3.7. НЕЧЕТКО-МНОЖЕСТВЕННЫЙ АНАЛИЗ МОСКОВСКОГО РЫНКА НАРУЖНОЙ РЕКЛАМЫ

Лабунец Л.В., д.т.н., с.н.с., проф. кафедры «Системы обработки информации и управления» МГТУ им. Н.Э. Баумана, зав. кафедрой «Информационные системы в экономике и управлении» НОУ ВПО «Российский новый университет»;
Лебедева Н.Л., главный специалист департамента информационных технологий ОАО «Банк ВТБ»;
Чижов М.Ю., коммерческий директор ЗАО «Миал-Холдинг»

Перейти на Главное МЕНЮ
Вернуться к СОДЕРЖАНИЮ

На основе методов интеллектуального анализа данных продемонстрирована взаимосвязь вероятностных и нечетко-множественных подходов к анализу и моделированию поступлений в московский бюджет от объектов наружной рекламы (ОНР). Представлена методика лингвистического анализа распределения поступлений в бюджет в виде аппроксимации гистограммы сглаженной сдвигом полигауссовской моделью. Рассмотрены основные этапы нечеткого логического вывода и кластеризации ОНР по критерию поступлений с помощью байесовского классификатора и метода главных компонент.

ВВЕДЕНИЕ

Быстрый рост объемов информации, содержащей в скрытой форме знания, необходимые для выявления закономерностей, прогнозирования рыночных тенденций и оценки финансовых рисков, привел к появлению интеллектуальных программных средств в виде систем поддержки принятия решений (СППР) [1]. Подсистема интеллектуального анализа данных СППР применяет современные методы и алгоритмы Data Mining. Методология этого раздела теории искусственного интеллекта предполагает формирование описательных и прогнозных моделей, основанных, например, на рациональном сочетании экспертного подхода к анализу данных, а также методов адаптивного моделирования нестационарных временных рядов (НВР) [7].

В статье [3] представлена методика анализа, моделирования и прогнозирования НВР помесячных поступлений в московский бюджет от объектов наружной рекламы (ОНР). Важной особенностью адаптивного направления анализа НВР является учет различных сценариев прогноза на основе гибких эконометрических моделей динамики временных рядов. Альтернативный экспертный подход к анализу исходной информации позволяет учесть ее неопределенность. Плодотворным в этом смысле является, на наш взгляд, применение систем, основанных на обработке знаний экспертов с помощью алгоритмов нечеткого логического вывода [2, 9]. Данная работа посвящена проблеме формирования описательных моделей эффективности рекламы на основе алгоритмов нечеткой кластеризации различных типов ОНР по уровню поступлений в бюджет.

1. МОДЕЛЬ МНОГОМЕРНЫХ ДАННЫХ

Трехмерная модель исходных данных представляет собой информационный куб [1] со следующими измерениями:

- время в виде последовательности месяцев $n = 1, 2, \dots, N$ ($N = 48$) в период с 2008 по 2011 гг. включительно;
- округ в виде упорядоченной последовательности административных округов г. Москвы $m = 1, 2, \dots, M$ ($M = 11$), представленных в табл. 1 по состоянию на 2011 г.;
- тип ОНР в виде упорядоченной последовательности типов наружной рекламы $k = 1, 2, \dots, K$ ($K = 30$), представленных в табл. 2.

Таблица 1

АДМИНИСТРАТИВНЫЕ ОКРУГА

m	Наименование
1	Общегородской заказ (ОГЗ)
2	Восточный административный округ (ВАО)
3	Западный административный округ (ЗАО)
4	Зеленоградский административный округ (ЗелАО)
5	Северный административный округ (САО)
6	Северо-восточный административный округ (СВАО)
7	Северо-западный административный округ (СЗАО)
8	Центральный административный округ (ЦАО)
9	Южный административный округ (ЮАО)
10	Юго-восточный административный округ (ЮВАО)
11	Юго-западный административный округ (ЮЗАО)

Кластерную структуру данных рационально исследовать в пространстве Евклида 11-и административных округов г. Москвы. Исходными признаками в этом случае являются компоненты вектора столбца

$$\vec{x} = (x_1, \dots, x_M)^T, \qquad (1)$$

в виде поступлений в московский бюджет от ОНР, консолидированных по месяцам в периоды:

- 2008 г. – $(1 \le n \le 12)$;
- 2009 г. – $(13 \le n \le 24)$;
- 2010 г. – $(25 \le n \le 36)$;
- 2011 г. – $(37 \le n \le 48)$.

Здесь и далее зависимость векторов от индекса времени n опускаем для упрощения записи.

Иными словами, для каждого года выборка обучающих примеров представляет собой блочную матрицу $X = (\vec{x}_1 \vdots \dots \vdots \vec{x}_K)$ размером $M * K$. Текущий столбец $\vec{x}_k = (x_{1k}, \dots, x_{Mk})^T$ матрицы – это поступления в бюджет от k-го типа рекламы в административные округа за фиксированный год.

Разведочный анализ данных рационально выполнять после предварительного преобразования Бокса-Кокса (БКП) обучающей выборки для каждого административного округа. Это преобразование позволяет сжать динамический диапазон признаков (1) и в определенной мере приблизит их к нормальному распределению. Для m-го административного округа БКП имеет вид [13].

