DOI: 10.34020/2073-6495-2020-1-255-267

УДК 311, 591.2

ТИПОЛОГИЧЕСКАЯ ГРУППИРОВКА НА ОСНОВЕ ДЕКОМПОЗИЦИИ СМЕСЕЙ ВЕРОЯТНОСТНЫХ РАСПРЕДЕЛЕНИЙ

Исмайылова Ю.Н., Хрущев С.Е.
Новосибирский государственный университет
экономики и управления «НИНХ»
E-mail: ismaiylova@gmail.com, s.e.hrushchev@edu.nsuem.ru

Смесь вероятностных распределений является математической моделью, которая позволяет адекватно описывать неоднородные данные. Задачей разделения смесей или задачей декомпозиции называется задача оценивания неизвестных параметров смешивающихся распределений. Несмотря на адекватность описания неоднородных данных, декомпозиция смесей представляет собой отдельную проблему, ввиду большого количества параметров, подлежащих оцениванию. В статье осуществлены историческая периодизация, систематизация и критический сравнительный анализ существующих методов и алгоритмов декомпозиции смесей вероятностных распределений, выявлены возможности и ограничения их применения для анализа реальных совокупностей. На основе существующих алгоритмов предложена методика разделения смесей произвольного известного количества вероятностных распределений и дальнейшей типологической группировки реальных социально-экономических совокупностей. В отличие от существующих методик предложен способ вычисления пороговых значений для определения границ типов и числа компонент смеси в случаях, когда оно неизвестно. На основе предложенной методики осуществлена типологизация субъектов Российской Федерации по уровню безработицы.

Ключевые слова: разделение смесей, метод моментов, метод максимального правдоподобия, EM-алгоритм, типологическая группировка, уровень безработицы.

TYPOLOGICAL GROUPING BASED ON DECOMPOSITION OF PROBABILITY DISTRIBUTIONS MIXTURES

Ismaiylova Yu.N., Khrushchev S.E.
Novosibirsk State University of Economics and Management
E-mail: ismaiylova@gmail.com, s.e.hrushchev@edu.nsuem.ru

A mixture of probability distributions is a mathematical model that allows to describe heterogeneous data. The task of separating mixtures or decomposition is the task of estimating the unknown parameters of miscible distributions. Despite the adequacy of the description of heterogeneous data, the decomposition of mixtures is a separate problem, due to the large number of parameters to be evaluated. The article carries out historical periodization, systematization, and a critical comparative analysis of existing methods and algorithms for decomposition of mixtures of probability distributions, identifies the possibilities and limitations of their application for the analysis of real populations. Based on existing algorithms, a method for separating mixtures of an arbitrary known number of probability distributions and a further typological grouping of real socio-economic aggregates is proposed. Unlike existing methods, a method for calculating threshold values to

determine the boundaries of types and the number of components of the mixture, in cases where it is unknown, is proposed. Based on the proposed methodology, a typology of the subjects of the Russian Federation by the level of unemployment in the Russian Federation is carried out.

Keywords: mixture separation, moments method, maximum likelihood method, EM-algorithm, typological grouping, unemployment rate.

1. ВВЕДЕНИЕ

В условиях быстрого турбулентного развития современного общества необходимо адекватно реагировать на политические, экономические и социальные изменения и сдвиги. Для этого необходимо обладать полной, достоверной статистической информацией о различных социальных и экономических процессах, которые, как правило, описываются смесями вероятностных законов, что приводит к необходимости знания значений параметров смесей распределений, т.е. к необходимости решения задачи о разделении или декомпозиции смесей вероятностных распределений. Декомпозиция смесей распределений позволяет разбивать всю исходную совокупность исследуемых объектов на однородные группы, что является одной из важнейших задач любого статистического исследования.

Основными методами выделения однородных групп в рамках совокупностной концепции являются типологическая группировка, портфельный анализ, историческая и параллельная периодизация [8, 9]. Но если рассматривать однородность объектов как возможность описания всех единиц частной совокупности одним вероятностным распределением, то при помощи методов декомпозиции смеси возможно разбивать исходную совокупность на однородные в этом смысле группы. Поэтому синтез типологической группировки и методов разделения смесей вероятностных распределений представляется актуальным направлением развития методологии в рамках совокупностной концепции. Задача разбиения исходной совокупности исследуемых объектов на однородные подгруппы является приоритетной при анализе данных любой природы: технических, биологических, социально-экономических.

2. ПОСТАНОВКА ЗАДАЧИ РАЗДЕЛЕНИЯ СМЕСЕЙ ВЕРОЯТНОСТНЫХ РАСПРЕДЕЛЕНИЙ

Предположим, что неоднородность совокупности порождается двумя или более законами распределения вероятностей и нам необходимо данную выборку разделить на однородные подгруппы. Задача разделения выборки на однородные подгруппы называется задачей разделения (декомпозиции) смеси и включает в себя несколько подзадач: оценивание параметров смесей, построение правила для определения совокупности, к которой принадлежит заданный элемент. В зависимости от порядка решения этих подзадач меняются методы их решения. Оценив сначала параметры смеси, можно на основе известных методов (например, дискриминантный анализ, байесовский наивный классификатор и др.) сформулировать решающее правило, таким образом, решив вторую подзадачу. Решение задачи в этом порядке требует у исследователя знания типов распределений совокупностей, вхо-

дящих в смесь, что позволяет применять классические методы, такие как метод моментов, метод максимального правдоподобия.

