УДК 316.1

КЛАСТЕРНАЯ ХАРАКТЕРИСТИЧЕСКАЯ СТРУКТУРА СОЦИУМА – ОСНОВА МАТЕМАТИЧЕСКОЙ СОЦИОЛОГИИ *

Колтунов Иосиф Анатольевич – кандидат физико-математических наук, старший научный сотрудник кафедры высшей математики и информатики механико-математического факультета

Самой современной областью математической обработки экспериментальных наблюдений, наиболее глубинной в теоретическом плане и наиболее обширной в плане практических приложений, является структурно-статистическая теория распознавания и классификации объектов (РИК). Математическая основа РИК – смесевая модель распределения признаков, построение отличительных характеристик и кластеризация (классификация) множества наблюдаемых объектов по этим характеристикам.
Применение РИК в социологии позволяет определить и построить отличительные характеристики членов общества, объективную структуру социума, а также наметить пути анализа динамики социальных структур.

Ключевые слова: смесевая модель вероятностных распределений признаков объектов; отличительные характеристики объектов социума (личностей); объективная кластерная структура социума; информативно эквивалентные матрицы наблюдений; динамическая теория развития социума.

Найсучаснішою областю математичної обробки експериментальних спостережень, найбільш глибинною у теоретичному плані і найбільш обширною в плані практичних додатків, є структурно-статистична теорія розпізнавання і класифікації об'єктів (РВК). Математична основа РВК - сумішева модель розподілу ознак, побудова відмінних характеристик і кластеризація (класифікація) безлічі спостережуваних об'єктів за цими характеристиками.
Застосування РВК в соціології дозволяє визначити і побудувати відмінні характеристики членів суспільства, об'єктивну структуру соціуму, а також намітити шляхи аналізу динаміки соціальних структур.

Ключові слова: сумішевих модель імовірнісних розподілів ознак об'єктів; відмінні характеристики об'єктів соціуму (особистостей); об'єктивна кластерна структура соціуму; інформативно еквівалентні матриці спостережень; динамічна теорія розвитку соціуму.

The most modern area of mathematical processing of the experimental observations, the most in-depth in theoretical perspective and the most extensive in terms of practical applications, is structural-statistical theory of recognition and classification of objects (RCO). The mathematical basis of the RCO is a blending model of distribution of features, building of distinctive characteristics and clustering (classification) of the set of observed objects by these characteristics. The use of RCO in sociology allows you to define and to build the distinctive characteristics of the members of society, the objective structure of society, as well as to identify the ways of analysis of the dynamics of social structures.

Keywords: *blending model of probability distributions of attributes of the objects, the distinctive characteristics of objects of society (individuals); objective cluster structure of society; informative equivalent matrixes of observation; dynamic theory of social development.*

С появлением компьютерной техники известные математические методы прикладной статистики получили реальное применение в социологических исследованиях для анализа опытных данных [1; 2; 3]. Окончился продолжительный, длящийся почти столетие, период предварительной подготовки и статистики, и социологии к совместной работе.

Предлагаемые в статье новые математические модели [4; 5], названные математической социологией, существенно расширяют и обобщают «аналитическую ветвь» в социологической науке, однако оставляют специалистам-социологам немало трудностей в сборе эмпирического материала для последующего анализа и, главное, в постановке жгучих социальных проблем.

* Мы представляем взгляд математика на потенциальные пути развития математической социологии (редакция Вестника)

И математикам, и социологам понятна важность и сложность наделения *объектов социума* (личностей) набором индивидуальных черт: физических, интеллектуальных, духовных, эстетических, морально-нравственных; набором определённых черт характера: инициативность, решительность, целеустремлённость, принципиальность; а также комплексом общественных признаков: социальный статус, образование, профессия, материальный достаток, семейное положение, отношение к трудовой деятельности.

Вышеуказанная проблема получения экспериментальных данных включает:
- сложность сбора **достоверных** данных о личностях;
- сложность преобразования качественных черт личности в количественные показатели, в том числе, возможность построения шкал балльности;
- сложность учёта динамики признаков личности в ходе эмпирического обследования и в последующем времени.

Актуальной практической целью статьи является разработка уникального комплекса компьютерных программ математической обработки социологической информации для анализа сложных современных общественных отношений на межгосударственном, межнациональном, межрелигиозном и межклассовом уровнях. Этот программный комплекс, основой которого будет новая математическая социология, основные разделы которой намечены в публикациях [4; 5], должен быть максимально автоматизированным и удобным в эксплуатации.

При условии, что проблема наделения числовыми признаками личностей в достаточной мере преодолена, построение моделей математической социологии, выполненное в данной работе, содержит решение двух алгоритмических задач и ввод в социологию новой **естественнонаучной парадигмы** в истолковании кластеров, находящихся в пространстве характеристик личностей социума:
- построение отличительных характеристик личностей;
- кластеризация социума на основе отличительных характеристик личностей;
- геометрическое, механическое и физическое истолкование кластеров в пространстве характеристик.