Таблица 2

ТИПЫ НАРУЖНОЙ РЕКЛАМЫ

k	Наименование
1	Витраж
2	Выносная щитовая конструкция (штендер)
3	Информационная конструкция предприятия и организации по обслуживанию населения на здании
4	Кронштейн на здании (площадью одной стороны более 1 кв. м и высотой более 1,5 м)
5	Кронштейн на здании (площадью одной стороны менее 1 кв. м и высотой менее 1,5 м)

k	Наименование
6	Кронштейн на несветовой опоре
7	Кронштейн на световой опоре
8	Крышная установка
9	Маркиза
10	Настенное панно (площадью более 2 кв. м)
11	Настенное панно (площадью менее 2 кв. м)
12	Носимая реклама
13	Объемно-пространственная установка
14	Объемно-пространственный объект (отдельно стоящий)
15	Отдельно стоящая реклама на флагах и других мягких полотнищах
16	Прочие нетиповые объекты (мусорные контейнеры, телефонные будки и пр.)
17	Реклама на остановках транспорта
18	Реклама на строительной сетке
19	Реклама на флагах и других мягких полотнищах
20	Реклама на флагах и других мягких полотнищах на здании, маркизы (площадью более 2 кв. м)
21	Реклама на флагах и других мягких полотнищах на здании, маркизы (площадью менее 2 кв. м)
22	Стационарная установка над проезжей частью
23	Транспарант-перетяжка
24	Тумба отдельно стоящая
25	Щит на временном ограждении
26	Щит на ограждении
27	Щит отдельно стоящий
28	Электронный экран на здании
29	Электронный экран на крыше
30	Электронный экран отдельно стоящий

$$Y_{mk}(g_m)=\begin{cases}\dfrac{1}{g_m}\left\{\left(X_{mk}+\varepsilon\right)^{g_m}-1\right\}, & g_m\neq 0\\ \ln\left(X_{mk}+\varepsilon\right), & g_m=0\end{cases};$$

$$k=1,2,\ldots,K.$$

где величину смещения ε выбирают из условий $X_{mk}+\varepsilon>0$, $m=1,2,\ldots,M$, $k=1,2,\ldots,K$. Оптимальное значение параметра g_m БКП удовлетворяет критерию максимума логарифма правдоподобия:

$$L(g_m)=\{g_m-1\}\sum_{k=1}^{K}\ln\{X_{mk}+\varepsilon\}-\frac{K}{2}\ln\{D(g_m)\};$$

$$D(g_m)=\sum_{k=1}^{K}\left\{Y_{mk}(g_m)-E_m(g_m)\right\}^2;$$

$$E_m(g_m)=\frac{1}{K}\sum_{k=1}^{K}Y_{mk}(g_m).$$

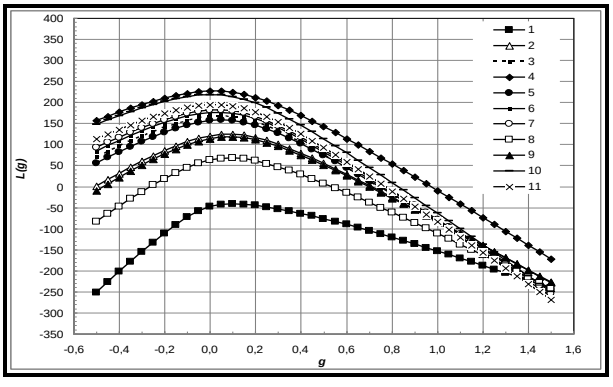

Рис. 1. Зависимость логарифма правдоподобия от параметра преобразования Бокса-Кокса

Зависимости логарифма правдоподобия $L(g_m)$ от параметра g_m для поступлений в московский бюджет от ОНР за 2008 г в 11-и административных округах и значения смещения $\varepsilon=10^{-8}$ иллюстрирует рис. 1.

Максимально правдоподобные оценки параметра БКП по округам в период с 2008 по 2011 гг. сведены в табл. 3.

Таблица 3

МАКСИМАЛЬНО ПРАВДОПОДОБНЫЕ ОЦЕНКИ ПАРАМЕТРА g_m БКП

Год	m										
	1	2	3	4	5	6	7	8	9	10	11
2008	0,1	0,1	0,05	0	0,05	0,05	0,05	0,1	0,1	0	0
2009	0,1	0,05	0,05	0	0,05	0	0,05	0,1	0,05	-0,05	0
2010	0,1	0,05	0	0	0	0	0	0,05	0,05	-0,1	0
2011	0,1	0	-0,05	-0,15	0	0	-0,05	0,05	0,05	-0,1	0

Результаты расчета показывают, что оптимальные оценки параметра БКП лежат в интервале:

$$-0,15\le g_m\le 0,1,\ m=1,2,\ldots,M.$$

Графики БКП для указанных выше значений параметра g_m демонстрирует рис. 2.

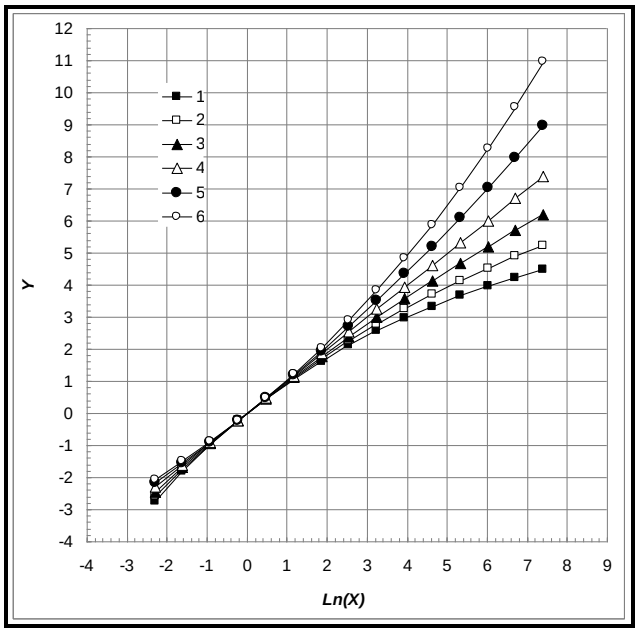

Рис. 2. БКП для различных значений параметра *g*

БКП для различных значений параметра ***g***:

1 – $g=-0,15$;

2 – $g=-0,1$;

3 – $g=-0,05$;

4 – $g=0$;

5 – $g=0,05$;

6 – $g=0,1$.

