*}^{\text{н},*} + + n_{2,*}^{\text{н},*} = 338$</td></tr>
</table>

Формула расчета критерия существенности приобретает вид

$$\text{C} = \frac{n_{1,1}^{\text{н},*} + n_{2,2}^{\text{н},*}}{n_{1,1}^{\text{н},*} + n_{2,2}^{\text{н},*} + n_{1,2}^{\text{н},*}} - \frac{n_{1,*}^{\text{н},*} n_{*,1}^{\text{н},*} + n_{2,*}^{\text{н},*} n_{*,2}^{\text{н},*}}{n_{1,*}^{\text{н},*} n_{*,1}^{\text{н},*} + n_{2,*}^{\text{н},*} n_{*,2}^{\text{н},*} + n_{1,*}^{\text{н},*} n_{*,2}^{\text{н},*}} =$$

$$= \frac{74+134}{74+134+26} - \frac{100\cdot 178 + 238\cdot(64+96)}{100\cdot 178 + 238\cdot(64+96) + 100\cdot(64+96)} = 0,89 - 0,78 = 0,11 .$$

Формула расчета критерия точности, соответственно,

$$\text{T} = \frac{n_{1,1}^{\text{н},*} + n_{2,2}^{\text{н},*}}{n_{1,1}^{\text{н},*} + n_{2,2}^{\text{н},*} + n_{1,2}^{\text{н},*}} = 0,89 .$$

В результате максимизации нижняя ГНИ для индекса Шеннона составила 1,68, нижняя граница для содержания кальция — 38,1 мг/л, верхняя — 56,3 мг/л.

3.3. Критерии проверки границ

1. Критерий проверки того, что *область* «*b*» *достаточно пуста в сравнении с областями «a» и «d»* ($\mathrm{T} > \mathrm{T}_{\text{мин}}$), преобразуется в два критерия:

$$\frac{0,5\cdot n_{1,1}^{\text{н},*} + n_{2,2}^{\text{н,н}}}{0,5\cdot n_{1,1}^{\text{н},*} + n_{2,2}^{\text{н,н}} + n_{1,2}^{\text{н,н}}} = \frac{0,5\cdot 74 + 54}{0,5\cdot 74 + 54 + 10} = 0,90 > \mathrm{T}_{\text{мин}};$$

$$\frac{0,5\cdot n_{1,1}^{\text{н},*} + n_{2,2}^{\text{н,в}}}{0,5\cdot n_{1,1}^{\text{н},*} + n_{2,2}^{\text{н,в}} + n_{1,2}^{\text{н,в}}} = \frac{0,5\cdot 74 + 80}{0,5\cdot 74 + 80 + 16} = 0,88 > \mathrm{T}_{\text{мин}},$$

где $\mathrm{T}_{\text{мин}}$ задано равным 0,85, т. е. в требование вида

$$\frac{0,5\cdot n_{1,1}^{\text{н},*} + n_{2,2}^{\text{н},l}}{0,5\cdot n_{1,1}^{\text{н},*} + n_{2,2}^{\text{н},l} + n_{1,2}^{\text{н},l}} > \mathrm{T}_{\text{мин}} \text{ при каждом } l \in \{\text{н},\text{в}\}.$$

Критерий преобразован ввиду наличия двух однотипных областей «*b*», требования к пустоте каждой из которых должны быть выполнены независимо друг от друга.

2. Проверка того, что *каждая из областей «a» и «d» должна содержать представительное количество точек.* Формально требование к количеству наблюдений в областях «*a*» и «*d*» должно быть изменено аналогично другим критериям, т. е. области определенного типа заменены на суммы областей определенного типа ($n_{1,1}/N > \mathrm{ПР}_{\text{мин}}$; $n_{2,2}/N > \mathrm{ПР}_{\text{мин}}$ заменено на $n_{1,1}^{\text{н},*}/N = 74/338 = 0,22 > \mathrm{ПР}_{\text{мин}}$; $n_{2,2}^{\text{н},*}/N = 134/338 = 0,40 > \mathrm{ПР}_{\text{мин}}$, где $\mathrm{ПР}_{\text{мин}}$ задана равной 0,20). Для двух областей типа «*a*» ввиду их неразделимости применимо только приведенное выше требование. Для двух областей типа «*d*» необходима проверка того, что каждая из областей «*d*» не пуста. Это обусловлено тем, что суммарная доля наблюдений может превышать минимальную представительность за счет высокого количества наблюдений в одной из областей типа «*d*» при пустоте второй области этого типа. В таком случае исследуемое распределение будет описано неверно. Таким образом, для проверки непустоты каждой из областей типа «*d*» ввиду их равноправности требование к представительности изменено на требования