Решение этих трёх задач изложено в пунктах 1, 2, 3 предлагаемой работы. Полученное решение позволяет дать математическое определение характеристической структуры социума.

Краткое заключение статьи – это попытка наметить перспективы применения первых результатов нового научного направления в фундаментальной и прикладной социологии.

1. Построение отличительных характеристик личностей.

Термины социологии, привычные и понятные специалистам, переведём в известные математические понятия [1 – 5].

Социум – множество объектов-векторов $Z = \{Z_k\}_{k=1}^{n}$; объект социума – личность Z_k.

Личности – вектора, компоненты которых вещественные числа $Z_k = \{z_{ik}\}_{i=1,k=1}^{m,n}$.

Индивидуальные и общественные признаки личности – компоненты z_{ik} векторов Z_k.

Упорядочим по наименованиям все признаки личностей; тогда социум $Z = \{z_{ik}\}_{i=1,k=1}^{m,n}$ - прямоугольная порядка ($m \times n$) матрица, каждая строка которой соответствует одному признаку для всех личностей, и каждый столбец которой соответствует одной личности со всеми признаками.

Поставленная социологическая задача состоит в том, чтобы вместо характеризации личностей социума большим количеством m исходных признаков ограничиться намного меньшим количеством p ($p \ll m$) других показателей $\{a_{lk}\}_{l=1}^{p}$ для всех личностей $k = \{1; n\}$, не теряя при этом отличительной информации каждой личности от других личностей социума.

Назовём введённый новый показатель A_k личности Z_k, $A_k = \{a_{lk}\}_{l=1}^{p}$ **вектором отличительных характеристик личности** Z_k.

Прежде чем перейти к математическому формализму задачи, заметим, что в социологии обычно изучают социумы личностей, проживающих компактно на заданном территориально названном обширном земельном участке (район, город, государство, материк, планета в целом). Для дальнейших математических построений в этом замечании существенно лишь то, что количество личностей социума является достаточно большим с тем, чтобы число объектов обучающей выборки было репрезентативным [4, 5].

Проблема уменьшения показателей личности означает, что вместо задания социума матрицей признаков $Z=\{z_{ik}\}_{i=1,k=1}^{m,n}$ нужно задать его прямоугольной матрицей характеристик $A=\{a_{lk}\}_{l=1,k=1}^{p,n}$ порядка $(p\times n)$, где $p\square m$, т.е. найти подходящее по определённому критерию информативности операторное преобразование $Z\to A$.

Решать задачу поиска такого преобразования будем в два этапа. Найдём вспомогательную матрицу $Y=G(Z)$; $Y=\{y_{ik}\}_{i=1,k=1}^{m,n}$ той же структуры, что и матрица Z с помощью оператора G, обладающего следующими двумя свойствами.

1. Каждый элемент y_{ik} новой матрицы Y является функционалом только от элементов одного собственного столбца $\{z_{ik}\}_{i=1}^{m}$ матрицы Z:

$$y_{ik}=\Phi(z_{1k},z_{2k},...,z_{mk}). \tag{1}$$

2. Все элементы одной строки матрицы Z преобразуются одним и тем же функционалом Φ, т. е. функционал $\Phi=\Phi_i$ зависит только от номера признака, но не зависит от номера личности. Такое преобразование обобщает до нелинейных те линейные преобразования элементов матрицы, которые оставляют эту матрицу в семействе подобных матриц.

Определение. Матрица Y, полученная с помощью оператора G: $(Y=G[Z])$, обладающего свойствами 1 и 2, называется **информативно эквивалентной** матрице наблюдений Z.

Если бы мы ограничились функцией $y_{ik}=\varphi_i(z_{ik})$ вместо функционала $y_{ik}=\Phi_i(z_{1k},z_{2k},...,z_{mk})$, то оператор $Y=\Phi(Z)$ соответствовал бы согласованному по всем личностям изменению шкал наблюдения признаков этих личностей.

Напомним, что каждая строка матрицы Y соответствует преобразованному признаку $Y_i=\{y_{ik}\}_{k=1}^{n}$ для всех личностей, а каждый столбец соответствует личности со всеми её преобразованными признаками $Y_k=\{y_{ik}\}_{i=1}^{m}$.

Перейдём к следующему этапу преобразования $Z\to A$, а именно, найдём преобразование $A=H(Y)$. Для этого построим обратный, но обратимый оператор H^{-1}: $Y=H^{-1}(A)$. Пусть каждый элемент матрицы $Y=\{y_{ik}\}_{i=1,k=1}$ является ординатой некоторой точки $X_i=\{x_{iv}\}_{v=1}^{q}$, принадлежащей координатной гиперплоскости $X=\{X_v\}_{v=1}^{q}$ размерности q пространства $R^{q+1}=(X,Y)$. При этом ординаты всех точек одного признака с номером i для разных личностей имеют одинаковую «q-мерную абсциссу» $X_i=\{x_{iv}\}_{v=1}^{q}$. Вектор X_i назовём **граундфактором** признака с номером i.