Удобным инструментом разведочного анализа данных является диаграмма рассеяния (ДР), содержащая ***K*** точек в ***M***-мерном пространстве преобразованных признаков $\vec{Y}_k=(Y_{1k},\ldots,Y_{Mk})^T$, $k=1,2,\ldots,K$. Здесь и в дальнейшем зависимость от параметра g_m БКП опускается. То-

пологию ДР удобно исследовать методом Grand Tur динамической визуализации многомерных данных [14]. Ортогональная проекция ДР на одну из гиперплоскостей в пространстве 11-и административных округов г. Москвы для поступлений в 2008 г. представлена на рис. 3.

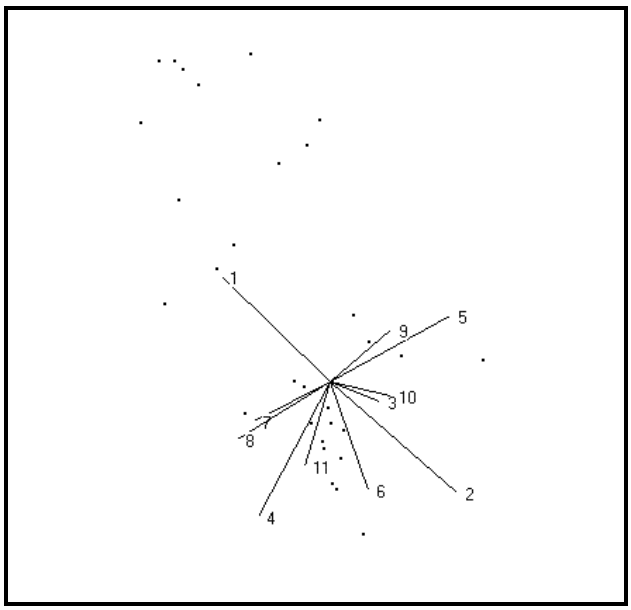

Рис. 3. Ортогональная проекция многомерной ДР на плоскость визуализации

Анализ формы ДР с различных ракурсов свидетельствует о наличии достаточно выраженных генеральных направлений в пространстве преобразованных признаков. Иными словами, имеет место значимая корреляция признаков $\vec{y} = (y_1, \ldots, y_M)^T$.

2. НЕЧЕТКО-МНОЖЕСТВЕНЫЙ И ВЕРОЯТНОСТНЫЙ ПОДХОДЫ К АНАЛИЗУ ДАННЫХ

Существенной проблемой формирования адекватных описательных статистических является наличие жесткого ограничения малого объема данных. В нашем случае гипотетический объем выборки K = 30 ассоциирован с количеством типов ОНР. Однако значения $x_{mk} = 0$, свидетельствующие об отсутствии поступлений в округа от некоторых типов рекламы, приводят к необходимости игнорировать аномально низкие величины признаков. Это снижает фактический объем данных. Известно, что в соответствии с правилом Старгеса (Sturges' rule) [15] объем выборки, необходимый для формирования гистограммы нормального распределения, содержащей как минимум пять разрядных интервалов, рассчитывают по формуле:

$$K_{min} = 2^{5-1} = 2^4 = 16 .$$

Рациональное решение проблемы малого объема выборки основано на нечетко-множественном подходе к анализу данных. В рамках такого подхода конструктивным является понятие квазистатистики, введенное в работе [9, с. 28]. В частности, консолидированные по годам выборки поступлений в бюджет от 30 типов ОНР в каждый из 11-и округов г. Москвы будем считать достаточными, чтобы с приемлемой степенью достоверности сформировать оценки законов распределения наблюдений.

С другой стороны, нет оснований кардинально разграничивать нечетко-множественный и вероятностный подходы к интеллектуальному анализу данных. В указанном смысле следует отметить результаты теоретического анализа, представленного в разделе 1.4.6 работы [10], интерпретирующего нечеткие множества как проекции случайных множеств. Плодотворное сочетание нечетко-множественной и вероятностной методологий может быть основано, на наш взгляд, на применении следующих моделей.

Во-первых, в качестве модели закона распределения наблюдений рационально выбрать гистограмму, сглаженную сдвигом (average shifted histogram, ***ASH***) [15]. Важно отметить, что ***ASH***-оценка по определению основана на сглаживании классической гистограммы окном данных. Фундаментальная процедура сглаживания позволяет значительно снизить требования к объему выборки в рамках понятия квазистатистики.

Во-вторых, достаточно адекватные результаты ***ASH***-оценивания выборочных данных позволяют сформировать байесовскую модель лингвистической переменной. Такого рода подход предполагает аппроксимацию ***ASH***-оценки конечной смесью стандартных распределений с помощью модифицированного ***EM***-алгоритма [5]. Естественным параметрическим описанием функций принадлежности в этом случае являются апостериорные вероятности ассоциации наблюдений с лингвистическими классами. Иными словами, нечеткость исходной информации, обусловленная малым объемом выборки, учитывается терм множеством лингвистической переменной. Анализу основных этапов формирования указанных выше моделей посвящены последующие разделы данной работы.