$$n_{2,2}^{\text{н,н}}/N = 54/338 = 0,16 > \mathrm{ПР}_{\text{мин}}/2,\ n_{2,2}^{\text{н,в}}/N = 80/338 = 0,24 > \mathrm{ПР}_{\text{мин}}/2.$$

Таким образом, для двух классов качества при поиске нижних границ норм по индикатору вместе с верхней и нижней границами норм по фактору должны быть выполнены два вида требований:

$$n_{1,1}^{*,*}/N > \mathrm{ПР}_{\text{мин}} \text{ и } n_{2,2}^{\text{н},l}/N > \mathrm{ПР}_{\text{мин}}/2 \text{ при каждом } l \in \{\text{н},\text{в}\}.$$

3. *Критерий проверки количества совместных наблюдений* индикатора и фактора должен быть преобразован исходя из того, что требование к наполненности каждой из областей «*a*» и «*d*» в абсолютных единицах при поиске верхней и нижней границ качественных классов должно соответствовать аналогичному требованию для двух классов качества.
Таким образом, исходя из требования к минимальному количеству совместных наблюдений для двух классов качества при расчете односторонних границ и требований к минимальной представительности каждой из четырех областей «*d*», получаем равенство

$$N_{\text{мин}}(2\text{ кл.}, 1\text{ гр.})\cdot \mathrm{ПР}_{\text{мин}} = N_{\text{мин}}(2\text{ кл.}, 2\text{ гр. факт.})\cdot \mathrm{ПР}_{\text{мин}}/2,$$

где $N_{\text{мин}}(2\text{ кл.}, 1\text{ гр.})$ — минимальное количество совместных наблюдений при расчете односторонних границ для двух классов качества, $\mathrm{ПР}_{\text{мин}}$ — минимальная представительность об-

ласти «*d*» при этом расчете; $N_{\text{мин}}\left(2\text{ кл., }2\text{ гр. факт.}\right)$ — минимальное количество совместных наблюдений при расчете одной границы нормы для индикатора вместе с верхними и нижними границами для двух классов качества; $\text{ПР}_{\text{мин}}/4$ — минимальная представительность области «*d*» при этом расчете.

Если, например, $N_{\text{мин}}\left(2\text{ кл., }1\text{ гр.}\right)$ задано равным 50, получаем

$$N_{\text{мин}}\left(2\text{ кл., }2\text{ гр. факт.}\right)=\frac{N_{\text{мин}}\left(2\text{ кл., }1\text{ гр.}\right)\cdot\text{ПР}_{\text{мин}}}{\text{ПР}_{\text{мин}}/2}=2\cdot N_{\text{мин}}\left(2\text{ кл., }1\text{ гр.}\right)=100.$$

Количество совместных наблюдений должно быть выше соответствующего параметра минимального количества совместных наблюдений:

$$N > N_{\text{мин}}\left(2\text{ кл., }2\text{ гр. факт.}\right).$$

4. *Доверительную вероятность результатов* определяют как вероятность того, что при независимости распределений двух характеристик, между которыми проводится поиск связи, не будут найдены границы норм при заданных параметрах минимальной точности и представительности. Доверительная вероятность результатов для пары «индекс Шеннона–кальций» составила 96,5 %.

3.4. Полноты факторов для найденных границ

В целом все полноты найденных границ изменяются аналогично критериям расчета и проверки границ, т. е. происходит замена количества наблюдений в области типа «*d*» на количество наблюдений в сумме областей этого типа. Общее количество «неприемлемых» значений индикатора получают, суммируя количество «неприемлемых» значений выше верхней границы нормы индикатора и ниже нижней границы нормы индикатора.

1. *Полнота фактора*. Отношение количества наблюдений, «недопустимых» по фактору и «неприемлемых» по индикатору, к общему количеству наблюдений, «неприемлемых» по индикатору (при любых значениях всех факторов):

$$\text{П}=n_{2,2}^{\text{н},*}\Big/N^{-}=134/429=0,31,$$

где N^{-} — количество наблюдений, «неприемлемых» по индикатору при любых значениях всех факторов. Стоит обратить внимание, что в данном случае N^{-} даже больше, чем N, так как всего в бассейне Нижней Волги было 612 наблюдений за значениями индекса Шеннона (из них 429 соответствуют «неприемлемым») и 338 наблюдений за содержанием кальция (из них 134 соответствуют «недопустимым»).

2. *Совместные полноты факторов*. Отношение количества «неприемлемых» наблюдений по индикатору, обусловленных одновременной «недопустимостью» значений по группе факторов, к общему количеству всех «неприемлемых» наблюдений по индикатору (при любых значениях всех факторов).
3. *Достаточность программы наблюдений за факторами.* Доля среди всех «неприемлемых» значений индикатора (при любых значениях всех факторов) таких значений, «неприемлемость» которых обусловлена «недопустимостью» значений хотя бы одного из исследуемых в программе наблюдений факторов. Для индекса Шеннона достаточность программы наблюдений составила 68 %.

Заключение

Разработанные алгоритмы реализованы в лицензионном программном продукте «Программа по установлению границ качественных классов для количественных характеристик систем и установлению взаимосвязи между характеристиками» [Программа по установлению…, 2012].

Описанный в данной работе метод ГКК позволяет решить проблему поиска связи между индикаторными характеристиками и влияющими на них факторами при одновременном действии на индикаторную характеристику множества факторов с преодолением многих возникающих при таком анализе трудностей (см. раздел «Введение»). Метод ГКК позволяет одновременно и взаимосогласованно рассчитывать границы классов качества по индикаторной характеристике и факторам среды, т. е. построить такие классификаторы, в которых переход значения фактора в более неблагоприятный класс приводит к аналогичному переходу для значения индикатора.

Поясним необходимость введения совместных полнот. Часто возникает ситуация, когда несколько факторов действуют на индикаторную характеристику совместно. В регрессионном анализе (при поиске связей между зависимой характеристикой и набором независимых факторов) для обозначения такой ситуации вводятся понятие «интеркоррелированность» или ее частный случай при тесной линейной связи между факторами — мультиколлинеарность. При наличии интеркоррелированности оценки параметров регрессионной модели и их дисперсии становятся неустойчивыми. В связи с этим при помощи различных алгоритмов проводят устранение интеркоррелированности. Суть этих алгоритмов сводится к исключению из анализа факторов, в наибольшей степени ответственных за интеркоррелированность. Таким образом, может возникать ситуация, при которой фактор объективно влияет на состояние индикатора, но из-за его корреляции с другим фактором он полностью выпадает из результатов анализа. Совместные полноты существуют далеко не для всех комбинаций факторов. Кроме того, при анализе совместных полнот имеет смысл обращать внимание на самые крупные из них. Анализировать совместные полноты менее 0,01 не имеет смысла, так как такие полноты могут быть обусловлены погрешностями в измерении факторов. Так, например, при исследовании влияния факторов среды на значения индекса Шеннона в бассейне Нижней Волги анализировали действие 51 фактора среды. Существенными из них оказались 14, для различных комбинаций негативного действия этих 14 факторов было найдено 91 значение совместных полнот, причем только 28 из этих значений были больше 0,01 и в сумме объясняли более 90 % «неблагополучия», обусловленного этими 14 факторами. Исследование значений факторов среды методами, предшествующими регрессионному анализу, например при помощи расчета «факторов инфляции дисперсии» (VIF, отражающего ортогональность каждого фактора ко всем остальным; чем он выше, тем сильнее линейная связь между этим и остальными факторами), требует рассмотрения 51 значения VIF. В результате анализа факторы с достаточно высокими VIF (как правило, более 5) должны быть исключены из дальнейшего регрессионного анализа. Однако такой подход может, например, привести к тому, что такой важный фактор, как расход воды (объем воды, протекающей через поперечное сечение водотока за единицу времени), будет исключен из анализа, так как характеризует разведение и естественно сильно коррелирует с большинством других факторов (чем выше расход в одной и той же точке наблюдения, тем ниже концентрации всех растворенных в воде веществ). В то же время данный фактор будет найден среди существенных при анализе методом ГКК.

Количество выделяемых классов качества в первую очередь зависит от количества имеющихся совместных наблюдений пары «индикатор–фактор». Если количество наблюдений достаточно велико (более 150), целесообразно проводить поиск начиная с 10 классов, при отсутствии результатов (если существенные факторы не будут найдены) — уменьшать число классов на 1 до тех пор, пока не будут получены результаты.

Перечислим особенности предложенного метода.

– Границы классов качества устанавливают не по экспертным оценкам, а путем расчетов, основанных на строго формальном методе анализа данных.
– Предложенный метод не использует никакие модельные предпосылки. Расчеты границ основаны только на частоте встречаемости тех или других значений характеристик в предыстории наблюдений за объектом.
– Метод не требует, чтобы распределение исходных данных удовлетворяло каким-либо статистическим критериям (например, чтобы они были распределены по нормальному или пуассоновскому законам).

– Метод позволяет рассчитать границу между допустимыми и недопустимыми значениями как для «слишком высоких», так и для «слишком низких» уровней факторов.
– Рассчитанные границы не универсальны, а имеют региональный и даже локальный характер в той степени, в которой рассчитаны по данным региональных или локальных наблюдений. Полученные границы не универсальны не только в географическом пространстве, но и во времени — они могут быть уточнены по мере накопления новых данных или изменений в исследуемом объекте.
– Полученные границы классов по индикатору и фактору взаимосогласованы.

Поиск связи предложенными методами осуществляют формально; метод отвечает на вопрос, какие связи существуют между исследуемым объектом и воздействующими на него факторами, но не дает ответа на вопрос, почему существуют эти связи. Результаты расчета служат основанием для дальнейшего содержательного анализа и интерпретации причин отсутствия или наличия связи.

Список литературы

Левич А. П., Булгаков Н. Г., Максимов В. Н. Insitu-методология оценки качества среды обитания: биоиндикаторы // Доклады по экологическому почвоведению. — 2013. — Вып. 18, № 1. — С. 23–36.

Левич А. П., Булгаков Н. Г., Максимов В. Н. Теоретические и методические основы технологии регионального контроля природной среды по данным экологического мониторинга. — М.: НИА-Природа, 2004. — 271 с.

Левич А. П., Булгаков Н. Г., Максимов В. Н., Рисник Д. В. In situ-технология установления локальных экологических норм // Вопросы экологического нормирования и разработка системы оценки состояния водоемов. — М.: Товарищество научных изданий КМК, 2011. — С. 32–57.

Левич А. П., Булгаков Н. Г., Максимов В. Н., Фурсова П. В. Insitu-методология оценки качества среды обитания: основные положения // Использование и охрана природных ресурсов в России. — 2012. — № 6. — С. 35–37.

Левич А. П., Булгаков Н. Г., Рисник Д. В. Экологический контроль окружающей среды по данным биологического и физико-химического мониторинга природных объектов // Компьютерные исследования и моделирование. — 2010. — Т. 2, № 2. — С. 199–207.

Левич А. П., Булгаков Н. Г., Рисник Д. В., Милько Е. С. Методические проблемы анализа экологических данных и пути их решения: метод локальных экологических норм // Доклады по экологическому почвоведению. — 2013. — Вып. 18, № 1. — С. 9–22.

Левич А. П., Забурдаева Е. А., Максимов В. Н., Булгаков Н. Г., Мамихин С. В. Поиск целевых показателей качества для биоиндикаторов экологического состояния и факторов окружающей среды (на примере водных объектов р. Дон) // Водные ресурсы. — 2009. — Т. 36. № 6. — С. 730–742.

Левич А. П., Рисник Д. В., Булгаков Н. Г., Леонов А. О., Милько Е. С. Методические вопросы применения показателей видового разнообразия фитопланктона для анализа качества вод Нижней Волги // Использование и охрана природных ресурсов. — 2010. — № 5. — С. 44–48; № 6. — С. 33–37.

Миркин Б. Г. Анализ качественных признаков и структур. — М.: Статистика, 1980. — 319 с.

Налимов В. В., Чернова Н. А. Статистические методы планирования экстремальных экспериментов. — М.: Наука, 1965. — 340 с.

Программа по установлению границ качественных классов для количественных характеристик систем и установлению взаимосвязи между характеристиками / *И. А. Гончаров, А. П. Левич, Д. В. Рисник.* Регистрационный номер в Роспатенте 2012616523 РФ (Программа для ЭВМ). — Зарегистрирована 19.07.2012.

Рисник Д. В., Левич А. П., Булгаков Н. Г., Бикбулатов Э. С., Бикбулатова Е. М., Ершов Ю. В., Конюхов И. В., Корнева Л. Г., Лазарева В. И., Литвинов А. С., Максимов В. Н., Мамихин С. В., Осипов В. А., Отюкова Н. Г., Поддубный С. А., Пырина И. Л., Соколова Е. А., Степанова И. Э., Фурсова П. В., Цельмович О. Л. Поиск связей между биологическими и физико-химическими характеристиками экосистемы Рыбинского водохранилища. Ч. 1. Критерии неслучайности связи // Компьютерные исследования и моделирование. — 2013а. — Т. 5, № 1. — С. 83–105.

Рисник Д. В., Левич А. П., Булгаков Н. Г., Бикбулатов Э. С., Бикбулатова Е. М., Ершов Ю. В., Конюхов И. В., Корнева Л. Г., Лазарева В. И., Литвинов А. С., Максимов В. Н., Мамихин С. В., Осипов В. А., Отюкова Н. Г., Поддубный С. А., Пырина И. Л., Соколова Е. А., Степанова И. Э., Фурсова П. В., Цельмович О. Л. Поиск связей между биологическими и физико-химическими характеристиками экосистемы Рыбинского водохранилища. Ч. 2. Детерминационный анализ // Компьютерные исследования и моделирование. — 2013б. — Т. 5, № 2. — С. 271–292.

Рисник Д. В., Рыбка К. Ю. О методе поиска сопряженностей между биологическими и физико-химическими характеристиками для натурных данных на примере экосистемы Рыбинского водохранилища // Материалы Всероссийской научно-практической конференции «Мологский край и Рыбинское водохранилище». — М.: МАКС Пресс, 2011. — С. 169–175.

Чесноков С. В. Детерминационный анализ социально-экономических данных. — М.: Наука, 1982. — 168 с.

Comrey A. L., Lee H. B. A first course in factor analysis (2nd ed.). — Hillsdale, NJ: Erlbaum. — 1992. — 430 p.

Goodman L. A., Kruskal W. H. Measures of association for cross classifications: I–IV // J. Amer. Stat. Assoc. — Vol. 49. — 1954. — P. 723–764; Vol. 54. — 1959. — P. 123–163; Vol. 58. — 1963. — P. 310–364; Vol. 67. — 1972. — P. 322–345.

Green S. B. How many subjects does it take to do a regression analysis? Multivariate Behavioral Research. — Vol. 26. — 1991. — P. 499–510.

Guttman L. An outline of the statistical theory of prediction // The Prediction of Personal Adjustment, Social Science Research Council / Horst P. and others (ed.). — Bulletin 48, New York, 1941. — P. 253–318.

Harris R. J. A primer of multivariate statistics (2nd ed.). — New York: Academic Press. — 1985. — 609 p.