КОНСУЛЬТАЦИИ

Г.Ф. Ромашкина, Г.Г. Татарова
(Тюмень, Москва)

КОЭФФИЦИЕНТ КОНКОРДАЦИИ В АНАЛИЗЕ СОЦИОЛОГИЧЕСКИХ ДАННЫХ

В статье рассматривается коэффициент конкордации как интегральный показатель похожести (схожести) совокупности ранжированных рядов. На конкретных примерах иллюстрируется познавательная возможность коэффициента как достаточно простого приема решения задач разного класса (факторизация переменных, кластеризация объектов, формирование усредненного ранжированного ряда).

Ключевые слова: прикладная статистика, взаимосвязь переменных, схожесть объектов, ранг, связанные ранги, ранжированный ряд, коэффициент связи, коэффициент конкордации.

Понятия близости, похожести, схожести между объектами или переменными являются базовыми в области теории и методов анализа социологических данных. Они лежат в основе постановки и решения таких классов задач, как *анализ взаимосвязей переменных* и *анализ схожести объектов*.

С содержательной точки зрения первый класс задач реализует в эмпирической социологии методологическую процедуру факторизации или факторного (факториального) анализа, а второй – процедуру типологизации или типологического анализа. С формально мате-

Гульнара Фатыховна Ромашкина – доктор социологических наук, профессор кафедры экономической социологии Тюменского государственного университета.
Галина Галеевна Татарова – доктор социологических наук, профессор, главный научный сотрудник Института социологии РАН.

матической позиции упомянутые задачи могут и не различаться и решаться посредством применения одних и тех же методов. Вместе с тем методологическая и практическая целесообразность диктует их различение. Факторному (факториальному) анализу и типологическому анализу как методологическим процедурам соответствуют разные системы языковых конструктов и различные логические схемы концептуализации и соответственно специфические процедуры использования на эмпирическом уровне [1; 2]. Такого рода сюжеты остаются за рамками статьи, ибо в ней делается упор лишь на математическую формализацию, на одно из средств, которым можно воспользоваться при реализации этих процедур на практике.

Анализ связи переменных или объектов посредством *коэффициента конкордации* следует отнести к так называемым «грубым» методам, необходимость применения которых возникает на начальных этапах решения разного рода задач анализа социологических данных. Например, для порождения (формулирования) гипотез о существовании факторной структуры связей в изучаемой системе переменных или гипотез о существовании однотипных групп объектов. Разумеется, такие гипотезы должны проверяться на правдоподобность в процессе дальнейшего анализа более тонкими математическими методами.

Если вопросам применения коэффициента конкордации для анализа экономических данных уделяется достаточное внимание в современных учебниках по эконометрике, например, в [3, с. 423–441; 4, с. 320–331], то этого не наблюдается в учебной литературе по анализу социологических данных. Следует особо подчеркнуть, что в пакете SPSS, активно используемом социологами, предусмотрены определенные возможности вычисления коэффициента конкордации. Вместе с тем, они недостаточны для решения целого ряда содержательных задач, которые могли бы быть реализованы посредством использования коэффициента конкордации.

Учитывая консультативный характер данной статьи и ее предназначенность исследователям, не имеющим математической подготовки, мы исходили из принципа наглядности и прозрачности процедуры применения рассматриваемого коэффициента. В этой связи решения содержательных задач иллюстрируются как в рамках пакета программ или среды Excel, так и SPSS.

Типы задач, решаемые посредством использования коэффициента конкордации

Каждый из выделенных выше двух классов задач внутренне неоднороден, т.е. в них можно выделить различные подклассы.

В первом классе задач – *анализ взаимосвязей переменных* (переменными называем и эмпирические индикаторы, и производные от них показатели) – представляется целесообразным выделение разных подклассов в зависимости, например, от *цели*, ради достижения которой решается задача. В качестве цели могут выступать: проверка гипотезы о существовании факторов в заданном исследователем смысле, поиск факторных синдромов[1] в структуре данных для формирования гипотез о существовании факторов, построение некоторого обобщенного показателя (индекса) и т.д.

Другим основанием для выделения подкласса может являться характер данных, исходных для решения задач «*анализа взаимосвязей переменных*». Это соотносится и со степенью формализованности данных (например, жесткоструктурированные, слабоструктурированные, неструктурированные), и их числовым – нечисловым характером. Возможны и другие основания для

[1] Факторным синдромом называем эмпирическую закономерность, своего рода совокупность симптомов, указывающих на существование фактора. Традиционно на эмпирическом уровне совокупность тесно взаимосвязанных эмпирических индикаторов интерпретируется как фактор. Мы предлагаем в рамках процедур поиска знаний о социальных факторах использовать наряду с термином «фактор» понятие факторного синдрома как адекватного большинству исследовательских ситуаций. Аналогичным образом можно ввести понятия типологического и причинного синдрома [1; 2].

выделения различных подклассов задач. В научной литературе уделяется большое внимание систематизации методических проблем, возникающих в процессе решения задачи «*анализа взаимосвязей переменных*» [5; 6].

Один из подклассов задач «*анализа взаимосвязей переменных*» определяется тем, что исходные данные представлены в виде матрицы, содержащей результаты ранжирования[1]. Далее будем ее называть *матрицей рангов*, в которой строки соответствуют объектам ранжирования, а столбцы – переменным, т.е. разным *основаниям ранжирования*, в клетках матрицы – значение ранга. Разумеется, в ранжированных рядах (столбцах матрицы) могут встречаться и связанные ранги.

Типичная задача «*анализа взаимосвязей переменных*» – поиск факторной структуры системы переменных или другими словами проверка гипотезы о существовании факторной структуры. Каждый ранжированный ряд можно рассматривать как точку в n-мерном пространстве объектов ранжирования. Тогда можно выделить три наиболее характерных типа структуры.