3. ИСКЛЮЧЕНИЕ АНОМАЛЬНЫХ ЗНАЧЕНИЙ ИЗ ДАННЫХ

Аномальные значения в данных существенно искажают результаты стандартного оценивания основных статистик. Одним из рациональных методов исключения влияния загрязнений выборки являются экспоненциально взвешенные оценки (ЭВО) характеристик положения и масштаба Л.Д. Мешалкина [12]. Одномерные робастные ЭВО математического ожидания (МО) $a_m(\lambda)$ и среднего квадрата отклонения (СКО) $s_m(\lambda)$, $m = 1, 2, \ldots, M$ представляют собой решение следующей системы нелинейных уравнений

$$\begin{cases} a_m(\lambda) = \sum_{k=1}^{K} q^{\lambda}\{d_{mk}(\lambda)\} Y_{mk} \Big/ \sum_{k=1}^{K} q^{\lambda}\{d_{mk}(\lambda)\}; \\ s_m^2(\lambda) = (1+\lambda)^* \\ * \sum_{k=1}^{K} q^{\lambda}\{d_{mk}(\lambda)\}\{Y_{mk} - a_m(\lambda)\}^2 \Big/ \sum_{k=1}^{K} q^{\lambda}\{d_{mk}(\lambda)\}, \end{cases} \quad (2)$$

где

$\lambda > 0$ – параметр эффективности статистик;

$q(d) = exp(-d/2)$ – экспоненциальная весовая функция;

$d_{mk} = (Y_{mk} - a_m)^2 / (2 s_m^2)$ – одномерная метрика Махалонобиса.

Здесь и далее зависимость статистик от параметра λ опускаем для упрощения записи.

Структура ЭВО обеспечивает автоматическое подавление выбросов в данных, если $\lambda > 0$. Аномально большие значения Y_{mk} формируют большие расстояния $\sqrt{d_{mk}}$, поэтому взвешиваются весами $q^{\lambda}(d_{mk})$, достаточно малыми, чтобы не вносить значимый вклад в общую сумму. В работах А.М. Шурыгина [12] показано, что ЭВО являются оценками минимума контраста, т.е. обеспечивают наименьшее значение критерия $s_m^{-\lambda/(1+\lambda)}\sum_{k=1}^{K} q^{\lambda}(d_{mk})$. Однако снижение эффективности ЭВО повышает их устойчивость к нарушению гипотезы нормальности плотности распределения вероятности (ПРВ).

Канонический вид системы уравнений (2) позволяет получать ее решение методом последовательных приближений. В качестве начальных значений характеристик положения и масштаба удобно выбрать оценки максимального правдоподобия

$$a_m = \frac{1}{K}\sum_{k=1}^{K} Y_{mk};$$

$$s_m = \frac{1}{K}\sum_{k=1}^{K}\left(Y_{mk} - a_m\right)^2.$$

Процесс сходимости алгоритма расчета ЭВО с параметром $\lambda = 1$ для характеристик положения и масштаба поступлений от ОНР за 2008 г. в случае общегородского заказа (m = 1) иллюстрирует рис. 4.

ЭВО удобны для формирования границ, отделяющих кластер «типичных» значений Y_{mk}, $k = 1, 2, \ldots, K$ от выбросов. Результаты моделирования показали, что в качестве границ рационально выбирать правило двух сигм, т.е. величины $(a_m \pm 2 s_m)$. В соответствии с указанным правилом выполнялось масштабирование данных:

$$y_{mk} = \left\{ Y_{mk} - Y_m^{(min)} \right\} \Big/ \left\{ Y_m^{(max)} - Y_m^{(min)} \right\},$$

$$k = 1, 2, \ldots, \tilde{K}_m.$$

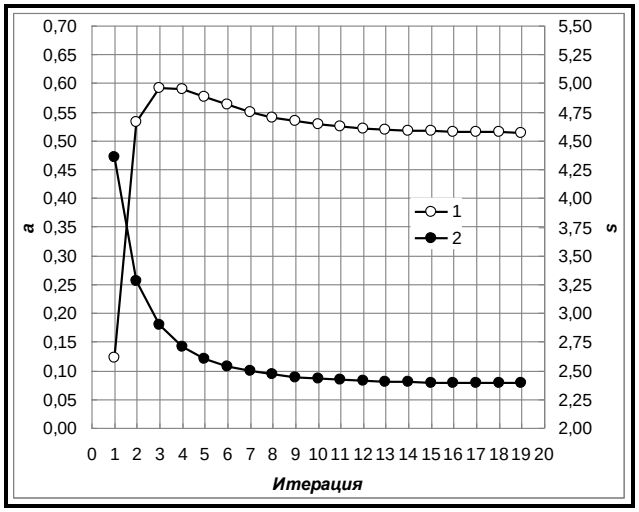

Рис. 4. ЭВО характеристик положения и масштаба: 1 – *МО*; 2 – *СКО*

Здесь

$Y_m^{(min)}$ и $Y_m^{(max)}$ – робастные оценки точной нижней и верхней границ кластера типичных значений;

$\tilde{K}_m$ – количество наблюдений в кластере.

4. ГИСТОГРАММА СГЛАЖЕННАЯ СДВИГОМ

В рамках понятия квазистатистики адекватной моделью закона распределения для выборки данных очищенных от загрязнений является ***ASH***-оценка. Процедуру сглаживания классической гистограммы окном данных описывает уравнение дискретной свертки:

$$f_m(j\Delta y_m) = \frac{1}{\tilde{K}_m \delta_m}\sum_{n=1-i_m}^{i_m-1} W_m(n)\vartheta_{j+n};$$

$$j = 0, \ldots, J_m;$$

$$J_m = 1 / \Delta y_m,$$

где

δ_m – субъективная оценка ширины разрядных интервалов (bins);

$\Delta y_m = \delta_m / i_m$ и i_m – ширина суженных интервалов (narrow bins) и их количество;

ϑ_j – количество наблюдений, попавших в j-й суженный интервал ($\vartheta_j = 0$, если $j < 0$ или $j > J_m$).