Каждой личности поставим в соответствие такую q-мерную гиперповерхность $y=B_k(X)$, чтобы отображение $X\to y$ было однозначным. Далее, пусть каждой личности соответствует набор p параметров $A_k=(a_{1k},a_{2k},...,a_{pk})$ такой, что $B_k(X)$ имеют один и тот же функциональный вид, но отличаются только векторным параметром A_k:

$$y=B(A_k,X)$$

Если граундфакторы X_i для всех признаков известны, то оператор H^{-1} построен:

$$y_{ik}=B(A_k,X_i)=B_i(A_k);\quad Y=H^{-1}(A) \tag{2}$$

Приравнивая правые части равенств (1) и (2), получим **общее уравнение отличительных характеристик**:

$$G(Z) = H^{-1}(A)$$

Решая это уравнение, имеем всю матрицу отличительных характеристик личностей социума:

$$A = H \cdot G(Z) \tag{3}$$

Чтобы построение преобразования $Z \to A$ стало доступным для практического использования в социологии, приведём два примера построения отличительных характеристик объектов с простыми преобразованиями G и H, рассмотренных в работе [5] в рамках выполнения проекта по дистанционному зондированию Земли.

1.1. Линейный вариант задачи

Выполним линейное калибровочное преобразование элементов матрицы Z с коэффициентами калибровки b_{0i} и b_{1i}, зависящими от признаков, но не зависящими от объектов:

$$y_{ik} = b_{0i} + b_{1i} \cdot z_{ik} \tag{4}$$

С другой стороны, предполагаем, что преобразованные признаки y_{ik} есть ординаты точек на гиперплоскостях $Y_k = A_k \cdot X$ объектов:

$$y_{ik} = a_{0k} + \sum_{l=1}^{p} a_{lk} \cdot x_{il} \tag{5}$$

В выражении (5) коэффициенты $A_k = \{a_{lk}\}_{l=1}^{p}$ являются векторами характеристик объекта с номером k; точка $X_i = \{x_{il}\}_{l=1}^{p}$ координатной гиперплоскости X размерности p есть граундфактор признака с номером i.

Приравняв правые части равенств (4) и (5), получим уравнение характеристик:

$$b_{0i} + b_{1i} \cdot z_{ik} = a_{ok} + \sum_{l=1}^{p} a_{lk} \cdot x_{il}, \tag{6}$$

в котором кроме искомых характеристик A_k присутствуют «мешающие» неизвестные параметры – калибровочные коэффициенты b_i и граундфакторы X_i.

Равенство (6) представляет собой систему $(m \times n)$ уравнений относительно $(p+2) \cdot m + (p+1) \cdot n$ неизвестных. При заданном p и достаточно больших m и n в системе (6) число неизвестных не превышает числа уравнений, однако вследствие её нелинейности эта система не имеет единственного решения.

Единственность решения достигается введением так называемых **опорных признаков**. На опорных признаках калибровочные коэффициенты и граундфакторы полагаются известными. Если количество опорных признаков взять равным числу искомых характеристик у каждого объекта, то получаем n систем $p+1$ линейных уравнений относительно $p+1$ неизвестных характеристик. При невырожденности матриц этих систем уравнений характеристики для всех объектов могут быть получены. Для проверки адекватности моделей (4) и (5) относительно наблюдаемой матрицы признаков $Z = \{z_{ik}\}$ в уравнении (6) при полученных характеристиках $\{A_k\}_{k=1}^{n}$ решим $(m - p - 1)$ систем из n уравнений относительно $p+2$ неизвестных граундфакторов и калибровочных коэффициентов для признаков, не являющихся опорными. При условии, что для m (число признаков) и $p+1$ (число характеристик) выполняется неравенство $m \geq p+1$, все системы уравнений переопределены. Если невязки во всех уравнениях оказались равными нулю, можно сделать вывод об адекватности моделей преобразований (4) и (5) относительно наблюдения $Z = \{z_{ik}\}$ и проблема полностью решена.

Однако на практике измерения признаков z_{ik} не являются безошибочными. Относительно ошибок в измерении признаков сделаем два вполне допустимых предположения: а) ошибка в z_{ik} является нормально распределённой случайной величиной, не зависящей от объекта; б) среднеквадратичная величина случайной ошибки для всех признаков априори известна и равна σ_i $(i = 1, 2, ..., m)$.

Укажем основные пункты алгоритма в решении задачи вычисления характеристик объектов при сделанных выше предположениях:

1) Применение метода максимального правдоподобия для оценки неизвестных параметров A, B, X системы уравнений (6) приведёт к методу наименьших квадратов [6; 7] :

$$\min_{a,b,x} \sum_{i=1}^{m} \frac{1}{b_{1i}^2 \cdot \sigma_i^2} \sum_{k=1}^{n} \left[b_{0i} + b_{1i} \cdot z_{ik} - a_{0k} - \sum_{l=1}^{p} a_{ik} \cdot x_{il} \right]^2 \tag{7}$$

2) Как указывалось выше, для получения единственного решения этой задачи нужно использовать метод опорных признаков. При назначении $p+1$ опорных признаков имеет смысл учесть два условия: а) чтобы величина $\Delta_i^2 = b_{1i}^2 \cdot \sigma_i^2$ была как можно меньше; б) чтобы попарные ковариационные моменты у опорных признаков были также как можно меньше – максимальная независимость.
3) Для удобства записи формул предполагаем, что опорными являются признаки с первыми номерами $i = 1, 2, ..., p+1$. Калибровочные коэффициенты выберем равными: $b_{0i} = 0$, $b_{1i} = 1$, $i = 1, 2, ..., p+1$. Это означает, что опорные признаки не калибруются. Граундфакторы этих признаков положим следующими. Признак с номером $p+1$ имеет граундфактором начало координат: $0, 0, ..., 0$; признаки с номерами $i = 1, 2, ..., p$ имеют граундфакторами точки «1» на осях с номерами i: $00, ..., 1, ..., 0$. Это означает, что на всех осях введены **эталонные единицы измерения.**
4) При вышеназванном выборе опорных признаков формула (7) переписывается следующим образом:

$$\chi^2_{\min} = \min_{a,b,X} \sum_{k=1}^{n} \left\{ \sum_{i=1}^{p} \left[\frac{1}{\sigma_i^2} \left(z_{ik} - a_{0k} - a_{1k} \right)^2 + \frac{1}{\sigma_{p+1}^2} \left(z_{p+1,k} - a_{0k} \right)^2 + \right. \right.$$
$$\left. \left. + \sum_{i=p+2}^{m} \frac{1}{b_{1i}^2 \cdot \sigma_i^2} \left(b_{0i} + b_{1i} \cdot z_{ik} - a_{0k} - \sum_{l=1}^{p} a_{ik} \cdot x_{il} \right)^2 \right\} \right. \tag{8}$$

5) В качестве начального приближения в решении задачи оптимизации (8) следует взять расчёт характеристик объектов, граундфакторов и калибровочных коэффициентов признаков из уравнения (6) методом опорных признаков в предположении отсутствия ошибок ($\sigma_i = 0$) в измерениях z_{ik}.
6) Если модели калибровочного и объектно-характеристического преобразований в уравнении характеристик (6) адекватны наблюдениям z_{ik}, то полученное в результате оптимизации значение квадратичного функционала $\chi^2_{\min}$ будет удовлетворять неравенству:

$$\chi^2_{\min} / N \le 1,$$

где $N = n \times m$ - число уравнений в системе (6).
7) Модель (5) можно попытаться сжать, т.е. уменьшить значение p. Если же неравенство пункта 6 не выполняется, модель нужно расширить, увеличив значение p. После этого все пункты 1 – 7 решения задачи расчёта характеристик объектов необходимо повторить.

1.2. Нелинейное калибровочное преобразование.

Калибровочное преобразование $Y = B_i(Z)$ элементов z_{ik} матрицы наблюдений Z, зависящее только от признаков, но не зависящее от объектов, будучи взаимно однозначным, не изменяет классификационной информативности откалиброванных признаков y_{ik}. Поэтому оно может быть нелинейным. Понятно, что равномерность измерительных шкал признаков – понятие относительное, точнее, условное. Более того, в вычислительном плане калибровочное преобразование G стоит сделать сколь угодно сложным с тем, чтобы объектно-характеристическое преобразование H в (6) максимально упростить по количеству характеристик объектов. В практической реализации метода построения отличительных характеристик можно предложить следующий компромисс: калибровочное

преобразование нелинейно, а калибровочные параметры являются линейными коэффициентами при нелинейных, но монотонных функциях $\varphi_j(z_{ik})$, $j=1,2,...,r$:

$$Y = B_i(Z) \Leftrightarrow y_{ik} = \sum_{j=1}^{r} b_{ji} \cdot \varphi_j(z_{ik}) \tag{9}$$

Тогда уравнение характеристик примет следующий вид:

$$\sum_{j=1}^{r} b_{ji} \cdot \varphi_j(z_{ik}) = a_{0k} + \sum_{l=1}^{p} a_{ik} \cdot x_{il} \tag{10}$$

Система уравнений (10) включает $(m \times n)$ уравнений с числом неизвестных $(1+p)\cdot n + (r+p)\cdot m$.