- Случайный разброс точек по всей области возможных значений, что означает отсутствие какой-либо связи между переменными в исследуемом пространстве.
- Расположение точек таково, что часть переменных образует скопление или «ядро» из близко лежащих точек, а остальные произвольно разбросаны вокруг этого ядра.
- Анализируемые упорядочения (ранжированные ряды) располагаются в пространстве несколькими относительно далеко отстоящими друг от друга «ядрами», что соответствует наличию

[1] Учитывая направленность статьи, напомним, что ранжирование – это процедура упорядочения объектов (объектов ранжирования) по возрастанию или убыванию в них некоторого свойства (основание ранжирования) и присвоения объектам ранга. В случае неразличимости объектов им присваиваются одинаковые ранги (связанные ранги, объединенные ранги). Связанные ранги могут быть и дробными. Число рангов равно числу объектов. Например, если четыре объекта претендуют на третье место, то каждому из них присваивается ранг, равный (3 + 4 + 5 + 6)/4 = 4,5.

нескольких групп переменных, имеющих высокую внутреннюю статистическую взаимосвязь и слабо связанных между собой.

Проиллюстрируем, как можно проанализировать эти три характерных типа структуры при помощи коэффициента, позволяющего оценить согласованность упорядочений рангов (далее – коэффициента конкордации). Если предполагается существование одного ядра, решение задачи можно организовать, например, как процедуру постепенного исключения и переменных (по столбцам), и объектов (по строкам) с тем, чтобы повысить согласованность упорядочений. В этом случае необходимо учитывать, что меняется и размерность пространства и длина ранжировок.

В случае предположения о существовании нескольких ядер, коэффициент конкордации будет рассчитываться для различных комбинаций переменных. Для проверки одного случая (комбинации) можно воспользоваться как пакетом SPSS, так и Excel. Поиск факторной структуры, выявление ядер можно организовать на основе простого «перебора» комбинаций переменных и исходя из логики постепенного увеличения значения коэффициента конкордации. Тогда на каждом шаге исключается переменная, дающая максимальный прирост значения коэффициента.

Возникает вопрос: зачем пытаться предварять или даже заменять известные и хорошо представленные, например, в SPSS процедуры типа факторного и кластерного анализа? Ответ на этот вопрос достаточно очевиден для тех, кто многократно пытался использовать эти процедуры для обработки социологических данных. Во-первых, необходимые для этих процедур типы шкал и типы распределений по переменным крайне редко встречаются на практике. Во-вторых, результаты такого анализа имеют иногда вид «слепого поиска» и крайне трудно интерпретируются, что и ограничивает применение этих методов прикладниками-социологами. Тогда как алгоритмы, позволяющие упростить такой поиск на первом этапе и, как следствие, получить пусть приблизительные, но внятно интерпретируемые результаты, бывают весьма полезны.

Во втором классе задач «*анализа схожести объектов*» возникает некий подкласс задач, когда объектами ранжирования выступают переменные, а основанием ранжирования – эмпирические объекты. Эта исследовательская ситуация характерна для случая экспертных опросов и тогда, когда речь идет о ранжировании экспертами переменных по степени их влияния на некий изучаемый социальный феномен (явление, процесс).

Одна из типичных задач, возникающая в рамках такой ситуации, – поиск однотипных экспертов. По сути речь идет о типологических синдромах. Другой типичной задачей «*анализа схожести объектов*» является построение среднего группового упорядочения переменных по всем экспертам или только некоторой однотипной группе экспертов, т.е. получение групповых оценок. Иными словами – это задача восстановления латентного и своего рода истинного упорядочения. Тогда правдоподобность гипотезы о согласованности мнений экспертов можно проверить с помощью коэффициента конкордации. Если это подтвердится, то для получения «среднего» ранжированного ряда можно воспользоваться и медианой, и средним арифметическим.

Таким образом, в любой из описанных задач возникает необходимость оценки согласованности упорядочений (ранжировок) в совокупности ранжированных рядов. Для оценки такой согласованности и применяется коэффициент согласия (обозначим его W). В литературе по прикладной статистике его называют коэффициентом конкордации (*лат.* Concordia – *согласие*) или коэффициентом конкордации Кендалла.

Коэффициент конкордации вводится для того, чтобы его значения менялись от 0 до 1. Если $W = 0$, то ранжировки в рядах считаются несогласованными (непохожими, несхожими). Если $W = 1$, то ряды являются согласованными. При этом они [упорядочения, ранжировки] могут полностью совпадать, и тогда по столбцам *матрицы рангов* наблюдается равномерное распределение, но обратное неверно [3, с. 438]. Распределение может быть равномерным в рядах, но ряды могут и не совпадать.

Специфика коэффициента конкордации состоит в том, что при его введении не накладываются ограничения на характер распределения рангов в строках *матрицы рангов*, например, в виде необходимости нормального распределения и линейности связи [7, с. 434].

Математическая специфика коэффициента конкордации[1]

Введение коэффициента конкордации в научный оборот (математические выкладки и обоснования) обычно приписывается Дж. Кендаллу. Вместе с тем, в литературе зачастую отсутствуют соответствующие ссылки, либо предлагаются ссылки на другие источники [3, с. 437; 4, с. 325; 8, с. 433]. Мы будем опираться на математические выкладки, предложенные в работе [7, с. 29; 9, с. 101]. Кратко приведем их.

Допустим, в общем случае, что имеются p переменных $\{x_1, \ldots, x_p\}$ и n объектов ранжирования. Матрицу рангов обозначим в виде $\{r^{(1)}, \ldots, r^{(p)}\}$. Предположим, что анализируются m переменных с номерами $k_1, k_2, \ldots, k_m$.

Тогда для *матрицы рангов* вводятся понятия ***среднего ранга*** (a) и ***вариации*** (S) относительно этого среднего.

$$a = \frac{1}{2} m \times (n+1) ;$$

$$S = \sum_{i=1}^{n} \left(\sum_{j=1}^{m} r_i^{k_j} - a\right)^2 . \quad (1)$$

Очевидно, что при полной согласованности ранжировок $S \to \max$, а при полной несогласованности $S \to \min$. При полной

[1] Исследователь-прикладник может пропустить в процессе чтения математические выкладки этого раздела статьи.