Параметр сглаживания δ_m согласуют с робастной оценкой Фридмана-Дьякониса [15]:

$$2\,IQ_m \Big/ \sqrt[3]{\tilde{K}_m},$$

где IQ_m – интерквартильный диапазон нормированных поступлений в бюджет в m-м административном округе за фиксированный год.

Кроме того, ширину разрядных интервалов классической гистограммы выбирают так, чтобы их количество было не менее пяти.

Окно данных $W_m(n)$ выбирают из условия:

$$\sum_{n=1-i_m}^{i_m-1} W_m(n) = i_m.$$

В этом случае ***ASH***-оценка распределения интегрируема с единицей. Такой нормировке удовлетворяет обобщенное окно вида:

$$W_m(n) = i_m\, Ker\,(n / i_m) \Big/ \sum_{l=1-i_m}^{i_m-1} Ker\,(l / i_m),$$

где $Ker(u)$ – положительная четная функция ядра, заданная на стандартном интервале [-1; 1] и интегрируемая с единицей.

Популярные модели ядерных функций приведены в [15, с. 140]. На рис. 5 представлена ***ASH***-оценка распределения нормированных поступлений в бюджет за 2008 г. для общегородского заказа.

В качестве модели функции ядра применялось трижды взвешенное ядро Епанечникова. Объем очищенной выборки данных составил $\tilde{K}_1 = 26$ наблюдений. Количество разрядных и суженных интервалов выбиралось 6 и 180 соответственно.

5. КОНЕЧНАЯ СМЕСЬ СТАНДАРТНЫХ РАСПРЕДЕЛЕНИЙ

Теоретически обоснованной методической основой процедуры лингвистического анализа гистограммы [8] является, на наш взгляд, модель конечной смеси стандартных распределений (рис. 5).

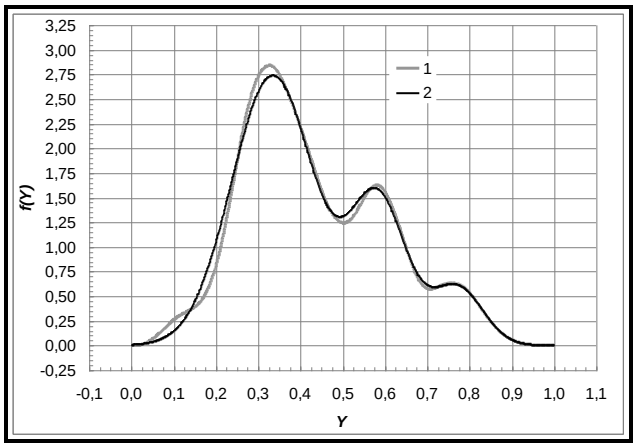

Рис. 5. Оценки распределения нормированных поступлений в бюджет: 1 – *ASH*; 2 – полигауссовская аппроксимация

$$\hat{f}_m\left(y\left|\vec{\theta}\right.\right)=\sum_{n=1}^{N_m} P_{mn}\,\varphi_{mn}\left(y\left|c_{mn},s_{mn}\right.\right),$$

$$\sum_{n=1}^{N_m} P_{mn}=1, \qquad (3)$$

где

P_{mn}, $n=1,2,\ldots,N_m$ и N_m – априорные веса лингвистических классов уровней доходности ОНР и их количество для m-го административного округа;

$\varphi_{mn}\left(y\left|c_{mn},s_{mn}\right.\right)$ – парциальные распределения с характеристиками положения c_{mn} и масштаба s_{mn} лингвистических классов. В дальнейшем там, где это не вызвано необходимостью, зависимость параметров модели (3) от номера m административного округа опускаем для сокращения записи.

В рамках представления (3) байесовский подход к формированию параметрической модели лингвистической переменной сводится к описанию функций принадлежности классов апостериорными весами:

$$w_n\left(y\left|\vec{\theta}\right.\right)=P_n\varphi_n\left(y\left|c_n,s_n\right.\right)\Big/\hat{f}\left(y\left|\vec{\theta}\right.\right),$$

$$n=1,2,\ldots,N. \qquad (4)$$

Важно отметить, что апостериорные веса надежности ассоциации текущего наблюдения с лингвистическими классами удовлетворяют условию нечеткого разбиения, т.е.

$$w_1\left(y\left|\vec{\theta}\right.\right)+w_2\left(y\left|\vec{\theta}\right.\right)+\cdots+w_N\left(y\left|\vec{\theta}\right.\right)=1,$$

и, таким образом, реализуют автоматическую мягкую классификацию наблюдений по уровням доходности ОНР. Это свойство позволяет модели функционировать в нестационарной среде в результате адаптации ее параметров по мере поступления новых данных.