С другой стороны, группировка элементов по группам (при помощи, например, методов кластерного анализа) позволяет оценить параметры смесей распределения стандартными методами.

В данной работе будет рассматриваться задача первого типа, математическая постановка которой выглядит следующим образом.

Пусть $\vec{x} = (x_1, x_2, ..., x_n)$ – наблюдаемое значение случайной выборки $\vec{X} = (X_1, X_2, ..., X_n)$, в которой X_1, X_2, …, X_n – независимые одинаково распределенные случайные величины с функцией распределения, тогда:

$$F(x) = \sum_{i=1}^{k} p_i \cdot F_{\theta_i}(x), \tag{1}$$

где $i = 1, \ldots, k$; $p_i \geq 0$ – смешивающие вероятности, удовлетворяющие условию $p_1 + p_2 + \ldots + p_k = 1$; $F_{\theta_i}(x)$ – функция распределения некоторой случайной величины, известным образом зависящая от неизвестного параметра θ_i, который может быть как скалярным, так и векторным.

Задача разделения (декомпозиции) смеси состоит в поиске оценок неизвестных параметров $\vec{\theta} = (p_1, ..., p_k, \theta_1, ..., \theta_k)$ размерностью d, где $d = 2k - 1$, в случае скалярного параметра θ_i, и $d = (m + 1)k - 1$, в случае, если θ_i – векторный параметр размерности m.

В случае смеси дискретных распределений формулу (1) можно записать как

$$\mathbb{P}(X = m) = \sum_{i=1}^{k} p_i \cdot \mathbb{P}_{\theta_i}(X = m),$$

где $\mathbb{P}_{\theta_i}(X = m)$ – вероятность принятия значения m случайной величиной с дискретным распределением, зависящим от параметра θ_i.

Если же в смеси участвуют только абсолютно непрерывные распределения, то формулу (1) удобнее использовать в следующем виде:

$$p(x) = \sum_{i=1}^{k} p_i \cdot p_{\theta_i}(x),$$

где $p_{\theta_i}(x)$ – плотность распределения абсолютно непрерывной случайной величины, зависящей от параметра θ_i.

3. ОБЗОР СУЩЕСТВУЮЩИХ МЕТОДОВ РАЗДЕЛЕНИЯ СМЕСИ ВЕРОЯТНОСТНЫХ РАСПРЕДЕЛЕНИЙ

Историю исследования задачи разделения смеси можно рассматривать с двух точек зрения: с точки зрения развития каждого из методов или же во временном разрезе. Рассматривая развитие методического аппарата для решения задачи декомпозиции смеси как исторического процесса, можно выделить четыре этапа. Сами этапы и их описание представлены в табл. 1.

Метод моментов. Одним из методов нахождения оценок неизвестных параметров распределения является метод моментов. Идея метода состоит в замене теоретических моментов их выборочными аналогами. Остановимся на нем более подробно. Пусть X_1, X_2, …, X_n – выборка объема n из некоторого распределения $\mathcal{F}_\theta$, которое известным образом зависит от

Таблица 1

Историческая периодизация развития теории разделения смесей вероятностных распределений

Этап	Основные ученые, внесшие значительный вклад в развитие методического аппарата	Характеристика этапа
1-й этап. 1897–1940 гг.	К. Пирсон, Б. Стрёмгрен, К. Бюрро	Первая формулировка задачи о разделении смеси. Решение задачи для частных случаев
2-й этап. 1940–1970 гг.	В.Ю. Урбах, К. Бхаттачария, Н. Дэй, В. Хассельблад, С. Джон	Более общая постановка задачи. Исследование вопроса о существовании решения в общей постановке. Решение задачи для более широкого класса распределений. Модификация стандартных методов для решения задачи разделения смеси. Исследование свойств решений
3-й этап. 1970–1985 гг.	М.И. Шлезингер, Н. Дэй, Б.Эфрон, А. Дэмпстер, Н. Лэйрд, Д. Рубин	Формулировка ЕМ-алгоритма. Исследование его свойств. Применение ЕМ-алгоритма для моделирования смесей, а также для решения различных прикладных задач
4-й этап. 1985 г. – н.в.	Б. Эверитт, Ж. Селю, Г. Гуверт, Р. Левайн, Х. Дилбу, В.Ю. Королев, В.В. Глинский	Разработка различных модификаций ЕМ-алгоритма. Первые применения алгоритмов для решения прикладных экономических задач

неизвестного параметра θ. Вычисляется один из существующих моментов случайной величины X_1, который является некоторой функцией от неизвестного параметра: $\mathbb{E}X_1^k = h(\theta)$. Решив это уравнение относительно параметра θ, получим $\theta = h^{-1}(\mathbb{E}X_1^k)$. В качестве оценки параметра берется величина $\theta^* = h^{-1}(\overline{X_k})$, в которой теоретический момент заменен своим выборочным аналогом. В случае, когда неизвестных параметров несколько, то в методе моментов берется не один момент $\mathbb{E}X_1^k$, а столько, сколько требуется для того, чтобы выразить все неизвестные параметры. Применение метода моментов для разделения смесей распределений в общем случае нецелесообразно в силу большого количества оцениваемых параметров и сложности получаемой системы уравнений. Однако в частных случаях, когда смесь состоит из двух компонент, метод моментов позволяет получать простые оценки неизвестных параметров.

Первые работы, посвященные разделению смесей и вообще формулировка самой задачи, принадлежат Пирсону (1897), который применил метод моментов к смеси двух нормальных распределений. Применение метода моментов позволяет свести задачу разделения смеси к решению системы алгебраических уравнений. В частности, в работе Пирсона полученная система уравнений сводилась к решению уравнения девятой степени.

Отметим, что в ряде работ рассмотрены модификации метода моментов. В [10, 14] для составления системы уравнений используются дробные моменты, в [4–6] – факториальные.

Метод максимального правдоподобия. Метод максимального правдоподобия состоит в том, что в качестве значения параметра берут такую величину, которая максимизирует вероятность получить при n испытаниях данную выборку $\vec{X} = (X_1, X_2, ..., X_n)$.

Решение задачи разделения смеси при помощи метода максимального правдоподобия рассматривается в литературе более полно и глубоко. В ра-

ботах, которые посвящены этому методу, развиваются те или иные итерационные схемы, позволяющие найти максимум функции правдоподобия. Ниже будут более подробно рассмотрены некоторые итерационные методы, позволяющие найти точку максимума функции правдоподобия.

Метод Ньютона–Рафсона. В большинстве известных методов решение задачи о нахождении точки максимума функции правдоподобия (логарифмической функции правдоподобия) сводится к решению задачи нахождения нулей производной.

Один из итерационных алгоритмов для нахождения нулей некоторой функции – это метод Ньютона–Рафсона. Он использует вектор-градиент логарифмической функции правдоподобия и ее линейное разложение в ряд Тейлора для нахождения оценки параметра $\vec{\theta}$ на $(k + 1)$-й итерации.

В случае, когда логарифмическая функция правдоподобия является вогнутой и унимодальной, то итерационный метод сходится к точке максимума. Скорость сходимости метода – квадратичная. В случае, когда условие вогнутости не выполнено, последовательность итераций сходится не от всех произвольных начальных приближений. Основным достоинством метода является его быстрая скорость сходимости, но в некоторых случаях реализация метода влечет за собой определенные трудности: на каждой итерации требуется вычисление информационной матрицы и решение системы уравнений, которое достигается за счет арифметических операций и требует время порядка d^3, т.е. при росте размерности вектора оцениваемых параметров время, необходимое для итерации метода Ньютона–Рафсона может стать очень большим [12]. Кроме того, для ряда задач метод Ньютона–Рафсона требует нецелесообразно точных начальных приближений, чтобы итерационная последовательность сходилась к точке глобального максимума. В задачах подобного типа последовательность может сходиться не только к точкам локального максимума функции правдоподобия, но и к ее седловым точкам, и локальным минимумам.

Квазиньютоновские методы. Широкое применение получили так называемые квазиньютоновские методы, которые являются своеобразным продолжением метода Ньютона–Рафсона. Оценка $\vec{\theta}$ на $(k + 1)$-й итерации также получается при помощи вектора-градиента логарифмической функции правдоподобия и ее линейного разложение в ряд Тейлора, но вместо прямого вычисления обратной матрицы к матрице Гессе используется ее некоторое приближение. Например, приближение может быть получено при помощи поправок первого или второго рангов, вносимых на каждой итерации [13].

Квазиньютоновские методы имеют преимущество перед методом Ньютона–Рафсона в том, что они не требует оценки матрицы Гессе на каждой итерации алгоритма и реализуются способами, которые требуют только порядка d^2 арифметических операций для решения системы d линейных уравнений. Однако эти методы также требуют достаточно точных начальных приближений для $\vec{\theta}$ и для обратной матрицы к матрице Гессе. Еще одно преимущество квазиньютоновских методов состоит в том, что при выполнении необходимых условий они всегда сходятся к локальным максимумам функции правдоподобия.

Но даже при всех этих достоинствах квазиньютоновские методы не всегда применимы в статистических приложениях [11]. В частности, обыч-

но на первом шаге матрица Гессе аппроксимируется единичной матрицей, что приводит к неоправданно завышенным или заниженным значениям логарифмической функции правдоподобия.

EM-алгоритм. Одной из наиболее используемых итерационных процедур, построенных для максимизации функции правдоподобия, является EM-алгоритм. Само название EM-алгоритма и его формулировка в том виде, в котором он существует сейчас, предложены в работе Дэмпстера, Лэйрда и Рубина [7], которая посвящена применению метода максимального правдоподобия к статистическому оцениванию по неполным статистическим данным. В этой работе были исследованы свойства EM-алгоритма, его сходимость, а также приведены примеры его применения к различным прикладным задачам.

EM-алгоритм – это итерационная процедура поиска оценок максимального правдоподобия вектора параметров $(p_1, ..., p_k, \theta_1, ..., \theta_k)$. Каждая итерация алгоритма состоит из двух этапов: на первом этапе (E-этапе, англ. Expectation – ожидание) определяется условное математическое ожидание логарифма функции правдоподобия при известных значениях наблюдаемых переменных. На втором этапе (M-этапе, англ. Maximization) вычисляется оценка максимального правдоподобия, которая используется для E-этапа на следующей итерации.

Пусть известно значение вектора параметров на m-й итерации EM-алгоритма: $\left(p_1^{(m)}, ..., p_k^{(m)}, \theta_1^{(m)}, ..., \theta_k^{(m)}\right)$, $m \geq 0$. Обозначим $g_{ij}^{(m)} = \dfrac{p_i^{(m)} \cdot f_{\theta_i^{(m)}}(x_j)}{\sum_{r=1}^{k} p_r^{(m)} \cdot f_{\theta_r^{(m)}}(x_j)}$, где $f_{\theta_i}(x)$ – плотность распределения, $i = 1, ..., k, j = 1, \ldots, n$. Тогда значения параметров на $(m + 1)$-й итерации EM-алгоритма определяются следующим образом:

$p_i^{(m+1)} = \dfrac{1}{n}\sum_{j=1}^{n} g_{ij}^{(m)}$, $\theta_i^{(m+1)}$ – оценки максимального правдоподобия неизвестного параметра θ_i, построенные по реализации $\vec{x} = (x_1, x_2, ..., x_n)$ выборки $\vec{X} = (X_1, X_2, ..., X_n)$, как если бы распределение каждого ее элемента задавалось вероятностями $g_{ij}^{(m)} \Big/ \sum_{j=1}^{n} g_{ij}^{(m)}$.

Кроме того, необходимо определить критерий остановки работы EM-алгоритма. Реализация EM-алгоритма может быть осуществлена на любом языке программирования.

SEM-алгоритм. Несмотря на неоспоримое достоинство EM-алгоритма – его простоту, он обладает существенным недостатком: сильной зависимостью результатов его применения от начального приближения, так как выбирает первый попавшийся локальный максимум. То есть являясь методом локальной оптимизации, он приводит не к глобальному максимуму функции правдоподобия, а к тому локальному максимуму, который является ближайшим к начальному приближению.

Таким недостатком не обладает SEM-алгоритм – Stochastic EM-algorithm (стохастический или случайный EM-алгоритм), идея которого заключается в случайном «встряхивании» наблюдений на каждой итерации.

Пусть известно значение вектора параметров на m-й итерации EM-алгоритма: $\left(p_1^{(m)}, ..., p_k^{(m)}, \theta_1^{(m)}, ..., \theta_k^{(m)}\right)$, $m \geq 0$, а также значения $g_{ij}^{(m)}$, причем $\sum_{i=1}^{k} g_{ij}^{(m)} = 1$, для каждого $j = 1, \ldots, n$ и при каждом m.

Для каждого $j = 1, \ldots, n$ генерируются векторы $\vec{y}_j^{(m+1)} = \left(y_{1j}^{(m+1)}, y_{2j}^{(m+1)}, ..., y_{nj}^{(m+1)}\right)$ как реализации случайных векторов с полиномиальным распределением с параметрами 1 и $g_{1j}^{(m)}, ..., g_{kj}^{(m)}$ ($g_{ij}^{(m)}$ – это вероятность того, что $y_{ij}^{(m+1)} = 1$). Таким образом, векторы $\vec{y}_j^{(m+1)}$ при каждом j имеют только одну ненулевую компоненту (равную единице).

Для каждого $i = 1, \ldots, k$ вычисляются $v_i^{(m+1)} = \sum_{j=1}^{n} y_{ij}^{(m+1)}$. Тогда значения параметров на $(m + 1)$-й итерации SEM-алгоритма определяются следующим образом:

$p_i^{(m+1)} = \dfrac{v_i^{(m+1)}}{n}$, $\theta_i^{(m+1)}$ – оценки максимального правдоподобия неизвестного параметра θ_i, построенные по реализации $\vec{x} = \left(y_{i1}^{(m+1)} x_1, y_{i2}^{(m+1)} x_2, ..., y_{in}^{(m+1)} x_n\right)$ с объемом выборки $v_i^{(m+1)}$, $g_{ij}^{(m+1)} = \dfrac{p_i^{(m+1)} \cdot f_{\theta_i^{(m+1)}}(x_j)}{\sum_{r=1}^{k} p_r^{(m+1)} \cdot f_{\theta_r^{(m+1)}}(x_j)}$.

