ПРИМЕНЕНИИ МЕТОДОВ МНОГОМЕРНОЙ КЛАССИФИКАЦИИ ПРИ ИЗУЧЕНИИ КАЧЕСТВА СЕЛЬСКОГО НАСЕЛЕНИЯ[1]

Бавыкина О.В. , Ускова О.Ф.
(Воронеж)

Излагается опыт использования математико-статистического классифицирования (с применением компьютерных технологий) качества сельского населения Воронежской области, что позволило эффективно обнаружить географические закономерности изменчивости медико-демографического состояния селян.

THE EXPERIENCE OF USING MATH-STATISTIC CLASSIFICATION THE QUALITY OF RURAL POPULATION

Bavykina O.V., Uskova O.F.

(Voronezh)

Abstract. The article describes the experience of using math-statistic classification (with employment of computer technologies) the quality of rural population of Voronezh region, which allows to detect effectively the geographical regularity of changeability of medical demographical state of seeds.

В практику научных исследований понятие "качество населения" введено в связи с поиском интегрального показателя оценки состояния общественного здоровья (или качества) населения, "выражающегося в способности населения выполнять биологические, социальные, экономические функции" (3). Проблема изучения качества населения возникает при решении

[1] Работа выполнена при финансовой поддержке РГНФ, проект № 02-06-00301 а/Ц

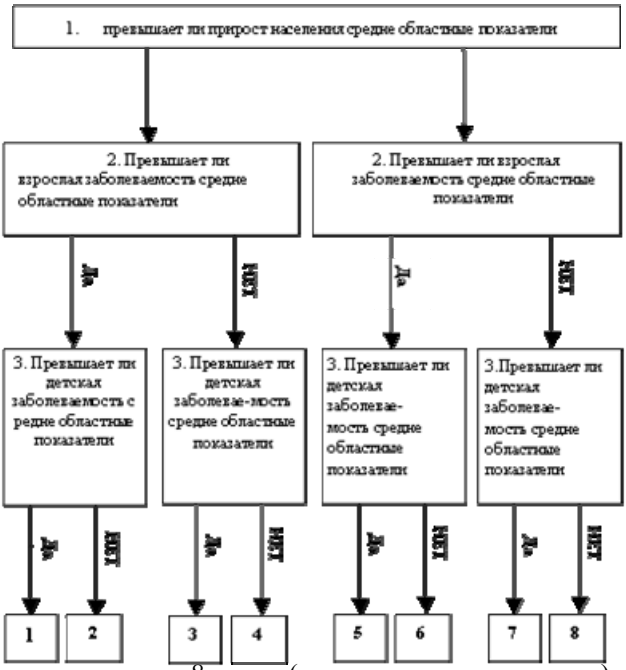
многих экологических, социально-экономических задач теоретического и прикладного назначения.

В современной экологии человека понятие "качество населения" часто привлекается как обобщенный диагностический индикатор среды "жизни", ее жизнепригодных и жизнеобеспечивающих свойств. Именно такой методологический подход использовался при изучении среды обитания сельских поселений Воронежской области. Теоретической моделью при этом послужило представление о социоэкосистеме "население – среда обитания", в которой экосистемные свойства составляющих компонентов в результате взаимодействия находятся в соответствии. Потому целесообразно и методологически оправдано было начать экосистемное исследование с установления территориальной дифференциации качества населения.

Формирование базы данных об объекте изучения предусматривало следующие методические приемы: 1) наблюдением охвачены поселения численностью более 500 человек; 2) отбор объектов наблюдения осуществлен по правилам выборочного метода и по схеме типической гнездовой случайной выборки в опоре на схемы географического районирования, что позволило наиболее реалистично и надежно охватить вниманием объекты с разнообразными свойствами. Объем выборки, гарантирующий обследование наиболее характерных и разнообразных особенностей сельской среды, рассчитан по правилам выборочного метода. Однако в целях достаточно надежного выявления наиболее тонких закономерностей объем выборки значительно завышен относительно статистически рассчитанного и составил 145 поселений; 3) данными о качестве населения стали медико-демографические показатели: средний прирост населения, заболеваемость на 1000 жителей, рождаемость и смертность. В итоге обретена массовая аналитическая информация, представляющая многомерную совокупность, трудно поддающаяся визуальному анализу. Возможность упростить и объективизировать анализ такой совокупности заключен в существующих методах многомерного классифицирования. Для реализации классификационного подхода привлекались два алгоритма.

Природа первого алгоритма (2) основывается на бинарном делении, в основу его положено сравнение трех параметров ка-

чества населения отдельных поселений со среднестатистическими данными этих же значений по области в целом (схема 1).



В итоге получено 8 групп (типов качества населения), различающихся своими характерными параметрами прироста населения, взрослой и детской заболеваемостью (таблица 1).

Таблица 1. Таблица характеристик групп, полученных в результате классификации методом деревьев

№ группы	Количество сел	Прирост (относительно средних показателей по области)	Взрослая заболеваемость (относительно средних показателей по области)	Детская заболеваемость (относительно средних показателей по области
1	26	больше в 1 - 9	больше в 1 - 7	больше в 1 – 15
2	0	-	-	-
3	33	больше в 1 – 9	меньше в 1 – 7	больше в 1 – 8
4	2	больше в 1 - 3	меньше в 3 –7	меньше в 2 – 3

5	34	меньше в 1 – 3	больше в 1 – 6	больше в 1 – 26
6	1	меньше в 2	больше в 2	меньше в 2
7	44	меньше в 1 – 3	меньше в 1 – 10	больше в 1 – 14
8	5	меньше в 1 – 2	меньше в 2 - 11	меньше в 1 – 2

По итогам классификации составлена карта. Различные группы сел, различающихся по медико-демографическому качеству населения, расположились по территории области упорядоченно: наиболее многочисленные группы создают четкие полосы (местами прерывающимися немногочисленными группами) субмеридианальной и субширотной ориентации. Обнаруженные закономерности размещения качества населения предположительно можно объяснить внутренними особенностями человеческой популяции – генетико-автоматическими процессами.

Процедура второй классификации формально ставилась следующим образом. Классификацию можно определить как объединение объектов в классы, опирающееся либо на сходство свойств, либо на взаимные связи между ними. В основу классификации должны быть положены свойства, неотъемлемо присущие всем объектам исходной генеральной совокупности. Мы будем следовать концепции "однородных" классов, понимая под этим, что любые два достаточно близких (сходных) объекта должны принадлежать одному классу.

Формально задача классификации ставится следующим образом. Задан массив X, состоящий из N точек, каждая из них характеризуется n-мерным вектором $x^i = (x^i_1, \ldots, x^i_n)$ в евклидовом пространстве R^n. Требуется разбить его на «классы», образующие «существенные сгущения». Имеется предположение, что таких классов получится по меньшей мере два.

Пусть на исходном массиве X определена мера близости $\mu(x,y)$. Одним из наиболее важных условий любой объектной классификации мы считает выполнение следующих (для случая двух классов A и B) неравенств

$$\inf_{x \in A,\, x \neq u} \mu(x, u) < \inf_{y \in B} \mu(y, u),$$

$$\inf_{x \in A} \mu(x, v) > \inf_{y \in B,\, y \neq v} \mu(y, v) \qquad (uA, vB)$$

Аналогичные условия можно выписать и для большего числа классов. Это замечание с одной стороны подчеркивает тот факт, что каждое скопление (сгущение) точек X в пространстве признаков должно целиком принадлежать одному классу, а с другой стороны – что при классификации нужно не только отыскивать «сгущения», но и следить за тем, чтобы границы классов проходили по «разрежениям». Отмеченные обстоятельства были отправными при разработке излагаемого ниже алгоритма.

Пусть для каждой точки $x^i \in X$ зафиксировано некоторое множество $U(x^i) \in R^n$, содержащее эту точку. Множество $U(x^i)$, будем называть окрестностью точки x^i , а количество $P(x^i)$ точек из X, содержащихся в $U(x^i)$ – плотностью массива X в точке x^i .

Пусть q – положительная константа. Множество $A \subset X$ назовем q-связным (или q-однородным), если для любых x, $y \in A$ существует конечная последовательность $x=x^1, x^2 \ldots x^k=y$ точек из A такая, что при любом i = 1, 2, … k-1 точки x^i и x^{i+1} являются q-близкими, т. е. $\mu(x, y) \le q$. Высотой множества $A \subset X$ назовем число max p(x)-min p(x), ($x \in A$), а абсолютной высотой A – число max p(x) ($x \in A$).

Пусть h – положительное число. Множество $A \subset X$ назовем h-почти связным, если все его q-связные компоненты (за исключением может быть одного) имеют высоту, меньшую h.

Множество $A \subset X$ назовем существенным сгущением X, если

а) A является q-связным,

б) абсолютная высота A не меньше h,

в) при любом $\tau \ge 0$ множество $A_\tau = \{x \in A : p(x) \ge \tau\}$ является h-почти связным,

г) при $\tau \ge \max p(x)-h$, ($x \in A$) справедливо неравенство

$\mu(A_\tau, (X \setminus A)_\tau) > q$

Здесь через μ (A, B) обозначается мера близости между множествами A и B в естественном смысле:

μ (A, B) = inf μ (x,y), где $x \in A$, $y \in B$.

Из определения нетрудно заметить, что пересечение различных существенных сгущений имеет абсолютную высоту, не большую h.

Пусть различные множества $U(x^i)$ не пересекаются. Если положить $p(x)=p(x^i)$ для любого $x \in U(x^i)$, то функция $p(x)$ является некоторым приближением плотности распределения массива X. Как легко видеть, каждое существенное сгущение X содержит одну из точек максимума функции $p(x)$.

Понимая теперь под классом существенное сгущение, мы сводим задачу классификации к выделению в X совокупности попарно непересекающихся существенных сгущений.

Величину q, с помощью которой мы определяем понятия близости и связности, удобно выбирать, когда заранее можно оценить порог q такой, что при $\mu(x, y) \le q_0$ элементы x, y можно считать схожими из соображений специфики задачи. В противном случае q необходимо варьировать, добиваясь наиболее четкого распределения X на непересекающиеся существенные сгущения. Следует отметить, что ниже мы заранее фиксируем число q, полагая его равным единице. Такой жесткий выбор будет в дальнейшем компенсироваться подбором подходящим масштабных единиц.

Известно, что для эффективной математической обработки большого массива эмпирических данных, как правило, полезно этот массив несколько «огрубить» (или усреднить) с тем, чтобы по возможности уменьшить влияние «информационного шума».

Описываемое огрубление будет определяться двумя целочисленными параметрами: r и h_0. Найдем величины:

$$h_k = \min_i x_k^i, H_k = \max_i x_k^i \ (k=1, 2, \dots n).$$

Тогда наименьший параллелепипед с осями, параллельными осям координат, содержащий множество X, описывается неравенством $h_k \le x \le H_k$ $(k=1, 2, \dots, n)$.

Далее, разобьем каждую сторону этого параллелепипеда на r равных частей и проведем через точки деления гиперплоскости так, чтобы весь параллелепипед разбился на r^n параллелепипедов с осями, параллельными осям координат. Теперь каждую точку $x^i \in X$ заменим ближайшей к началу координат вершиной z^i того маленького параллелепипеда, в котором эта точка лежит, причем координатами точки z^i будет считать целые числа от 0 до $r - 1$, порожденные заданным разбиением параллелепипеда. Точнее говоря, где квадратные скобки означают взятие целой части.

Если r достаточно велико, то переход от массива X к массиву Z (со старыми координатами) почти не изменяет структуру множества X.

В полученном массиве Z, состоящем из N точек с целочисленными координатами, многие точки совпадают.

$$Z_k^i = \left[\frac{\left(x_k^i - h_k \right)}{H_k - h_k} \right], \; k = 1, 2, ..., n$$

Поэтому мы рассмотрим еще массив W, который получается из массива Z отождествлением одинаковых точек. Для любого $\omega \in W$ обозначим через $p(\omega)$ число всех точек $z \in Z$, совпадающих с ω. Нетрудно видеть, что величина $p(\omega)$ совпадает с плотностью исходного массива X в любой точке x^i, попавшей при огрублении в ω, если окрестностью x^i считать элементарный параллелепипед (без некоторых граней), содержащих x^i. Таким образом, функцию p можно считать «приближением» плотности исходного массива.

Переход к целочисленным координатам означает выбор новых единиц на осях координат. Можно было бы для каждой оси ввести свой параметр r_k с тем, чтобы единицы на всех осях оказались одинаковые. Выбирая на некоторых осях маленькие единичные отрезки, мы тем самым придаем этим координатам большое значение при классификации.

Отметим еще, что, очевидно, степень огрубления задачи обратно пропорциональна параметру r.

Второй этап огрубления связан с параметром h_0. Мы исключим из массива W все точки ω, для которых $p(\omega) < h_0$. Новый массив обозначим через V. Таким образом, «редкие» точки массива не будут участвовать в классификации. Мы отнесем их к «нейтральному» классу, который обозначим через X_0. Конечно, при выборе параметра h_0 следует позаботиться о том, чтобы в нейтральный класс попало не слишком много точек.

Результаты классификации, представленные на географической карте, выявили иной пространственный порядок: типические особенности в качестве сельского населения территориально соответствуют природной среде, аномальные и немного-

численные обнаруживают привязанность к экономико-техническим объектам.

Таким образом, в итоге классифицирования обретен сконцентрированный материал, который позволил выводить обобщенные характеристики выделенных особенностей качества населения и подойти к объяснению обнаруженных различий.

Литература

1. Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрики. – М., 1998. – 524 с.

2. Ашмарин И.П., Васильев Н.Н., Абросов В.А. Быстрые методы статистической обработки и планирование эксперимента. – Изд-во Ленинградского ун-та, 1971. – 79 с.

3. Джонстон Дж. Эконометрические методы. – М., 1980. – 260 с.

4. Драхвелидзе П.Г., Марков Е.П. Delphi – среда визуального программирования. – СПб, 1996. – 468 с.

5. Езекиэл М, Фокс К. Методы анализа корреляций и регрессий. – М., 1996. – 314 с.

6. Мисевич К.М., Рященко С.В. Географическая среда и условия жизни населения Сибири. – Новосибирск., Наука, 1988.

7. Ускова О.Ф., Ермоленко Н.Н. Автоматическая классификация объектов окружающей среды с комплексом свойств // Экология. Экологическое образование. Нелинейное мышление. Сб. трудов III международной конференции из серии "Нелинейный мир". – М., 1998. – с. 244 – 249