Оптимальную оценку вектора параметров:

$$\vec{\theta}=\left(P_1,\ldots,P_N,c_1,\ldots,c_N,\ s_1,\ldots,s_N\right)$$

модели (3) получают с помощью эффективного в вычислительном отношении ***EM***-алгоритма. Стандартным критерием оптимальности параметров модели является функционал правдоподобия Фишера:

$$\vec{\theta}\Big|_{opt}=\arg\max_{\theta}\left\{L\left(\vec{\theta}\right)\right\};$$

$$L\left(\vec{\theta}\right)=\int_0^1 \ln\left\{\hat{f}\left(y\left|\vec{\theta}\right.\right)\right\}f(y)\,dy.$$

Рациональным критерием оптимальности является также функционал расстояния Бхатачария [5]:

$$\vec{\theta}\Big|_{opt}=\arg\min_{\theta}\left\{D\left(\vec{\theta}\right)\right\};$$

$$D\left(\vec{\theta}\right)=-\ln\left\{\int_0^1\sqrt{\hat{f}\left(y\left|\vec{\theta}\right.\right)f(y)}\,dy\right\}.$$

Решением указанных выше задач условной оптимизации является система нелинейных уравнений [5]:

$$\begin{cases} P_n=\dfrac{1}{R\left(\vec{\theta}\right)}\displaystyle\int_0^1 w_n\left(y\left|\vec{\theta}\right.\right)dy; \\ c_n=\dfrac{1}{P_nR\left(\vec{\theta}\right)}\displaystyle\int_0^1 y\,w_n\left(y\left|\vec{\theta}\right.\right)r\left(y\left|\vec{\theta}\right.\right)dy; \\ s_n^2=\dfrac{1}{P_nR\left(\vec{\theta}\right)}\displaystyle\int_0^1 y^2 w_n\left(y\left|\vec{\theta}\right.\right)r\left(y\left|\vec{\theta}\right.\right)dy-c_n^2, \end{cases}$$

$$n=1,2,\ldots,N. \qquad (5)$$

Здесь $r\left(y\left|\vec{\theta}\right.\right)$ – весовая функция, определяемая функционалом качества оценок параметров конечной смеси:

$$r\left(y\left|\vec{\theta}\right.\right)=\begin{cases} f(y), & \text{для правдоподобия;} \\ \sqrt{\hat{f}\left(y\left|\vec{\theta}\right.\right)f(y)}, & \text{для расстояния,} \end{cases}$$

$$R\left(\vec{\theta}\right)=\int_0^1 r\left(y\left|\vec{\theta}\right.\right)dy.$$

Канонический вид системы уравнений (5) позволяет аппроксимировать ***ASH***-оценку распределения нормированных поступлений в бюджет за фиксированный год в текущем административном округе параметрической моделью (3) с помощью итерационной процедуры [6, с. 1299], реализующей метод последовательных приближений.

Выбор количества классов N терм множества лингвистической переменной уровней доходности ОНР и начальных оценок параметров модели (3) в значительной степени является субъективным и основывается на мнении эксперта. Это замечание справедливо и в случае явно выраженной модальной структуры ***ASH***-оценки, поскольку объединение соседних лингвистических классов с функциями принадлежности $w_n\left(y\left|\vec{\theta}\right.\right)$ и $w_{n+1}\left(y\left|\vec{\theta}\right.\right)$ не составляет труда. В силу свойства нечеткого разбиения функция принадлежности консолидированного кластера приобретает вид:

$$\tilde{w}_n\left(y\left|\vec{\theta}\right.\right)=w_n\left(y\left|\vec{\theta}\right.\right)+w_{n+1}\left(y\left|\vec{\theta}\right.\right),$$

а количество классов N терм множества уменьшается на единицу.

Популярной на практике является полигауссовская модель конечной смеси распределений (3), в которой в качестве парциальных применяют нормальные распределения:

$$\varphi_n\left(y \left| c_n, s_n\right.\right) = \frac{1}{s_n\sqrt{2\pi}} exp\left\{-\left(y - c_n\right)^2 / \left(2 s_n^2\right)\right\}.$$

Полигауссовская аппроксимация ***ASH***-оценки распределения нормированных поступлений в бюджет за 2008 г. для общегородского заказа представлена на рис. 5. Явно выраженная модальная структура распределения свидетельствует о наличии трех классов уровней доходности ОНР – низкой, средней и высокой. Байесовскую модель (4) соответствующей трехуровневой лингвистической переменной демонстрирует рис. 6.

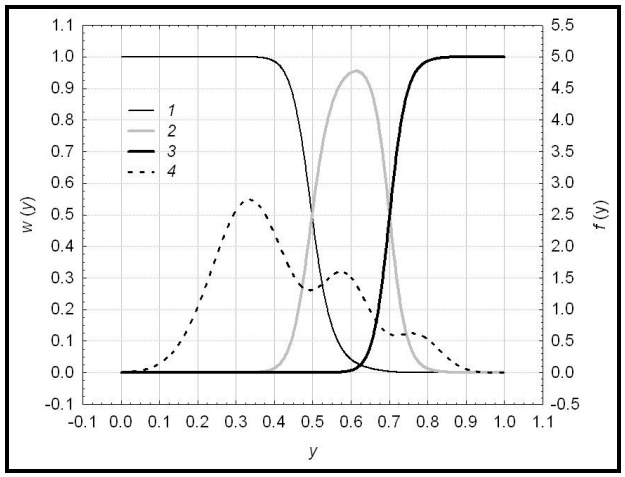

Рис. 6. Байесовская модель лингвистической переменной: 1, 2, 3 – функции принадлежности классов низкой, средней, высокой «доходности» ОНР; 4 – полигауссовская аппроксимация ASH-оценки распределения

Процесс сходимости ***EM***-алгоритма обучения модели (3) по критерию правдоподобия иллюстрирует рис. 7. Из графиков видно, что сходимость параметров модели к оптимальным значениям достигается практически за 10 итераций как по критерию правдоподобия Фишера – ***L***, так и по критерию расстояния Бхатачария – ***D***.