После оценки неизвестных параметров смеси распределений важной задачей является задача определения числа компонент смеси, если оно заранее неизвестно. В этом случае предлагается использовать информационный критерий Акаике (*AIC*), предложенный в 1971 г. Х. Акаике и исследованный затем в его работах 1973, 1974 и 1983 гг. и работах других авторов [1–3]. В качестве показателя качества модели Акаике предложил использовать следующую статистику:

$$AIC = -2 \ln f\left(x; \hat{\theta}(x)\right) + 2d.$$

Этот показатель является мерой «несогласия» модели и реальных данных, т.е. чем меньше значение *AIC*, тем лучше модель.

При этом включение в модель дополнительных параметров только увеличивает правдоподобие модели и, следовательно, уменьшает первое слагаемое в *AIC*. Однако при этом увеличивается второе слагаемое, «штрафующее» за использование дополнительных параметров.

Руководствуясь информационным критерием Акаике, оптимальное число компонент смеси можно определить следующим образом:

$$k_{\text{опт}} = \arg\min_k \left\{-\ln f\left(x; \hat{\theta}(x)\right) + (m+1)k - 1\right\},$$

где m – размерность параметра θ_i.

После определения числа компонент смеси распределений необходимо вычислить пороговые значения показателя, зная которые, можно определять принадлежность объекта к той или иной компоненте смеси. Предлагается считать объект со значением показателя x, принадлежащим i-й компоненте смеси, если $\dfrac{p_i \cdot F_{\theta_i}(x)}{\sum_{l=1}^{k} p_l \cdot F_{\theta_l}(x)} > \dfrac{p_j \cdot F_{\theta_j}(x)}{\sum_{l=1}^{k} p_l \cdot F_{\theta_l}(x)}$ для всех $i \neq j$, т.е. если i-я

компонента дает наибольший вклад в $F(x)$. Это условие выполняется, когда $p_i \cdot F_{\theta_i}(x) > p_j \cdot F_{\theta_j}(x)$. Таким образом, определение пороговых значений сводится к решению $(k-1)$ неравенства типа $p_i \cdot F_{\theta_i}(x) > p_j \cdot F_{\theta_j}(x)$, а затем к нахождению пересечения полученных множеств. Для большинства известных распределений подобные неравенства возможно решить в явном виде. В случае же, когда смесь состоит из двух компонент, нахождение пороговых значений сводится к решению одного неравенства: $p_1 \cdot F_{\theta_1}(x) > p_2 \cdot F_{\theta_2}(x)$.

Стоит отметить, что в случае дискретных распределений неравенства удобнее решать в виде $p_i \cdot \mathbb{P}_{\theta_i}(X = x) > p_j \cdot \mathbb{P}_{\theta_j}(X = x)$. А в случае абсолютно непрерывных распределений $p_i \cdot p_{\theta_i}(x) > p_j \cdot p_{\theta_j}(x)$.

Результаты сравнительного анализа рассматриваемых методов декомпозиции смесей распределений представлены в табл. 2.

Таблица 2

Сравнительный анализ рассматриваемых методов декомпозиции смесей распределений

Методы	Преимущества	Недостатки
Метод моментов	Простота вычисления оценок для двухкомпонентных смесей Состоятельность оценок	Сложность вычисления оценок для смесей с тремя и более компонентами Для выборок небольшого объема может давать оценки, которые существенно хуже оценок максимального правдоподобия
Метод Ньютона–Рафсона	Быстрая скорость сходимости	Большая вычислительная стоимость каждой итерации Требует большого количества памяти Обладает неустойчивостью по начальным данным Может сходиться не только к локальным максимумам функции правдоподобия, но и к минимумам и седловым точкам
Квазиньютоновские методы	Сходится к локальным максимумам функции правдоподобия Меньшая вычислительная стоимость каждой итерации по сравнению с методом Ньютона–Рафсона	Обладает неустойчивостью по начальным данным приводит к неоправданно завышенным или заниженным значениям функции правдоподобия Достаточно большая вычислительная стоимость каждой итерации
EM-алгоритм	Численно стабилен Легко программируется и не требует большого количества памяти Низкая вычислительная стоимость каждой итерации, что может компенсировать большое количество итераций по сравнению с другими алгоритмами Реализован в ППП STATISTICA	Обладает медленной скоростью сходимости для ряда задач Не гарантирует сходимости к глобальному максимуму при наличии нескольких максимумов Обладает неустойчивостью по начальным данным В некоторых задачах E-шаг может быть аналитически трудноразрешимым
SEM-алгоритм	Численно стабилен Обладает устойчивостью по начальным данным Как правило, сходится глобальному максимуму функции правдоподобия Быстрая скорость сходимости	Существует ряд смесей, для которых алгоритм не эффективен

Проведенный сравнительный анализ показывает, что для решения задачи разделения смеси наиболее эффективным является использование SEM-алгоритма. В ряде случаев можно использовать EM-алгоритм, ввиду удобства его использования (так как EM-алгоритм реализован в ППП STATISTICA), а также метод моментов в силу его простоты.

4. ТИПОЛОГИЧЕСКАЯ ГРУППИРОВКА НА ОСНОВЕ ДЕКОМПОЗИЦИИ СМЕСЕЙ ВЕРОЯТНОСТНЫХ РАСПРЕДЕЛЕНИЙ

Проведенный обзор исследований, ряд выполненных и апробированных практических работ позволили сформировать авторское видение методического подхода к разделению смесей распределений.

Предложенный методический подход реализуется в несколько этапов.

I. Статистическое наблюдение.

Исследование социально-экономического явления может проводиться на макро-, мезо- и микроуровнях. В зависимости от уровня исследования данные необходимо собирать в разрезе стран, субъектов Российской Федерации или муниципальных образований. При выборе показателя, отвечающего за характеристику социально-экономического явления или процесса, следует придерживаться следующих правил:

– показатель должен разрабатываться органами официального статистического учета;

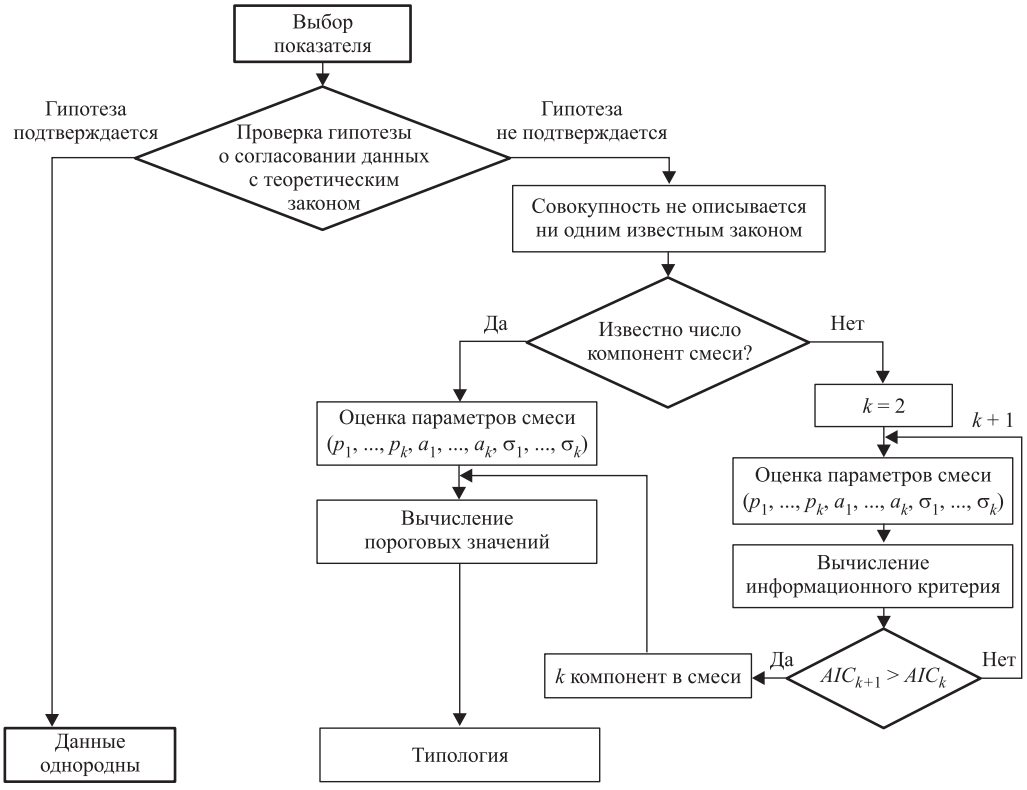
– не рекомендуется использовать показатель, оценка которого проводится экспертным путем;

– для различных единиц совокупности данные должны быть сопоставимы, т.е. скорректированы либо с учетом размера территории, либо с учетом численности населения

II. Разведывательный анализ.

Этот этап реализуется в несколько шагов. На первом шаге на основе теоретических предположений, а также при помощи гистограммы и визуального анализа выдвигается гипотеза о законе распределения, которому подчиняется распределение значений данного показателя, а также вычисляются оценки максимального правдоподобия для неизвестных параметров распределения. Далее осуществляется проверка соответствия реальных данных предполагаемому закону распределения при помощи одного из критериев согласия (критерий Пирсона, критерий Колмогорова–Смирнова). Здесь же вычисляется значение информационного критерия Акаике (AIC_1) для данной выборки и данного распределения. В случае, когда основная гипотеза по критерию согласия не отвергается, данные можно считать однородными. В случае, когда основная гипотеза отвергается в пользу альтернативной, выдвигается предположение о смеси нескольких законов распределений и дальнейшее исследование осуществляется по схеме, представленной на рисунке.

Оценка параметров смеси распределений может осуществляться при помощи любого из трех перечисленных методов, но стоит отметить, что SEM-алгоритм дает наиболее устойчивые результаты, а метод моментов разумно применять в случае, когда число компонент смеси равно двум.



Методика типологизации на основе разделения смесей вероятностных распределений

Следующий шаг после определения числа компонент смеси – определение пороговых значений для каждой компоненты, который сводится к решению системы, состоящей из $(k - 1)$ неравенства.

III. Типологическая группировка.

На основе результатов предыдущего шага проводится разделение всех объектов наблюдения на k типов. Вычисленные числовые характеристики для каждой компоненты смеси позволяют дать качественное описание каждого типа.

5. ТИПОЛОГИЯ СУБЪЕКТОВ РОССИЙСКОЙ ФЕДЕРАЦИИ ПО УРОВНЮ БЕЗРАБОТИЦЫ

На основе предложенного подхода осуществлена типологизация субъектов Российской Федерации по уровню безработицы.

I этап. Данные собраны в разрезе субъектов Российской Федерации. Уровень безработицы рассчитывался как численность безработного населения в процентах от общей численности населения субъекта. Информационный массив для исследования сформирован на основе статистической информации, представленной на официальном сайте Федеральной службы государственной статистики.

II этап. Проверено согласование реальных данных с нормальным и логнормальным законами распределения при помощи критерия согласия

Пирсона. Все гипотезы проверялись на уровне значимости 0,05. За исследуемый период эмпирическое распределение не соответствует предполагаемому теоретическому распределению.

Это позволило сделать предположение, что показатели описываются смесью двух или более нормальных распределений. Задача декомпозиции смеси для всех показателей осуществлялась SEM-алгоритмом. Затем при помощи критерия Акаике определялось количество компонент смеси. Для всех исследуемых лет распределение показателей описывается двухкомпонентной смесью нормальных распределений, причем вторая компонента характеризуется более высокими показателями числовых характеристик – среднего значения и стандартного отклонения.

III этап. Наличие смеси при описании уровня безработицы позволяет разделить все субъекты РФ на два типа: «занятые» регионы – регионы с низким уровнем безработицы, распределение которых с большей вероятностью описывается первой компонентой смеси, и «безработные» регионы – регионы с высоким уровнем безработицы, распределение которых с большей вероятностью описывается второй компонентой смеси. Для типологической группировки для каждого года были рассчитаны пороговые значения и сформулировано решающее правило: в случае, когда уровень безработицы превышает пороговое значение, субъект относится к «безработным» регионам. В табл. 3 за исследуемый период представлены «безработные» субъекты Российской Федерации, а также пороговое значение для каждого года. Курсивом выделены устойчивые субъекты РФ, которые всегда попадали во вторую группу регионов.

Таблица 3

«Безработные» субъекты РФ

Год	Пороговое значение (x_b)	Субъект РФ
2005	13,48	Республика Калмыкия, Республика Дагестан, *Республика Ингушетия*, Кабардино-Балкарская Республика, Карачаево-Черкесская Республика, Республика Тыва
2010	23,98	*Республика Ингушетия*, Чеченская Республика
2011	18,16	*Республика Ингушетия*, Чеченская Республика
2012	10,88	Республика Калмыкия, Республика Дагестан, *Республика Ингушетия*, Чеченская Республика, Республика Алтай, Республика Тыва
2013	10,76	Республика Калмыкия, Республика Дагестан, *Республика Ингушетия*, Чеченская Республика, Республика Алтай, Республика Тыва
2014	11,65	*Республика Ингушетия*, Чеченская Республика
2015	11,43	*Республика Ингушетия*, Карачаево-Черкесская Республика, Чеченская Республика, Республика Тыва
2016	11,35	*Республика Ингушетия*, Карачаево-Черкесская Республика, Чеченская Республика, Республика Алтай, Республика Тыва
2017	8,12	Республика Карелия, Республика Адыгея, Республика Калмыкия, Республика Дагестан, *Республика Ингушетия*, Кабардино-Балкарская Республика, Карачаево-Черкесская Республика, Республика Северная Осетия–Алания, Чеченская Республика, Курганская область, Республика Алтай, Республика Бурятия, Республика Тыва, Забайкальский край, Иркутская область, Еврейская автономная область

Необходимо отметить, что несмотря на снижение уровня безработицы по Российской Федерации в целом, тем не менее количество регионов, попавших в «безработные», растет, следовательно, растет и дифференциация между субъектами РФ по указанному показателю.

6. ЗАКЛЮЧЕНИЕ

В работе предложен подход к определению неоднородных данных, основанный на смесях вероятностных распределений. Разработанная на основе этого подхода методика позволяет выявлять неоднородность данных по заданному показателю, а также позволяет выполнить корректную типологию объектов наблюдения.

Проведение статистических исследований в рамках предложенного методического подхода позволяет проводить типологические группировки территориальных образований любого уровня, которые могут быть использованы при разработке и корректировке органами государственной власти документов стратегического планирования, региональных и федеральных программ экономического развития, и принятия управленческих решений, направленных на уменьшение дифференциации территорий.

Литература

1. *Aitkin M., Aitkin I.* Efficient computation of maximum likelihood estimates in mixture distributions, with reference to overdispersion and variance components. In Proceedings XVIIth International Biometric Conference, Hamilton, Ontario. Alexandria, Virginia: Biometric Society. 1994. P. 123–138.
2. *Akaike H.* Information theory and an extension of the maximum likelihood principle. in: B.N. Petrov and F. Csake (eds.) Second International Symposium on Information Theory. Akademiai Kiado, Budapest, 1973. P. 267–281.
3. *Akaike H.* A Bayesian analysis of the minimum AIC procedure. Ann. Inst. Statist. Math. 1978. Vol. 30A. P. 9–14.
4. *Blischke W.R.* Estimating the parameters of mixtures of binomial distributions. J. Amer. Statist. Assoc. 1964 – 59. № 306. P. 510–528.
5. *Blischke W.R.* Moment estimators for the parameters of a mixture of two binomial distributions. Ann. Math. Stat. 1962 – 33. № 2. P. 444–454.
6. *Cohen A.C.* Estimation in mixtures of discrete distributions. Proc Int Symp. Classical and Contagious Discrete Distrib. Montreal, 1963. P. 373–378.
7. *Dempster A.P., Laird N.M., Rubin D.B.* Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society B. 1977 – 39. P. 1–38.
8. *Glinskiy V., Serga L., Chemezova E., Zaykov K.* Clusterization economy as a way to build sustainable development of the region. Procedia CIRP 13. «13th Global Conference on Sustainable Manufacturing – Decoupling Growth from Resource Use». 2016. P. 324–328.
9. *Glinskiy V., Serga L., Khvan M.* Assessment of environmental parameters impact on the level of sustainable development of territories. Procedia CIRP 13. «13th Global Conference on Sustainable Manufacturing – Decoupling Growth from Resource Use». 2016. P. 626–631.
10. *Joffe A.D.* Mixed exponential estimation by the method of half moments. Appl. Statist. 1964 – 13. № 2. P. 91–98.
11. *Lange K.* A quasi-Newton acceleration of the EM algorithm. Statistica Sinica. 1995. 5. P. 1–18.

12. *McLachlan, Geoffrey J., Krishnan Thriyambakam, Ng, See Ket*. The EM Algorithm, Papers / Humboldt-Universität Berlin, Center for Applied Statistics and Economics (CASE), 2004, 24.
13. *Redner R.A., Walker H.E*. Mixture densities, maximum likelihood and the EM algorithm. SIAM Review. 1984. 26. P. 195–239.
14. *Tallis G.M., Light R*. The use of fractional moments for estimating the parameters of a mixed exponential distribution. Teehnometniics. 1968 – 10. № 1. P. 161–175.

Bibliography

1. *Aitkin M., Aitkin I*. Efficient computation of maximum likelihood estimates in mixture distributions, with reference to overdispersion and variance components. In Proceedings XVIIth International Biometric Conference, Hamilton, Ontario. Alexandria, Virginia: Biometric Society. 1994. P. 123–138.
2. *Akaike H*. Information theory and an extension of the maximum likelihood principle. in: B.N. Petrov and F. Csake (eds.) Second International Symposium on Information Theory. Akademiai Kiado, Budapest, 1973. P. 267–281.
3. *Akaike H*. A Bayesian analysis of the minimum AIC procedure. Ann. Inst. Statist. Math. 1978. Vol. 30A. P. 9–14.
4. *Blischke W.R*. Estimating the parameters of mixtures of binomial distributions. J. Amer. Statist. Assoc. 1964 – 59. № 306. P. 510–528.
5. *Blischke W.R*. Moment estimators for the parameters of a mixture of two binomial distributions. Ann. Math. Stat. 1962 – 33. № 2. P. 444–454.
6. *Cohen A.C*. Estimation in mixtures of discrete distributions. Proc Int Symp. Classical and Contagious Discrete Distrib. Montreal, 1963. P. 373–378.
7. *Dempster A.P., Laird N.M., Rubin D.B*. Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society B. 1977 – 39. P. 1–38.
8. *Glinskiy V., Serga L., Chemezova E., Zaykov K*. Clusterization economy as a way to build sustainable development of the region. Procedia CIRP 13. «13th Global Conference on Sustainable Manufacturing – Decoupling Growth from Resource Use». 2016. P. 324–328.
9. *Glinskiy V., Serga L., Khvan M*. Assessment of environmental parameters impact on the level of sustainable development of territories. Procedia CIRP 13. «13th Global Conference on Sustainable Manufacturing – Decoupling Growth from Resource Use». 2016. P. 626–631.
10. *Joffe A.D*. Mixed exponential estimation by the method of half moments. Appl. Statist. 1964 – 13. № 2. P. 91–98.
11. *Lange K*. A quasi-Newton acceleration of the EM algorithm. Statistica Sinica. 1995. 5. P. 1–18.
12. *McLachlan, Geoffrey J., Krishnan Thriyambakam, Ng, See Ket*. The EM Algorithm, Papers / Humboldt-Universität Berlin, Center for Applied Statistics and Economics (CASE), 2004, 24.
13. *Redner R.A., Walker H.E*. Mixture densities, maximum likelihood and the EM algorithm. SIAM Review. 1984. 26. P. 195–239.
14. *Tallis G.M., Light R*. The use of fractional moments for estimating the parameters of a mixed exponential distribution. Teehnometniics. 1968 – 10. № 1. P. 161–175.