несогласованности ранжировок сумма в (1) будет равна a, если m четное, и максимально близка к a при нечетном m. При малых значениях m и n ($2 \le m \le 20$, $3 \le n \le 7$) степень согласованности можно проверять по математическим таблицам, которые можно найти в справочниках по математической статистике.

Коэффициент конкордации вводится как отношение «реального» к «идеальному». На этот принцип опирается целый ряд математических конструктов (формул). Под «реальным» понимается значение вариации в *матрице рангов*, а под идеальным – возможно максимальное значение вариации (полная согласованность ранжировок). Нетрудно доказать [9, с. 101], что максимальное значение равно:

$$S\max = \frac{1}{12} m^2 (n^3 - n). \tag{2}$$

Отметим, что наличие связанных рангов приводит к появлению дробных значений.

Коэффициент конкордации при отсутствии связанных рангов находится по достаточно простой формуле:

$$W = \frac{S}{1/12 * m^2 (n^3 - n)}. \tag{3}$$

Известно, что величина $m(n-1)W$ (для $n > 7$) имеет χ^2 распределение [7, с. 115; 8, с. 441] с числом степеней свободы $f = n - 1$.

Тогда, если окажется, что

$$\chi^2{}_p = m(n-1)W > \chi^2_T, \tag{4}$$

можно сделать вывод о том, что ранжировки согласованы при заданном уровне значимости. Даже малое значение коэффициента позволяет сделать вывод о статистически значимой связи между переменными.

Необходимо отметить, что использование критерия χ^2 имеет существенное неудобство, связанное с тем, что его верхняя граница стремится к бесконечности при возрастании объема выборки n. С учетом этого факта С.А. Айвазян [3, с. 445] предлагает вместо критерия Пирсона проводить проверку статистической значимости исследуемой связи (между ранжированными рядами) на основе критерия Фишера. Утверждается, что в условиях отсутствия связи распределение случайной величины

$$\lambda_F = \frac{1}{2} \ln \frac{(m-1)W}{1-W} \qquad (5)$$

приближенно описывается распределением Фишера с числом степеней свободы числителя $\nu_1 = n - 1 - 2/m$ и знаменателя $\nu_2 = (m-1)\nu_1$.

В этой же работе утверждается, что строгих рекомендаций по построению доверительных интервалов для истинного значения *W* в условиях наличия ранговых связей (связей переменных) к настоящему времени не имеется.

Коэффициент конкордации является многомерным аналогом коэффициента ранговой корреляции Спирмена (обозначим его как $\rho_{k_1k_2}$), в частности, при $m = 2$ он пропорционален коэффициенту Спирмена [3, с. 438], т.е.

$$W = \frac{1}{2}(\rho_{k_1k_2} + 1). \qquad (6)$$

Но это верно только для случая отсутствия связанных рангов, что объясняется природой коэффициентов, поскольку обе характеристики являются линейной функцией от числа так называемых инверсий (нарушений монотонности ранговых упорядочений), имеющихся в сравнении последовательностей рангов. Коэффициент конкордации не принимает отрицательные значения, и это объясняется следующим обстоятельством. В отличие от случая парных связей, противоположные понятия «согласованности и несогласованности» при *m* больше двух утрачивают симметричность (относительно нуля), так как ранжированные ряды могут полностью совпадать, но *полностью не совпадать* не могут.

Вычисление коэффициента конкордации в ситуации наличия связанных рангов производится по более сложной формуле:

$$W = \frac{S}{\frac{1}{12} * m^2(n^3 - n) - m\sum_{j=1}^{m} T_j}, \text{ где } T_j = \frac{1}{12}\sum_{\gamma=1}^{l}(t_\gamma^3 - t_\gamma), \qquad (7)$$

T_j – поправочный коэффициент для j-й переменной. Он вычисляется по всем l «случаям» неразличимости объектов. При этом t_γ – число неразличимых объектов одного «случая». Если связанные ранги отсутствуют, то поправочный коэффициент равен нулю.

Согласно приведенным выше формулам (5), (7) при больших значениях m и n статистически значимым будет очень малое по величине значение коэффициента конкордации.

Обсуждение вопроса о поведении выборочного значения коэффициента конкордации при повторении выборок заданного объема n и при отсутствии связи между анализируемыми m переменными можно найти в работе [3, с. 439].

Коэффициент согласия нечасто рассматривается в современных работах по математической статистике, а там, где он приводится, изложение ограничивается формулами, в которых не учитываются связанные ранги, а критерий (коэффициент конкордации можно назвать и критерием согласия) ограничивается требованием стремления W к единице. Тогда как в абсолютном выражении W может оказаться очень малым, но его значение будет статистически значимым для проверки гипотезы о равномерном распределении рангов (согласии ранжировок).

Приведем примеры из литературы, иллюстрирующие неточности в описаниях коэффициента и несколько затрудняющие практическое его использование и, в частности, в социологических исследованиях.

В работе [10, с. 598] утверждается, что выборочная версия коэффициента Кендалла находится по формуле:

$$W = \frac{12}{m^2(n^3 - n)} \sum_{i=1}^{n} \left(\sum_{j=1}^{m} R_i(j) - \frac{m(n+1)}{2} \right)^2, \tag{8}$$

где $R_i(j)$ – ранг i-го наблюдения j-й случайной переменной.

Вместе с тем, легко проверить, что это верно только для случая отсутствия связанных рангов. При этом не приводится критерий для проверки значимости, а, как известно, проверка гипотез о связи переменных посредством W при больших m и n не имеет смысла.