6. НЕЧЕТКИЙ ЛОГИЧЕСКИЙ ВЫВОД

Адекватное ранжирование ОНР по уровню поступлений в бюджет при наличии неполных или достаточно скудных исходных данных рационально выполнять с помощью алгоритмов нечеткого логического вывода. Одним из основных этапов формирования моделей приближенных рассуждений является агрегирование нечеткой информации об объекте анализа. В нашем случае это композиция средних уровней доходности лингвистических классов. Иными словами, комплексный показатель эффективности различных типов ОНР естественно представить в виде байесовской модели учета неопределенности:

$$y_{mk}^{(AC)} = \sum_{n=1}^{N_m} c_{mn}\, w_{mn}\left(y_{mk} \left| \vec{\theta}_m\right.\right),$$

$$m = 1, 2, \ldots, M,$$

$$k = 1, 2, \ldots, K,$$

где c_{mn} и y_{mk} – средняя доходность (узловая точка [8; 9, с. 157]) ***n***-го лингвистического класса и нормированные поступления в городской бюджет от ***k***-го типа ОНР в ***m***-м административном округе за фиксированный год.

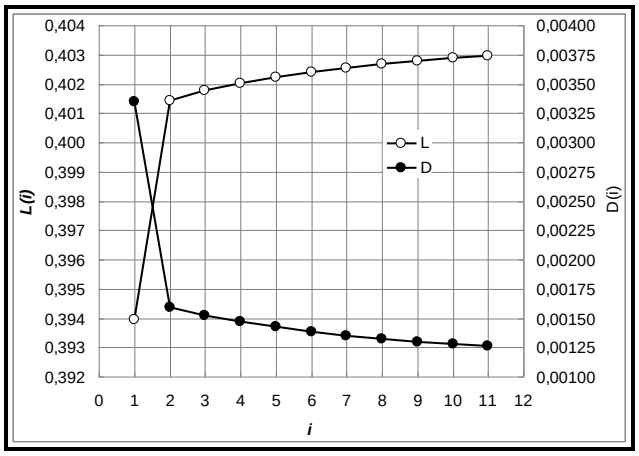

Рис. 7. Сходимость*EM*-алгоритма по итерациям *i* обучения модели

Последовательность значений $y_{mk}^{(AC)}$, $k = 1, 2, \ldots, K$, упорядоченная по убыванию, ранжирует различные типы ОНР по их доходности в ***m***-м административном округе за фиксированный год. Кроме того, байесовский классификатор, зачисляющий ***k***-й тип ОНР в ***n***-й класс доходности по правилу победителя, т.е.

$$w_{mn}\left(y_{mk} \left| \vec{\theta}_m\right.\right) > w_{mj}\left(y_{mk} \left| \vec{\theta}_m\right.\right),$$

$$j = 1, 2, \ldots, N_m,\ j \neq n.$$

реализует лингвистическую кластеризацию ОНР. В качестве примера в табл. 4-6 представлены результаты такого рода кластеризации по уровням поступлений в бюджет за 2008 г. для общегородского заказа.

Следующий уровень иерархии нечеткого логического вывода связан с формированием одного или нескольких обобщенных показателей эффективности различных типов ОНР по городу в целом. На этом этапе агрегируют комплексные показатели $y_{mk}^{(AC)}$ по административным округам. В теории принятия решений известны различные методы экспертного оценивания [11] весов важности частных критериев для их линейной свертки. Рациональной методической основой для формирования подобного рода обобщенных показателей является, на наш взгляд, метод главных компонент (МГК).

В рамках МГК процедура формирования информативных признаков сводится к анализу собственных значений и векторов ковариационной матрицы данных. Робастные ЭВО вектора математического ожидания $\vec{A}(\lambda)$ и ковариационной матрицы $B(\lambda)$ для ***M***-мерной выборки векторов:

$$\vec{Y}_k = \left\{ y_{1k}^{(AC)}, \ldots, y_{Mk}^{(AC)} \right\}^T,\ k = 1, 2, \ldots, K$$

получают в результате решения системы нелинейных уравнений [12, с. 75].

$$\begin{cases} \vec{A}(\lambda) = \sum_{k=1}^{K} q^{\lambda}\{D_k(\lambda)\}\vec{Y}_k \Big/ \sum_{k=1}^{K} q^{\lambda}\{D_k(\lambda)\}; \\ B(\lambda) = (1+\lambda)^* \\ {}^* \sum_{k=1}^{K} q^{\lambda}\{D_k(\lambda)\}\{\vec{Y}_k - \vec{A}(\lambda)\}\{\vec{Y}_k - \vec{A}(\lambda)\}^T / \\ / \sum_{k=1}^{K} q^{\lambda}\{D_k(\lambda)\}, \end{cases}$$

например, методом последовательных приближений. Здесь

$$D_k(\lambda) = \{\vec{Y}_k - \vec{A}(\lambda)\}^T B^{-1}(\lambda)\{\vec{Y}_k - \vec{A}(\lambda)\},$$

$k = 1, 2, \ldots, K$.

M-мерная метрика Махаланобиса в пространстве административных округов г. Москвы для средних нормированных поступлений от различных типов ОНР.

Анализ собственных значений v_m, $m = 1, 2, \ldots, M$ ЭВО ковариационных матриц $B(\lambda)$ для параметра эффективности статистик $\lambda = 1$ за 2008-й, 2009-й, 2010-й и 2011 гг. показал, что эффективный ранг ***r*** этих матриц по критерию:

$$e^2(r) = (v_{r+1} + \cdots + v_M)/(v_1 + \cdots + v_M) \le 0{,}1$$

не превышают четырех. В качестве примера рис. 8 иллюстрирует тенденцию убывания собственных значений матрицы $B(\lambda)$ за 2008 г.