Пусть линейное калибровочное преобразование (4) входит частным случаем в нелинейное (9), т.е. $\varphi_1(z_{ik}) = 1$; $\varphi_2(z_{ik}) = z_{ik}$, а остальные функции $\varphi_j(z_{ik})$ будут произвольными непрерывно дифференцируемыми монотонными функциями аргумента z_{ik} (например, степенными с нечётными показателями степени). Для получения единственного решения системы уравнений (10), как и прежде, воспользуемся методом опорных признаков. Число опорных признаков по-прежнему равно числу характеристик объектов $(p+1)$. Калибровочные коэффициенты зафиксируем в следующем виде:

$$b_{1i} = 0; \qquad \{b_{j1} = 0\}_{j=3}^{r}; \quad b_{2i} = 1$$

Это означает, что опорные признаки , как и прежде, не калибруются. Граундфакторы опорных признаков берутся: X_1 в начале координат и X_i $(i = 2,3,...,p+1)$ отсекают единичные отрезки на всех осях координатной гиперплоскости. Вычисления ведутся по вышеуказанной схеме (1 -7). Для удобства записи последующей формулы (11) введём обозначение частной производной:

$$\frac{\partial}{\partial z_{ik}}\left[B_i(Z)\right] = y'_{ik}(b)$$

При неточном измерении признаков z_{ik} метод максимального правдоподобия приводит к следующей задаче оптимизации:

$$\chi^2_{\min} = \min_{a,b,X} \sum_{k=1}^{n} \left\{ \sum_{i=1}^{p} \left[\frac{1}{\sigma_i^2}\left(z_{ik} - a_{0k} - a_{ik}\right)^2 \right] + \frac{1}{\sigma_{p+1}^2}\left(z_{p+1,k} - a_{0k}\right)^2 + \right.$$
$$\left. + \sum_{i=p+2}^{m} \frac{1}{\left[y'_{ik}(b)\cdot\sigma_i\right]^2} \left[\sum_{j=1}^{r} b_{ji}\cdot\varphi_j(z_{ik}) - a_{0k} - \sum_{l=1}^{p} a_{lk}\cdot x_{il} \right]^2 \right\} \tag{11}$$

При этом, если выполняется для функционала $\chi^2_{\min}$ (11) неравенство

$$\chi^2_{\min} / N \le 1, \tag{12}$$

то модели для калибровочного G (9) и для объектно-характеристического H (5) преобразований полагаем адекватными наблюдениям z_{ik} в уравнении характеристик (10). Если в проверяемом неравенстве (12) стоит обратный знак, то обе модели нужно попытаться расширить, последовательно увеличивая r и p.

На более детальном статистическом анализе адекватности моделей останавливаться не будем: такой анализ потребует отдельного исследования.

2. Кластеризация социума на основе отличительных характеристик личностей.

Методы кластеризации социума на заданных признаках личностей неоднократно применялись в социологии [1, 2]. Однако использование этих методов в социологической практике ограничивалось

решением выборочных частных задач, имело черты маргинальности и не могло быть направлено на решение фундаментальных проблем по нескольким причинам.

1. Кластеризация проводилась по небольшому количеству заранее выбираемых исходных признаков.
2. Результаты кластеризации носили эвристический характер и не имели критериев надёжности и достоверности с математической точки зрения.
3. Определение структуры социума на основе такой кластеризации имело бы весьма неоднозначный формат в зависимости от выбранных исходных признаков личностей. Таким образом, понятие кластера (класса) в социуме осталось бы субъективным и неопределённым.

Значительный объём выбора в качестве исходных признаков личности (например, в несколько десятков) с добавлением к этому объёму по необходимости дополнительных признаков; затем сжатие пространства этих признаков до небольшого (в первом десятке) числа отличительных характеристик; и, наконец, кластеризация социума по этим характеристикам, содержащим в себе всю необходимую информацию об исходных признаках – вот путь социологического исследования, предлагаемый в этой работе.

Прежде чем расчленять социум на кластеры (в определённом смысле на «однородные группы» личностей), введём математическое понятие такой «группы» на основе теории вероятностей [9]. Согласно классической предельной теореме Ляпунова, если наблюдаемое случайное событие является суммарным следствием большого числа независимых случайных «равновеликих» событий, то оно имеет закон распределения вероятностей, близкий к нормальному закону Гаусса. Итак, математическая модель однородной группы личностей социума такова: вероятностное распределение характеристик личностей одного кластера является гауссовым, так как личности одного кластера – это последовательность наблюдений случайной величины как следствия многих причин указанного выше постулируемого характера.

Социум в целом является последовательностью независимых случайных величин – характеристик личностей из различных однородных групп с различными распределениями Гаусса. Согласно теореме о полной вероятности социум имеет плотность распределения в виде смеси гауссовых распределений [10 - 14].

Прежде чем перейти к математическим формулировкам, согласуем буквенные обозначения и определения предлагаемой работы с цитируемыми авторскими статьями [4; 5].

Наблюдение рандомизированной обучающей выборки характеристик (а не признаков) личностей из социума представляет собой случайную матрицу $Z = \{z_{ik}\}_{i=1,k=1}^{p,n}$

Наблюдаемый случайный объект Z_k имеет плотность распределения вероятностей

$$\varphi_k = \sum_{r=1}^{v} q_r \cdot N(Z_k, A_r, G_r), \tag{13}$$

где v - число гауссовых компонент (кластеров в социуме); q_r - вероятность объекта попасть в кластер с номером r; A_r - вектор средних значений кластера с номером r; G_r - ковариационная матрица кластера с номером r;

$N(Z_k, A_r, G_r)$ - плотность распределения Гаусса с вектором средних значений A_r и матрицей ковариаций G_r:

$$N(Z_k, A_r, G_r) = \frac{1}{(2\pi)^{\frac{m}{2}} |G_r|^{\frac{1}{2}}} \cdot \exp\left[-\frac{1}{2} (Z_k - A_r)^T \cdot (G_r)^{-1} \cdot (Z_k - A_r) \right] \tag{14}$$

На вероятности q_r наложено ограничение

$$\sum_{r=1}^{v} q_r = 1, \quad q_r > 0 \tag{15}$$

2.1. Вывод итерационных формул EM-алгоритма

В предлагаемом варианте алгоритма кластеризации объектов полагаем наблюдаемые объекты Z_k независимыми. Тогда для плотности распределения $\Phi(z)$ матрицы наблюдений Z справедливо представление:

$$\Phi(Z) = \prod_{k=1}^{n} \varphi_k(Z_k) \tag{16}$$

или для *логарифмической функции правдоподобия* (ЛФП):

$$l(\alpha) = \ln \Phi\left(Z^0\right) = \sum_{k=1}^{n} \ln \varphi_k\left(Z_k^0\right), \tag{17}$$

где Z^0 - реализация случайной величины Z.

Под совокупностью регулярных параметров α ЛФП $l(\alpha)$ понимаем q_r, A_r, G_r для всех значений r. При построении уравнений правдоподобия [5] необходимо учесть ограничение (15) на коэффициенты q_r (метод Лагранжа). Неизвестное количество гауссовых компонент (кластеров) v на первом этапе оценки параметров α считаем заданным, однако к этой проблеме вернёмся в пункте 2.2 .

Необходимые условия экстремума ЛФП $l(\alpha)$ (17) с учётом ограничения (15) приводят к системе уравнений правдоподобия, записанной в форме, удобной для итераций.

$$\omega_r^k = \frac{q_r \cdot N\left(Z_k^0, A_r, G_r\right)}{\varphi_k\left(Z_k^0\right)}; q_r = \frac{U_r}{W}; U_r = \sum_{k=1}^{n} \omega_r^k; W = \sum_{r=1}^{v} U_r$$

$$A_r = \frac{\sum_{k=1}^{n} \omega_r^k \cdot Z_k^0}{U_r}; G_r = \frac{\sum_{k=1}^{n} \omega_r^k D_r^k}{U_r}; \tag{18}$$

$$D_r^k = \left(Z_k^0 - A_r\right)\left(Z_k^0 - A_r\right)^T$$

Получив оценки $\alpha = \left(q_r, A_r, G_r\right)$ параметров оптимизации ЛФП $l(\alpha)$, в качестве функции принадлежности личности Z_k кластеру с номером r применим известную формулу гипотез Байеса:

$$P\left(H_r \mid Z_k\right) = \frac{q_r \cdot N\left(Z_k, A_r, G_r\right)}{\sum_{k=1}^{v} q_r \cdot N\left(Z_k, A_r, G_r\right)} \tag{19}$$

в предположении, что количество кластеров v социума известно.

2.2. Оценивание количества кластеров социума.

В работах [15,16,17] приводится целое семейство информационных методов оценки числа кластеров социума, которое называют методами «принципа оштрафованного правдоподобия».

Самые известные и простые представители этого семейства: Акайке информационный критерий (АИК) и Байесовский информационный критерий (БИК). Предлагается простейший информационный критерий БИК, который часто называют критерием Шварца. Формула критерия Шварца:

$$БИК(v) = 2l(\alpha) - \Omega \cdot \ln(n)$$

с числом компонент v, числом Ω искомых параметров α распределения , числом объектов обучающей выборки n, ЛФП $l(\alpha)$. Максимум БИК(v) даёт «истинное» количество кластеров v социума.

В работе [5] рассмотрен метод оценки числа кластеров, основанный на вычислении и применении информационной матрицы Фишера (ИМФ). Хотя этот метод требует большего объёма компьютерных вычислений, знание его может быть полезным для расчёта надёжности и достоверности характеристической кластеризации личностей социума. Если наблюдаемая случайная матрица Z объектов имеет плотность распределения характеристик $\Phi(Z)$, то ИМФ для неизвестных параметров α плотности распределения наблюдения называется интеграл:

$$I(\alpha) = \int_{B(Z)} \nabla_\alpha\left[\ln \Phi(Z)\right] \cdot \nabla_\alpha^T\left[\ln \Phi(Z)\right] \cdot \Phi(Z)\, dZ$$

$$B(Z) = \left\{Z : \Phi(Z) \neq 0\right\}, \tag{20}$$

где символ ∇ означает градиент функции, взятый по всем её неизвестным параметрам α; ∇_α^T означает транспонирование вектора ∇_α.

При независимости наблюдаемых объектов Z_k ($k=1,...,n$) для ИМФ выполняется равенство:

$$I(\alpha)=\sum_{k=1}^{n}\int_{b(Z)}\nabla_{\alpha}\left[\ln\varphi_k(Z,\alpha)\right]\cdot\nabla_{\alpha}^{T}\left[\ln\varphi_k(Z,\alpha)\right]\cdot\varphi_k(Z,\alpha)\,dZ \qquad (21)$$

где область $b(Z)\subset R^m$ есть проекция области $B(Z)$ в $R^m\subset R^{(n\times m)}$.

Имея ИМФ, можно оценить точность оценок метода максимального правдоподобия (ММП) для параметров α, полученных ранее в итерационном процессе. Известна асимптотическая формула [5]

$$\lim_{n\to\infty} I(\alpha)\cdot K(\alpha)=E,$$

где E - единичная матрица, $K(\alpha)$ - ковариационная матрица оценок ММП. На диагонали $K(\alpha)$ стоят квадраты среднеквадратичных ошибок (СКО) в ММП, в том числе квадраты СКО коэффициентов q_r. Обозначим СКО q_r символом Δq_r. Отношение $q_r/\Delta q_r=\xi q_r$ назовём значимостью q_r. Выбрав определённый порог δ, назовём кластер значимым, если $\xi q_r\ge\delta$, и незначимым, если $\xi q_r<\delta$.
Ликвидируя незначимые кластеры, получаем оценку для параметра v: v равняется числу значимых кластеров социума.

3. Геометрическое, механическое и физическое истолкование кластеров в пространстве характеристик

Полученные параметры $\alpha=\left\{v,\left(q_r,A_r,G_r\right)\right\}_{r=1}^{v}$ являются параметрами ковариационных эллипсоидов гауссовых распределений кластеров в p-мерном пространстве характеристик личностей. При этом параметры A_r и G_r указывают местоположение, размер, форму и ориентацию каждого эллипсоида в пространстве характеристик, параметры q_r являются нормированными массами эллипсоидов; параметр v - число эллипсоидов в пространстве характеристик личностей социума.
Определение. Совокупность всех параметров $\alpha=\left\{v,\left(q_r,A_r,G_r\right)\right\}_{r=1}^{v}$ назовём **характеристической структурой социума.**

Создание такого многомерного геометрического образа социума позволяет провести сравнение различных социумов по их географическому местоположению и проследить их изменение во времени. Если априори выработаны предпочтения социологов в бытии и в сознании личностей одного социума по сравнению с другим, можно визуально отметить у каждого социума лишние и недостающие кластеры, а также произвести расчёт пересечений и разниц сопоставимых кластеров из различных социумов.

Проводить визуальный анализ в P - мерном пространстве характеристик вполне возможно с помощью вывода на дисплей компьютера многомерных эллипсоидов в формате PD (P > 3). Вывод в этом формате означает получение на экране всех трёхмерных проекций многомерного эллипсоида [18].

Изменение характеристической структуры социума во времени может происходить **самопроизвольно** либо с помощью **управления**. Кластеры социума будут перемещаться, изменяться в размерах, форме и пространственной ориентации, пересекаться, взаимно поглощаться, исчезать и появляться. Тело P - мерного эллипсоида во многом аналогично материальному телу. Это тело имеет **массу** и **плотность.** Перемещаясь в пространстве характеристик, оно имеет **скорость**, **импульс** и **кинетическую энергию**. Так как **частицы** (личности) внутри этого тела тоже движутся (либо хаотично, либо упорядоченно), эллипсоид обладает **внутренней энергией** и **термодинамическими параметрами.** В зависимости от характера движения личностей внутри кластера (хаос либо упорядоченность) можно ввести понятие **фазового состояния** кластера (**твёрдое тело**, **жидкость**, **газ**, **плазмоид**). Предположение об «однородности групп» личностей кластера не является незыблемым. Не исключено, что кластер с определённого момента времени приобретает внутреннюю структуру, т.е. включает несколько эллипсоидов.

Если рассматривать задачу кластеризации как алгоритм таксономии (алгоритм разделения социума на изолированные группы личностей), то можно убедиться, что таксоны будут сложной неправильной не эллипсоидальной формы. Это ещё раз указывает на возможность более детального расчленения каждого кластера личностей.

Заключение.

Причина самопроизвольного (стихийного) изменения во времени структуры социума пока не ясна. Очевидно, что между телами различных кластеров и внутри тел кластеров присутствуют силы **полевого** взаимодействия в характеристическом пространстве. Математические модели («физические законы») этих сил не известны. Локальный во времени учёт изменений структуры социума возможен с помощью динамической модели распознавания [19; 20]; и это в настоящее время всё, что можно предложить.

Последующие совместные усилия математиков, естественников и гуманитариев, проведенные исследования после изучения развития во времени характеристических структур социумов, дадут материал для построения новых математических моделей в пространстве характеристик личностей. Не исключено, что это будут аналоги классической или квантовой механики. Более того, так как социум является **открытой системой**, в которой наблюдаются **необратимые процессы**, решающее слово в будущей математической социологии скажет синергетика [21].

Чем и когда завершится построение новой науки? Начать имеет смысл сегодня.

Пользователями этого программного продукта станут специалисты различных профессий – технических и гуманитарных – независимо от компьютерной квалификации и математической подготовки.

Необходимым средством достижения поставленной цели является организация группы разработчиков комплекса; профессия этих разработчиков – прикладная математика, информатика и социология.

Литература:

1. Ядов. В. А. Стратегия социологического исследования. Описание, объяснение, понимание социальной реальности / А.В. Ядов / – М. Добросвет, 1999. – 596 с.

2. Орлов А. И. Теория принятия решений / А.И. Орлов / – М. Экзамен, 2006. – 576 с.

3. Кондович В. Ю.. Соціологія (схеми, таблиці, слайди) / В.Ю. Кондович / - Чернігів, ЧДТУ, 2012, - 394 с.

4. Боровик А. И. Построение субоптимального метода РИК / Боровик А. И.,Деркач С. В.,Колтунов И. А. // Вестник ХНУ им. В.Н. Каразина, № 926. – Х: Наука, 2010. – с. 61-75

5. Колтунов И. А. Новые статистические методы РИК для автоматического дешифрирования дистанционных наблюдений. / Колтунов И. А., Думин А. Н., Катрич В. А., Наумов Р. Р. // Вестник ХНУ им. В. Н. Каразина, № 966. – Х: Наука – 2011 – с. 37- 49.

6. Линник Ю. В. Метод наименьших квадратов и основы математико-статистической теории обработки наблюдений / Ю.В. Линник / – М. : Б. и., - 336 с. – 1962.

7. Рао К. Р. Линейные статистические методы и их применение / К.Р. Рао – М.: Наука, 1968 – 548 с.

8. Крамер Г. Математические методы статистики [пер с англ.] / Г. Краммер / - М., 1975. – 648 с.

9. Гнеденко Б. В. Курс теории вероятностей / Б.В. Гнеденко / - М., 1965. – 400 с.

10. Колтунов И. А. Статистическая классификация наблюдений с полимодальными распределениями / Колтунов И. А., Кондратьева Л. М., Монастырёв А. П. // Статистические проблемы управления. Вильнюс. – 1987. - №78. – с. 83-121.

11. McLachlan G.J. The EM–Algorithm and Extensions / G.J.McLachlan, T.Krishnan / - New York:Wiley. - 1997.

12. Колтунов И. А. Применение смесевых моделей вероятностных распределений для обработки изображений и распознавания образов / И.А. Колтунов / – Изд-во БелГУ, 2004. – с. 122.

13. Koltunov A. Mixture density separation as a tool for high-quality interpretation of multisource remote sensing data and related issues / Koltunov A., Ben-Dor E. // International Journal of Remote Sensing, 2004. – V 25, PP. 3275-3299.

14. Koltunov Y., et al. Detection and recognition of objects by multispectral sensing. / Y. Koltunov // US Patent, 6837617, January, 2005. – 25 P.

15. Schwarz G. Estimation the dimension of a model / G. Schwarz // The Annals of Statistics. – 1978. – V. 6(2). – P. 461-464.

16. Kass R. E. Bayes factors / Kass R. E., Raftery A. E. // Journal of American Statistical Association. – 1995. – V. 90. – P. 773-795.

17. Fraley C. How many clusters? Which clustering method? Answers via model-based cluster analysis / Fraley C., Raftery A. E.// The Computer Journal. – 1998. – V. 41 (8). – P. 578-588.

18. Колтунов И. А. Алгоритмы программ для визуального анализа исходных наблюдений и результатов их обработки в проекте компьютерного комплекса РИК ДДН. / Колтунов И. А., Нерода И. В. // Современные проблемы математики и её приложения в естественных науках и информационных технологиях. Тезисы докладов межд.конф. Харьков. 2012. – с. 61.

19. Koltunov Y. Dynamic Detection Model and its application for perimeter security, intruder detection, and automated target / Koltunov Y., Koltunov A. //In: Infrared Technology and Applications XXIX, Proc. SPIE, 2003. – V. 5074, PP. 777-787.

20. Koltunov A. On timeliness and accuracy of wildfire detection by the GOES WF- ABBA algorithm over California during the 2006 fire season. Remote Sensing of Environment / Koltunov A., Ustin S. L., Prins E. M. / - 127 (2012) PP. 194-209.

21. Пригожин И. Порядок из хаоса. Новый диалог человека с природой / Пригожин И., Стенгерс И. / - М. 2003., – с. 310.