В широко известном учебнике [4] приводится формула для расчета уточненного критерия Пирсона для случая связанных рангов:

$$\chi_P^{\ 2} = \frac{S}{\frac{1}{12} m \times n \times (n-1) - \frac{1}{n-1} \times \sum_{j=1}^{m} T_j}.$$

Нетрудно проверить, что при прямой подстановке в формулу (4) величины (7) получаем несколько иной результат:

$$\chi_P^{\ 2} = \frac{S}{\frac{1}{12} m \times n \times (n+1) - \frac{1}{n-1} \times \sum_{j=1}^{m} T_j}.$$

Отметим, что при *m* и *n,* больших 40, применение упрощенных формул (8) для вычисления коэффициента конкордации без введения поправочных коэффициентов и проверки на статистическую значимость может привести к существенным ошибкам.

Три типичных примера исследовательских ситуаций

Приведем примеры трех типичных исследовательских ситуаций, в которых использование коэффициента конкордации представляется целесообразным. В первом из них объектами ранжирования являются переменные, основание ранжирования формулируется исследователем априори и задача состоит в получении единой (групповой) упорядоченности переменных по степени их влияния на некоторый целевой признак. Пример носит модельный характер.

Аналогичный пример приводится в работе [11, с. 247]. Речь идет об анализе матрицы рангов, где строки соответствуют объектам ранжирования (22 футболиста). Три крупные спортивные газеты оценивали этих футболистов по порядковой шкале от 1 до 6 (к примеру, 1 – «соответствует мировому уровню», 6 – «отработал свои деньги»). Тем самым в трех столбцах матрицы – ранги футболистов. Этот пример отличается от нашего модельного примера тем, что в роли переменных выступают футболисты, но задача одна и та же – получение единой групповой упорядоченности.

Во втором примере объекты ранжирования – респонденты, основание ранжирования – отношение респондентов к некоторым социальным феноменам. Задача состоит в определении влияния на согласованность ответов по ряду вопросов, относящихся к этим феноменам, социально-демографических характеристик респондентов.

Третий пример соотносится с широко распространенным случаем необходимости проверки значимости выводов исследователя, полученных на основании какого-либо метода. Для такого рода проверки может быть использован и коэффициент конкордации.

Пример 1.

В табл. 1 приведена матрица рангов с результатами опроса экспертов. Каждый из них ранжирует переменные по степени их влияния на некоторый целевой признак. Для случаев неразличимости переменных введены связанные ранги. В нашем случае n = 6 (число объектов ранжирования), а m = 7 (число ранжированных рядов).

Таблица 1

МАТРИЦА РАНГОВ

Эксперты (j)	Объекты ранжирования (переменные)					
	r_i^1	r_i^2	r_i^3	r_i^4	r_i^5	r_i^6
1	1,5	5	1,5	3	4	6
2	1	3	2	6	4	5
3	6	5	1	5	2	4
4	2	3	1	4,5	4,5	6
5	1	2	5	6	4	3
6	2	3	1	5,5	5,5	4
7	1,5	3,5	1,5	5	6	3,5
Сумма рангов	15	24,5	13	35	30	31,5
$(\sum_{j=1}^{m} r_i^{k_j} - a)^2$	90,25	0,0	132,25	110,25	30,25	49

Для решения поставленной задачи (получения единого упорядочения или среднего ранжированного ряда) оценим степень согласованности мнений экспертов. Для этого вычислим значение коэффициента конкордации. Средний ранг матрицы и вариация соответственно равны:

$$a = \frac{1}{2} m \times (n+1) = 24{,}5$$

$$S = \sum_{i=1}^{n} \left(\sum_{j=1}^{m} r_i^{k_j} - a \right)^2 = 412 \qquad (1)$$

В исходных ранжированных рядах наблюдаются связанные ранги (см. данные по экспертам с № 1, 4, 6, 7), поэтому необходимо вычислить поправочные коэффициенты. Их значения приведены в табл. 2 для экспертов, у которых наблюдаются связанные ранги.

Таблица 2

ЗНАЧЕНИЯ ПОПРАВОЧНЫХ КОЭФФИЦИЕНТОВ

Эксперты	Поправочные коэффициенты	
	$T_j = t^3 - t$	$T_j/12$
1	2^3-2=6	0,5
4	2^3-2=6	0,5
6	2^3-2=6	0,5
7	(2^3-2)+(2^3-2)=12	1

$$W = \frac{S}{\frac{1}{12} * m^2 (n^3 - n) - m \sum_{j=1}^{m} T_j} = \frac{12 * 415}{7^2 (6^3 - 6) - 7(0{,}5 + 0{,}5 + 0{,}5 + 1)} = 0{,}49.$$

Оценим статистическую значимость этого значения коэффициента конкордации, вычисленного по формуле (7).

Согласно критерию (4) $\chi_p^2 = 17{,}15$. Табличное значение критерия для числа степеней свободы, равного 5, и уровня значимости $\alpha = 0{,}01$ равно $\chi_T^2 = 15{,}09$ (для $\alpha = 0{,}05$ равно $\chi_T^2 = 11{,}07$). Поскольку $\chi_P^2 > \chi_T^2$, гипотеза о согласованности мнений экспертов принимается.

Обоснование способа построения единого варианта упорядочения приводится, например, в [3, с. 427]. Идея алгоритма основана на том, что если принимается гипотеза о согласованности мнений, то, ранжируя суммы рангов по переменным, можно построить единый вариант упорядочения объектов.

В нашем случае для шести переменных X^1, X^2, X^3, X^4, X^5, X^6 вычислим по каждому столбцу матрицы рангов (табл. 1) характеристики распределения оценок переменных экспертами.

Таблица 3

ХАРАКТЕРИСТИКИ РАСПРЕДЕЛЕНИЯ ОЦЕНОК ПЕРЕМЕННЫХ

Характеристики	Переменные					
	X^1	X^2	X^3	X^4	X^5	X^6
Среднее арифметическое	2,14	3,43	1,86	4,93	4,29	4,36
Медиана	1,5	3	1,5	5	4	4
Мода	1,5	3	1	6	4	6
Ранг по группе	2	3	1	6	4	5

Последняя строка в таблице получена ранжированием «суммы рангов» (см. соответствующую строку табл. 1). В нашем примере точность единого варианта упорядочения достаточно высока (99%). Например, ранг переменной $X^1 = 2$. Такой же результат получается как на основе применения медианы, так и среднего арифметического рангов. В определенной степени можно утверждать, что в нашем случае по среднему арифметическому рангов групповая оценка переменной получается с большей точностью. Что касается моды, то ее применять не следует, так как могут появиться случаи неразличимых рангов и случаи, когда мода не существует.

Отметим, что приведенный расчет не слишком сложен и может быть реализован как в Excel, так и в специализированных статистических пакетах. Эти же данные были обработаны нами в SPSS 12.0.1, и получены следующие результаты[1].

[1] Лицензия на пакет SPSS 12.0.1 № 116245, ТюмГУ, г. Тюмень.

Табл. 4 и 5 иллюстрируют соответствующие распечатки пакета. Расчеты в SPSS проводятся без учета связанности рангов, поэтому значения самого коэффициента конкордации и Хи-квадрат отличны от их значений, приведенных выше. На больших выборках возможны значимые ошибки, если не учитывать связанность рангов. Этот факт важен при применении стандартных процедур SPSS. Однако результаты можно уточнить, используя поправки на связанность рангов.

Таблица 4

РАНГИ

	Средний ранг	Относительный ранг
$r1$	2,14	2
$r2$	3,43	3
$r3$	1,86	1
$r4$	4,93	6
$r5$	4,29	4
$r6$	4,36	5

Таблица 5

СТАТИСТИКА КРИТЕРИЯ *W* КЕНДАЛЛА

N (число упорядочений)	7
Статистика *W* Кендалла	0,465
Хи-квадрат	16,276
Число степеней свободы	5
Асимптотическая значимость	0,006

Значимость согласованности упорядочений высока и равна 0,006. Эксперты достаточно единодушны. Тогда для получения единого группового упорядочения можно воспользоваться (см. табл. 4) средним рангом по всем экспертам. В свою очередь, ранжирование этих средних рангов и дает усредненное (относительный ранг) по группе упорядочение. Тем самым получен результат, почти совпадающий с данными табл. 3.

Пример 2.

Приведем фрагмент из исследования, проведенного в 2002–2003 гг. в Тюменской области по вопроснику, разработанному Центром конфликтологии Института социологии РАН (рук. проекта Е.И. Степанов). Объем выборки (1241 чел.) репрезентирует взрослое население юга Тюменской области.

Поставим задачу определения влияния некоторых переменных (назовем их переменными-факторами) на согласованность ответов респондентов на следующие вопросы (в скобках приведены обозначения переменных x_{38}, x_{39}, x_{40}, x_{41}).

- Как Вы полагаете, какую роль органы местного самоуправления играют в жизни Вашего региона? (x_{38})
    1. Очень значительную.
    2. Довольно значительную.
    3. Затрудняюсь ответить.
    4. Довольно незначительную.
    5. Совсем незначительную.

- Как Вы думаете, региональная (областная) Администрация прислушивается к общественному мнению при решении злободневных проблем? (x_{39})
    1. Да.
    2. Скорее да, чем нет.
    3. Не могу сказать точно.
    4. Скорее нет, чем да.
    5. Нет.

- Одобряете ли Вы вмешательство власти в конфликты между бизнес-структурами? (x_{40})
    1. Да.
    2. Скорее да, чем нет.
    3. Не могу сказать точно.
    4. Скорее нет, чем да.
    5. Нет.

- Как Вам кажется, может ли простой человек реально отстоять свои законные права перед местными и региональными органами власти? (x_{41})
  1. Может.
  2. Скорее может, чем нет.
  3. Не могу сказать точно.
  4. Скорее не может, чем может.
  5. Не может.

В качестве переменных-факторов в поставленной задаче использовались следующие:

- x_1 – род занятий (с 11 вариантами ответа);
- x_2 – возраст (пять возрастных категорий, закодированных как: $\{1, 2, 3, 4, 5\}$;
- x_3 – образование (четыре уровня образования: $\{1$ – начальное, 2 – среднее, 3 – среднее специальное, 4 – высшее$\}$;
- x_4 – пол респондента (принимает два значения: 1 – мужской, 2 – женский).

Задача решалась в несколько этапов. На *первом* этапе была сформирована матрица рангов, т.е. получены ранжированные ряды по каждой из четырех целевых переменных, имеющих порядковый уровень измерения. Объекты ранжирования – респонденты. Значение коэффициента конкордации для матрицы рангов позволило сделать вывод о том, что гипотеза о согласованности рядов не подтверждается. Вместе с тем, невозможно сделать и обратный вывод – о несогласованности рядов, ибо коэффициент конкордации устремился к нулю, а соответствующая проверка по критерию хи-квадрат (4) показала значения, устремляющиеся в бесконечность. Проверка по формулам (5) также привела к пренебрежимо малым значениям.

Ситуация изменилась при переходе к отдельным группам респондентов, выделенным в зависимости от значений переменных-факторов, влияние которых на согласованность ответов по целевым переменным нас и интересует.

На *втором* этапе решения задачи анализ, проведенный по отдельным группам, выделенным по полу, роду занятий и образованию респондентов, не подтвердил значимость согласованности. Однако подтвердилась согласованность ответов по матрицам рангов, соответствующих возрастным группам. Получили статистически значимую согласованность ответов по целевым переменным x_{38}, x_{39}, x_{41} при фиксировании уровня переменной x_2, ибо коэффициент конкордации принимал статистически значимое значение, что соответствовало удовлетворению критерия (5).

Забегая вперед, отметим, что одна из целевых переменных была исключена из рассмотрения. Введение переменной x_{40} в матрицы рангов (их было много в соответствии с группами) приводило к существенному уменьшению коэффициента конкордации.