Таблица 4

ОНР С ВЫСОКИМ УРОВНЕМ ПОСТУПЛЕНИЙ В ГОРОДСКОЙ БЮДЖЕТ

Тип конструкции	*k*	y_{1k}	w_{11}	w_{12}	w_{13}	$y_{1k}^{(AC)}$
Щит отдельно стоящий	27	1,23395241	0,000000	0,000000	1,000000	0,770188
Транспарант-перетяжка	23	1,05954391	0,000000	0,000000	1,000000	0,770188
Настенное панно (площадью более 2 кв. м)	10	1,00000002	0,000000	0,000000	1,000000	0,770188
Крышная установка	8	0,96001369	0,000001	0,000032	0,999967	0,770182
Кронштейн на световой опоре	7	0,89895473	0,000002	0,000278	0,999720	0,770135
Электронный экран отдельно стоящий	30	0,78391064	0,000099	0,025169	0,974732	0,765435
Кронштейн на здании (площадью одной стороны более 1кв.м и высотой более 1,5 м)	4	0,72123955	0,001604	0,282399	0,715998	0,716648
Тумба отдельно стоящая	24	0,71903489	0,001757	0,302869	0,695374	0,712751
Реклама на остановках транспорта	17	0,70634273	0,002893	0,435171	0,561936	0,687500

Таблица 5

ОНР СО СРЕДНИМ УРОВНЕМ ПОСТУПЛЕНИЙ В ГОРОДСКОЙ БЮДЖЕТ

Тип конструкции	*k*	y_{1k}	w_{11}	w_{12}	w_{13}	$y_{1k}^{(AC)}$
Стационарная установка над проезжей частью	22	0,697657	0,003936	0,533458	0,462607	0,668655
Информационная конструкция предприятия и организации по обслуживанию населения на здании	3	0,607643	0,035579	0,953546	0,010875	0,576263
Реклама на флагах и других мягких полотнищах	19	0,533415	0,229593	0,770221	0,000186	0,526039
Настенное панно (площадью менее 2 кв. м)	11	0,531127	0,243079	0,756759	0,000162	0,522683
Кронштейн на здании (площадью одной стороны менее 1 кв.м и высотой менее 1,5 м)	5	0,518575	0,328578	0,671349	0,000073	0,501415
Электронный экран на крыше	29	0,515978	0,348624	0,651315	0,000061	0,496430

Таблица 6

ОНР С НИЗКИМ УРОВНЕМ ПОСТУПЛЕНИЙ В ГОРОДСКОЙ БЮДЖЕТ

Тип конструкции	*k*	y_{1k}	w_{11}	w_{12}	w_{13}	$y_{1k}^{(AC)}$
Отдельно стоящая реклама на флагах и других мягких полотнищах	15	0,46345953	0,796002	0,203997	0,000001	0,385219
Выносная щитовая конструкция (штендер)	2	0,43031309	0,942021	0,057979	0,000000	0,348924
Объемно-пространственная установка	13	0,42467768	0,954512	0,045488	0,000000	0,345819
Щит на временном ограждении	25	0,40706334	0,979532	0,020468	0,000000	0,339600
Маркиза	9	0,39303412	0,989563	0,010437	0,000000	0,337107
Кронштейн на несветовой опоре	6	0,38571665	0,992739	0,007261	0,000000	0,336318
Объемно-пространственный объект (отдельно стоящий)	14	0,37392815	0,996016	0,003984	0,000000	0,335503
Реклама на флагах и других мягких полотнищах на здании, маркизы (площадью более 2 кв. м)	20	0,34779001	0,999013	0,000987	0,000000	0,334758
Электронный экран на здании	28	0,33877843	0,999402	0,000598	0,000000	0,334662
Прочие нетиповые объекты (мусорные контейнеры, телефонные будки и пр.)	16	0,20824337	1,000000	0,000000	0,000000	0,334513
Щит на ограждении	26	0,12049077	1,000000	0,000000	0,000000	0,334513
Витраж	1	0,11459128	1,000000	0,000000	0,000000	0,334513
Реклама на флагах и других мягких полотнищах на здании, маркизы (площадью менее 2 кв. м)	21	0,00000006	1,000000	0,000000	0,000000	0,334513
Носимая реклама	12	-0,15619906	1,000000	0,000000	0,000000	0,334513
Реклама на строительной сетке	18	-1,70538556	1,000000	0,000000	0,000000	0,334513

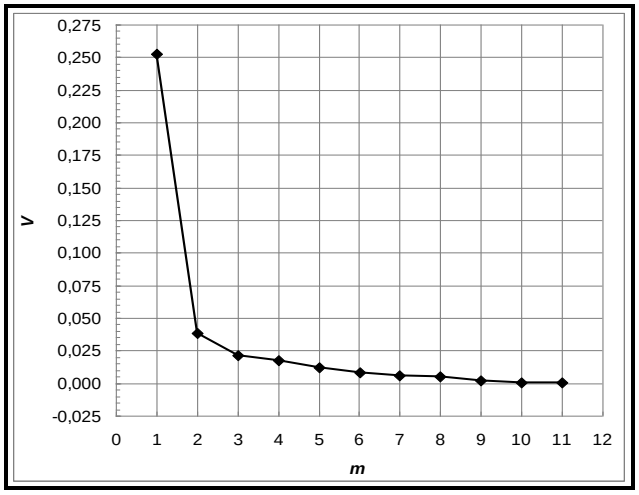

Рис. 8. Собственные значения ковариационной матрицы

Собственные значения ЭВО ковариационной матрицы $B(\lambda)$ за указанные выше годы сведены в табл. 7.

Таблица 7

СОБСТВЕННЫЕ ЗНАