Многомерные статистические совокупности и их метрический анализ*

Г.К. Каменев[1], И.Г. Каменев[1,2]

[1] Вычислительный центр им. А.А. Дородницына Федерального исследовательского центра «Информатика и управление» Российской академии наук, г. Москва, Россия
[2] Негосударственное частное образовательное учреждение высшего образования «Синергия», г. Москва, Россия

Аннотация. В статье рассматривается проблема анализа распределений в многомерном пространстве характеристик субъектов в общественных науках. Обосновывается необходимость разработки математических методов, учитывающих как меру (частотность), так и метрику. Исследуются методы метрического анализа надежности для генеральной совокупности на основе построения (ε, δ)-сетей (Шеннон). В качестве примера приведен анализ данных ООН по человеческому развитию.
Ключевые слова: *многомерный статистический анализ, генеральная совокупность, выборка, метрические сети, визуализация данных, индексы ООН для развития человечества.*
DOI: 10.14357/20790279180207

Введение

Многие теоретические и прикладные задачи в общественных науках сегодня сводятся к анализу многомерных множеств: количественных характеристик субъектов (потребителей, избирателей, работников, предприятий, сообществ и др.). Исходя из сочетаний характеристик, субъекты подразделяются на категории (классы), которым приписывается определенная модель поведения. Для построения таких моделей используются данные выборочных обследований. В последние годы все чаще звучит их критика: социологи ставят вопрос о наличии систематических ошибок в выборке [1]. В современной постиндустриальной экономике и обществе существенно выросла значимость «меньшинств», то есть групп, отличающихся по своим свойствам от большинства: количественно немногочисленных, но метрически значимых. Необходимо оценивать не вероятность ошибки в каждой характеристике отдельно, а вероятность ошибок в сочетании характеристик. Для этого предлагается проводить метрический анализ надежности многомерных социологических выборок

1. Алгоритм изучения многомерного пространства характеристик

Многомерное пространство характеристик представляет собой совокупность свойств, одновременно приписываемых субъекту (актору, агенту). Закономерности поведения субъекта определяются сочетанием его изучаемых характеристик. Каждая характеристика должна быть не качественной, а количественной (предпочтительно, непрерывной). Пространство характеристик представляет собой развитие факторного анализа в рамках поведенческого подхода: каждая характеристика представляет собой фактор (вход), а поведение субъекта – функцию (выход). Тот же принцип может применяться и к институтам, и механизмам, для которых корректным термином на выходе будет не «поведение», а «функционирование».

Изучение многомерного пространства характеристик предлагается осуществлять, четко следуя алгоритму действий. 1) Сбор данных по характеристикам. 2) Отбор характеристик. 3) Построение многомерного пространства. 4) Восстановление генеральной совокупности. 5) Структурный и визуальный анализ объектов в многомерном пространстве. 6) Обобщение характеристик и изучение особенностей поведения выделенных классов. Выделенные группы субъектов следует подробно описать, выделив ключевые характеристики, определенное (сходное) значение которых определяет близость их расположения.

2. Метрический анализ выборки

Рассмотрим теперь вопрос о построении и визуализации многомерных генеральных сово-

* Работа выполнена при финансовой поддержке РНФ, проект 14-11-00432.

купностей (Statistical population), лежащих в основе многомерных социологических выборок. В качестве генеральной совокупности мы будем рассматривать случайную векторную величину $a=(a_1, \dots a_m)\in\mathbf{R}^m$, принимающую конечное, возможно, очень большое число различных значений из множества $A=\{a^1, \dots a^L\}\subset\mathbf{R}^m$. Пусть на совокупности A задана некоторая многомерная случайная величина Ξ, в самом общем виде отражающая процесс получения социологической информации. Нас будет интересовать не закон распределения этой величины, а способ метрического восстановления генеральной совокупности A по некоторому подмножеству из A и независимой выборке объема N.

Универсальным средством аппроксимации сложных метрических объектов являются метрические ε-сети. Пусть A и U – непустые подмножества пространства $\mathbf{R}$ с метрикой ρ. Множество U называется метрической ε-*сетью для* A, если любая точка A расположена на расстоянии не большем, чем ε, от некоторой точки U. Если конечное U является метрической ε-сетью для A, то множество A имеет *аппроксимацию* в виде ε-*покрытия*, состоящего из системы шаров радиуса ε с центрами в точках U. В настоящей статье будет использован подход, разработанный в [2] для стохастически заданных множеств, т.е. для множеств с заданной на них вероятностной мерой. Предполагается, что имеется возможность получения независимых выборок значительного объема, т.е. случайных точек, принадлежащих ограниченному носителю в заданном метрическом пространстве. Заметим, что при стохастическом подходе к построению метрических сетей и покрытий возникает дополнительная проблема надежности оценок точности и качества.

Пусть задано некоторое конечное множество $T\subset\mathbf{R}$, которое мы будем в дальнейшем называть *базой* сети или покрытия. Обозначим через $(T)_\varepsilon$ его открытую ε-окрестность, т.е. множество $(T)_\varepsilon = \{a\in\mathbf{R}: \exists t\in U,\ \rho(t, a)<\varepsilon\}$, обозначим также $[T]_\varepsilon = \{a\in\mathbf{R}: \exists t\in T,\ \rho(t, a)\le\varepsilon\}$. Пусть на борелевской σ-агебре $\mathbf{B}(A)$ задана вероятностная мера μ(×), $\mu(A)=1$. Множество T называется (ε, δ)-*сетью* для A (в смысле определения К. Э. Шеннона), если $\mu(A/[T]_\varepsilon)\le\delta$, т.е. любая точка A, за исключением множества меры δ, расположена на расстоянии не большем, чем ε, от некоторой точки U [3]. Методы построения (ε, δ)-сетей для стохастически заданных множеств рассмотрены в [2]. В случае больших данных (тысячи и миллионы точек) эти методы основаны на разделении выборки на базовую и тестовую часть, которая используется для оценки с заданной надежностью меры δ генеральной совокупности, лежащей вне ε-окрестности базовой части. В случае big data (миллиарды точек), используется метод глубоких ям для фильтрации потока случайных точек с целью получить равномерное распределение отобранных базовых точек, ε-окрестность которых с заданной надежностью является (ε, δ)-сетью генеральной совокупности. При разделении на базу и тестовую выборки малого объема (сотни точек) могут быть использованы различные методы генерации статистического бутстрепа.

После построения аппроксимации генеральной совокупности в виде совокупности метрических шаров, соответствующих заданной точности, анализ такого множества может проводиться средствами Диалоговых Карт Решений, разработанных в Вычислительном Центре РАН в рамках метода аппроксимации и визуализации неявно заданных множеств [5].

3. Пример

Рассмотрим данные ООН по различным индексам развития человечества (Human Development Data) за 1990-2015 годы (http://hdr.undp.org/en/data). Объектами наблюдения в данном случае являются страны. Выделим из них ряд показателей, связанных с человеческим капиталом. Наиболее полные данные имеются за 2015 год. Требуется метрически охарактеризовать всю совокупность стран с точки зрения выбранных показателей и, по возможности, выделить в ней содержательные структуры. В этом случае мы имеем дело с выборкой малого объема. Прежде всего, заметим, что часть показателей в рассматриваемой выборке в большей или меньшей степени коррелируют друг с другом, т.е. увеличение одного из них влечет за собой увеличение другого (положительная корреляция) или уменьшение другого (отрицательная) практически во всех точках выборки. Это позволяет ограничиться в дальнейшем исследованием одного показателя из этой группы, например, Education Index.

Рассмотрим аппроксимацию с точностью около 7% генеральной совокупности в пространстве трех не коррелирующих параметров (рис. 1). Можно выделить 4 компоненты генеральной совокупности.

I. Страны с хорошим уровнем социально-демографических показателей (высокими Education Index и Public health expenditure и низким Inequality in income). Эта структура находится в нижнем правом углу рис.1 и включает слои светлой штриховки. Условно ее можно описать как кластер с ограничениями: Education Index ≥ 0.8, Inequality in income ≤ 20, Public health expenditure ≥ 5.

II. Уникальный кластер на основе одного выборочного элемента. Этот кластер находится в центре рис. 1 и имеет светлую штриховку, соответствующую максимальному значению показателя Public health expenditure. Этот кластер можно описать следующим образом: Education Index ≈ 0.56, Inequality in income ≈ 23, Public health expenditure ≈ 10.8 (максимум).

III. Основной массив стран с удовлетворительным уровнем (низкими и средними значениями показателя) Inequality in income. Эта структура генеральной совокупности не является кластером и ее сложно формально отделить от первых структур. Условно его можно описать ограничениями: Inequality in income ≤ 40, Public health expenditure < 5.

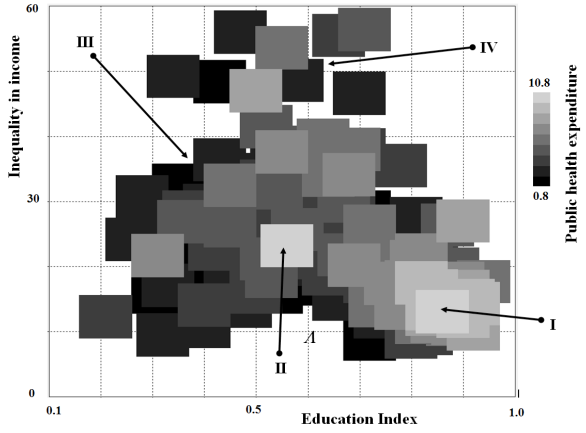

Рис. 1. Метрические компоненты генеральной совокупности

Заметим, что в этой части генеральной совокупности имеется следующая особенность: отсутствуют страны со средним уровнем Education Index и низким уровнем Inequality in income.

IV. Кластер «аутсайдеров» с высоким уровнем показателя Inequality in income. Условно его можно охарактеризовать ограничением: Inequality in income > 40. Заметим, что в этой части генеральной совокупности присутствуют страны со средним уровнем Education Index и низким уровнем Inequality in income. Именно в этой части наблюдается недостаток данных, которые желательно дополнить на основе сбора государственной статистики. Рассмотренный пример показывает, что на основе научно обоснованного пространственного распределения становится возможным изучение структур (как большинства, так и меньшинств), обладающих схожим набором характеристик, и особенностей их поведения (функционирования), относящихся к ним субъектов, что составляет предмет особого исследования.

Литература

1. *Саганенко Г. И.* Структура эмпирического результата в социологии и проблема его надежности // Социология: методология, методы, математическое моделирование, 1994. № 3-4. С. 5-22.
2. *Kamenev G.K.* Approximation of Completely Bounded Sets by the Deep Holes Method // Comput. Maths. Math. Phys. 2001. Vol.41. N11. Pp. 1667-1675.
3. *Shannon K.* The Mathematical Theory of Communication // The Bell System Technical Journal, Vol. 27, pp. 379–423, 623–656, July, October, 1948.
4. *Lotov A.V., Bushenkov V.A., Kamenev G.K.* Interactive Decision Maps. Approximation and Visualization of Pareto Frontier. Appl. Optimization. V. 89. Kluwer Academic Publishers. Boston / Dordrecht / New York / London. 2004.

Каменев Георгий Кириллович. Вычислительный центр им. А.А. Дородницына ФИЦ ИУ РАН, г. Москва, Россия. Ведущий научный сотрудник. Доктор физико-математических наук. Количество печатных работ: 53 (в т.ч. 6 монографий). Область научных интересов: выпуклая геометрия, многокритериальные решения. E-mail: gkk@ccas.ru.

Каменев Иван Георгиевич. Вычислительный центр им. А.А. Дородницына ФИЦ ИУ РАН, г. Москва, Россия. Негосударственное частное образовательное учреждение высшего образования «Синергия», г. Москва, Россия. Младший научный сотрудник, доцент. Кандидат экономических наук. Количество печатных работ: 23. Область научных интересов: экономика труда. E-mail: igekam@gmail.com

Multidimensional statistical sets and their metric analysis

G.K. Kamenev[1], I.G. Kamenev[1,2]

[1] Federal research center of Informatics and Management Dorodnicyn Computing centre of the Russian science academy

[2] Private Educational Institution of Higher Education «Synergy»

Abstract. The article describes mathematical analysis methods of distributions in multidimensional space of subject's characteristics in social sciences. The necessity is motivated to take into account measure (frequency) and metric. Methods of metric reliability analysis for the general population are considered using (ε, δ)-sets (Shannon). As an example UN Human Development Data are analyzed.

Keywords: *multidimensional statistical analysis, general population, sociological sample, metric net, approximation, data visualization, UN human development data.*

DOI: 10.14357/20790279180207

References

1. *Saganenko G. I.* Struktura empiricheskogo rezultata v sotsiologii i problema ego nadezhnosti [Empirical result structure in sociology and its reliability problem] // Sotsiologiya: metodologiya, metodyi, matematicheskoe modelirovanie [Sociology: methodology, methods, mathematical modeling], 1994. Pp. 3-4. C. 5-22
2. *Kamenev G.K.* Approximation of Completely Bounded Sets by the Deep Holes Method // Comput. Maths. Math. Phys. 2001. Vol.41. N11. Pp. 1667-1675.
3. *Shannon K.* The Mathematical Theory of Communication // The Bell System Technical Journal, Vol. 27, pp. 379–423, 623–656, July, October, 1948.
4. *Lotov A.V., Bushenkov V.A., Kamenev G.K.* Interactive Decision Maps. Approximation and Visualization of Pareto Frontier. Appl. Optimization. V. 89. Kluwer Academic Publishers. Boston / Dordrecht / New York / London. 2004.

Kamenev Georgy Kirillovich. Federal Research Center of Computer Science and Control of the Russian academy of science, Moscow, Russia, Vavilova str, 40. Leading Research Fellow. Ph.D., Dr. Hab. in Applied Mathematics. Senior researcher. Publications number: 53 (6 monograph.). Scientific interests: convex geometry, multiple criteria decisions. E-mail for correspondence: gkk@ccas.ru

Kamenev Ivan Georgievich. Federal Research Center of Informatics and Management of the RAS, Moscow, Russia, Vavilova str, 40. Junior Research Fellow. Ph.D. in Economic. Private Higher Educational Institution «Synergy», Moscow, Russia, Izmailovsky val str., 2. Lecturer. Publications number: 23. Scientific interests: labor economics. E-mail: igekam@gmail.com