В общих чертах алгоритм опирался на выделение отдельных подмножеств (групп) исходного множества респондентов, для которых в *матрицах рангов* наблюдался наибольший разброс рангов. Правомерность такого алгоритма обосновывается тем, что в основе вывода коэффициента конкордации лежит гипотеза о равномерном распределении в столбцах матрицы рангов. При таком распределении, как уже отмечалось, вариация в матрице рангов $S \to \max$.

Реализацию этого простого вычислительного алгоритма допускает как пакет SPSS (в меню – Вычислить), так и широко распространенный среди непрофессионалов пакет программ Excel. Вместе с тем, стандартные процедуры SPSS не позволяют провести вычисления в полном объеме.

В табл. 6 приводится один вариант расчетов по формулам (2) и (7) для группы респондентов – мужчин со средним образованием, возраст которых старше 45 лет (число таких респондентов равно 21 чел.). Введение переменной x_{40} в матрицу рангов привело к существенному уменьшению коэффициента конкордации. В этом случае $W = 0{,}092$.

Процедура вычислений полностью автоматизирована в среде Excel – вводится только исходная матрица данных, а матрица рангов (основания ранжирования – переменные, объекты ранжирования – респонденты) формируется автоматически.

Таблица 6

ДАННЫЕ ДЛЯ ВЫЧИСЛЕНИЯ КОЭФФИЦИЕНТА КОНКОРДАЦИИ

Респонденты	Переменные			Ранжированные ряды			Суммы рангов по строкам	S
	x_{38}	x_{39}	x_{41}	$r^{(38)}$	$r^{(39)}$	$r^{(41)}$		
1	1	0	2	1	1	9	11	182,89
2	1	4	2	1	15	9	25	0,23
3	2	2	5	10	6	20	36	131,70
4	4	5	1	17	18	1	36	131,70
5	3	3	2	11	8	9	28	12,08
6	1	2	1	1	6	1	8	273,04
7	3	3	1	11	8	1	20	20,46
8	1	3	2	11	8	9	28	12,08
9	5	5	2	18	18	9	45	419,27
10	1	1	2	1	2	9	12	156,85
11	3	3	1	11	8	1	20	20,46
12	1	3	2	1	8	9	18	42,56
13	3	1	2	11	2	9	22	6,37
14	5	1	1	18	2	1	21	12,42
15	1	5	2	1	18	9	28	12,08
16	1	1	2	1	2	9	12	156,85
17	1	5	1	1	18	1	20	20,46
18	3	3	1	11	8	1	20	20,46
19	5	4	2	18	15	9	42	305,42
20	5	4	5	18	15	20	53	810,89
21	1	3	1	1	8	1	10	210,94

Так как для каждой переменной наблюдаются связанные ранги, вычисляются поправочные коэффициенты по соответствующим формулам. К примеру, по первой переменной ранг «1» встречается девять раз, ранг «11» шесть раз, а ранг «18» четыре раза. Тогда поправка будет равна:

$T^{(38)} = (9^3 – 9 + 6^3 – 6 + 4^3 – 4)/12 = 82{,}5$.

Поправки, соответствующие двум другим переменным, будут равны:

$T^{(39)} = (6^3 – 6 + 7^3 – 7 + 3^3 – 3)/12 = 52{,}5$;

$T^{(41)} = (8^3 – 8 + 7^3 – 7 + 2^3 – 2)/12 = 152{,}5$.

В нашем случае $n = 21$, $m = 3$. Тогда вариация будет равна:

$S = 2959{,}24$, а коэффициент конкордации:

$W = 2959{,}24 / (3^2*(21^3 – 21)/12 – 3*(82{,}5 + 52{,}5 + 152{,}5))$, т.е. $W =$ **0,49.**

Тогда значение χ_P^2 = **29,26**.

Число степеней свободы $f = n – 1 = 20$, соответственно χ_T^2 = **28,41** для уровня значимости, равного 0,1. Так как $\chi_P^2 > \chi_T^2$, то критерий (4) удовлетворяется на уровне погрешности 10%. Для других уровней значимости нельзя сделать такой вывод.

Величина коэффициента конкордации подтверждает умеренную тесноту связи между целевыми переменными. Тем самым можно сделать вывод о том, что мужчины со средним образованием, возраст которых старше 45 лет, образуют гомогенную группу, т.е. согласованно отвечают на вопросы, которые мы обозначили как целевые переменные (первый, второй и четвертый из них).

В завершение рассмотрения *примера 2* отметим, что предложенный алгоритм не заменяет методы кластерного или факторного анализа, а скорее предваряет их, позволяя оценить степень гомогенности исходного массива данных или подмассивов.

Применение к исходным данным алгоритмов кластерного анализа, заложенного, например, в SPSS, не гарантирует получение прозрачно интерпретируемого результата ввиду «слепого» характера алгоритмов, заложенных в таких пакетах.

В предложенном алгоритме производится «персонифицированный» отбор, т.е. мы последовательно ищем ответ на вопрос: согласованно ли отвечают на наши вопросы вполне определенные группы респондентов. Применение простейшего алгоритма поиска, например, метода скорейшего спуска, позволяет довольно быстро выбрать наиболее согласованно отвечающую группу. Этот метод основан на принципе максимизации (или минимизации) целевой функции (в нашем случае это вариация *S матрицы рангов*), а переменные на каждом шаге выбираются из условия скорейшего приближения.

Пример 3.

Данный пример относится к ситуации, в которой исследователь использует коэффициент конкордации для проверки достоверности результатов, полученных другими методами, например, «работоспособности» некоторых введенных им индексов. В нашем примере речь идет об индексе качества жизни, построенном на основе субъективных оценок респондентов по их удовлетворенности различными сторонами жизнедеятельности.

Эмпирическая база, на которую мы опираемся, та же, что и в *примере 2* – результаты опроса населения Тюменской области, проведенного в 2002 г. В вопросник конфликтологического мониторинга для выявления самооценок населения удовлетворенностью качеством жизни был включен вопрос табличного вида:

Удовлетворены ли Вы...? *(ответ дается по каждой строке таблицы).*

Далее предлагалась таблица с десятью различными сторонами жизнедеятельности (см. табл. 7). Варианты ответа кодировались от 1 – *совсем не удовлетворен* до 4 – *вполне удовлетворен*.

При дальнейших рассуждениях будем исходить из некоторой логики исследовательского поиска, типичной для социологических исследований. В табл. 7 представлены распределения опрошенных по степени их удовлетворенности различными сторонами их жизнедеятельности. Это те данные, которые визуально воспринимает исследователь для предварительных выводов об удовлетворенности качеством жизни.

Таблица 7

РАСПРЕДЕЛЕНИЯ ОПРОШЕННЫХ ПО СТЕПЕНИ ИХ УДОВЛЕТВОРЕННОСТИ РАЗЛИЧНЫМИ СТОРОНАМИ ЖИЗНЕДЕЯТЕЛЬНОСТИ, % к числу опрошенных

Удовлетворены ли Вы…		Степень удовлетворенности			
		Совсем не удовлетворен	В основном не удовлетворен	В основном удовлетворен	Вполне удовлетворен
1	Тем, как питаетесь	3,9	11,5	42,8	41,7
2	Тем, как одеваетесь	8,4	20,1	45,2	28,5
3	Медицинским обслуживанием	20,8	31,2	34,1	13,9
4	Системой образования	15,6	27,7	42,8	13,9
5	Вашим доходом	19,1	30,7	40,1	10,1
6	Вашим жилищем	14,7	21	38,5	25,8
7	Вашим отдыхом	17,5	29,2	37,2	16,1
8	Вашей работой	9,7	15,8	51,8	22,6
9	Вашими жизненными перспективами	13,1	24,7	44,4	17,8
10	Тем, как складывается Ваша жизнь в целом	7,9	18,4	53,6	20

Очевидно, что сделать содержательные выводы по этой таблице весьма сложно, и тем более оценить значимость этих выводов. Попытка сравнить средние величины также не позволяет сделать сколько-нибудь значимых выводов. В табл. 8 приведены данные, где n – число ответивших, x – среднее арифметическое по ответившим, σ – среднеквадратическое отклонение.

Таблица 8

ЗНАЧЕНИЯ ОПИСАТЕЛЬНЫХ СТАТИСТИК

Удовлетворены ли Вы…		n	$\bar{x}$	σ
1	Тем, как питаетесь	1241	3,22	0,803
2	Тем, как одеваетесь	1240	2,91	0,805
3	Медицинским обслуживанием	1239	2,41	0,968
4	Системой образования	1236	2,55	0,916
5	Вашим доходом	1088	2,41	0,910
6	Вашим жилищем	1238	2,75	0,998
7	Вашим отдыхом	1238	2,52	0,960
8	Вашей работой	1231	2,87	0,871
9	Вашими жизненными перспективами	1233	2,67	0,916
10	Тем, как складывается Ваша жизнь в целом	1238	2,86	0,826

На основе визуального восприятия данных табл. 8 можно сделать вывод о средней степени удовлетворенности респондентов сложившейся ситуацией. Ограничить анализ такой интерпретацией недопустимо. Прежде всего потому, что вычисление средней арифметической по порядковым шкалам некорректно. Тогда можно предложить несколько иную логику анализа удовлетворенности качеством жизни.

Введем индекс удовлетворенности *Ind* отдельной стороной жизнедеятельности. Определим его следующим образом:

$$Ind = (P|_{k=4} + (P|_{k=3} - P|_{k=2}) \times 0{,}5 - P|_{k=1}) / 50,$$

где, например, $P\big|_{k=4}$ – доля респондентов (в процентах), ответивших, что они *полностью удовлетворены*, что соответствует коду 4.

Значение этого индекса меняется от –2 (случай, когда все респонденты *совершенно не удовлетворены*) до 2 (случай, когда все *полностью удовлетворены*). Значение *Ind* = 0 соответствует случаю, когда половина респондентов удовлетворена, а половина не удовлетворена.

На рис. 1 приведены значения индекса для нашего случая.

Менее всего население Тюменской области удовлетворено медицинским обслуживанием (*Ind* = –0,109), а также доходом (*Ind* = –0,08). Интересным является факт высокой удовлетворенности питанием (1,069), ибо этот результат существенно отличается от результата, полученного по другим российским регионам (например, по Туле и Ярославлю). Этот факт характерен для относительно благополучного региона.

Возникает вопрос, насколько значим полученный вывод и можно ли его подтвердить посредством других методов анализа данных? Для ответа на этот вопрос используем коэффициент конкордации. Рассмотрим *матрицу рангов*, в которой объектами ранжирования выступают виды жизнедеятельности (в нашем случае их десять), а основанием ранжирования – респонденты. В табл. 9 приведены значения средних рангов по всем респондентам и полученное на их основе групповое упорядочение (относительные ранги). Следует отметить, что в анализе участвуют только те респонденты, у которых есть оценки по всем десяти видам жизнедеятельности.

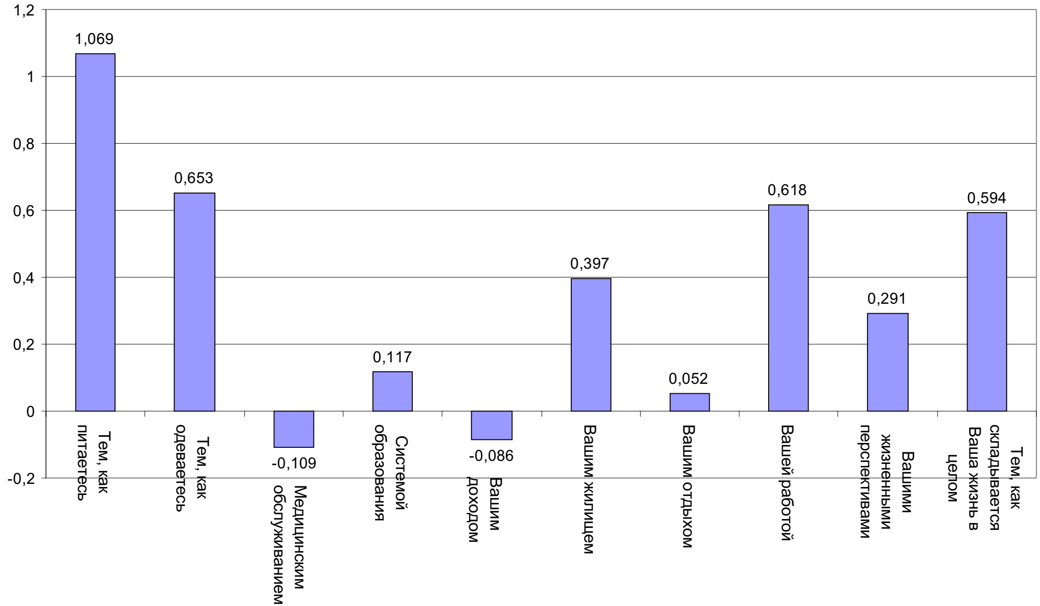

Рис 1. Значения индекса удовлетворенности качеством жизни

Таблица 9

ЗНАЧЕНИЯ РАНГОВ

Удовлетворены ли Вы…		Средний ранг	Относительный ранг
1	Тем, как питаетесь	7,26	10
2	Тем, как одеваетесь	6,13	9
3	Медицинским обслуживанием	4,48	2
4	Системой образования	4,98	4
5	Вашим доходом	4,39	1
6	Вашим жилищем	5,60	6
7	Вашим отдыхом	4,78	3
8	Вашей работой	6,03	8
9	Вашими жизненными перспективами	5,34	5
10	Тем, как складывается Ваша жизнь в целом	6,00	7

Средний ранг показывает среднее положение переменной на шкале от 1 до 10, относительный ранг – относительное положение переменной на той же шкале. Ранг 1 присваивается переменной с наименьшей степенью удовлетворенности, ранг 10 – с наибольшей степенью удовлетворенности.

Далее были проведены расчеты, аналогичные тем, которые приводились при рассмотрении *примера 1* и *примера 2*. В *примере 3* они проведены при помощи пакета программ SPSS. На основе критерия конкордации W Кендалла была проверена гипотеза о согласованности ранжировок. Эта гипотеза подтверждается на уровне, превышающем 99,9% (табл. 10).

Таблица 10

СТАТИСТИКА КРИТЕРИЯ

N	1076
Статистика W Кендалла	0,124
Хи-квадрат	1200,05
Степени свободы	9
Асимптотическая значимость	0,000

Критерий согласия по коэффициенту конкордации Кендалла показывает максимальный уровень значимости, поэтому групповое упорядочение различных сфер жизнедеятельности на основе средних рангов можно считать обоснованным.

Результаты (эмпирические закономерности), показанные на рис. 1 и в табл. 9, в принципе совпадают и поэтому могут быть основанием для более глубокого и детального анализа.

Расхождение результатов применения двух подходов к анализу удовлетворенности различными сторонами жизнедеятельности наблюдается в двух случаях (фактически это один случай). Первый случай относится к удовлетворенности медицинским обслуживанием (относительный ранг равен 2, а по значению индекса эта сфера выходит на первое место по неудовлетворенности). Второй случай относится к удовлетворенности доходом (относительный ранг равен 1), а по значению индекса выходит на второе место по неудовлетворенности.

Объяснение этого факта достаточно простое. Оказалось, что в ответах на эти вопросы имелось малое количество выбросов, которые внесли вклад в значение индекса удовлетворенности, но не оказали влияние на ранжирование и оценки согласованности ответов.

ЛИТЕРАТУРА

1. *Татарова Г.Г.* Метаметодика анализа данных как средство концептуализации // Россия реформирующаяся: Ежегодник – 2003 / Отв. ред. Л.М. Дробижева. М.: ИС РАН, 2003.

2. *Татарова Г.Г.* Система языковых конструктов анализа социологических данных // Вестн. Рос. ун-та др. нар. Сер. Социология. 2003. С. 35–46.

3. *Айвазян С.А., Мхитарян В.С.* Прикладная статистика и основы эконометрики. М.: ЮНИТИ, 1998.

4. Теория статистики / Под ред. Р.А. Шмойловой. М.: Финансы и статистика, 1999.

5. *Толстова Ю.Н.* Анализ социологических данных. М.: Научный мир, 2000.

6. *Аргунова К.Д., Татарова Г.Г.* Выбор стратегии анализа взаимосвязи признаков // Математические методы анализа и интерпретация социологических данных. М.: Наука, 1989. С. 61–94.

7. *Бондарь А.Г., Статюха Г.А.* Планирование эксперимента в химической технологии (основные положения, примеры и задачи). Киев: Вища школа, 1976.

8. *Кремер Н.Ш.* Теория вероятностей и математическая статистика. М.: ЮНИТИ-ДАНА, 2000.

9. *Рузинов Л.П.* Статистические методы оптимизации химических процессов. М.: Химия, 1972.

10. Вероятность и математическая статистика: Энциклопедия / Под ред. А.М. Прохорова. М.: Научное издательство «Большая российская энциклопедия», 1999. С.537–538.

11. *Бююль А., Цефель П.* SPSS: искусство обработки информации: Анализ статистических данных и восстановление скрытых закономерностей / Пер. с нем. СПб.: ООО «ДиаСофт», 2002. С. 246–247.

12. *Орлов Г.М., Шуметов В.Г.* Модель электоральных предпочтений: методология построения // Социологические исследования. 2001. № 1. С. 127–141.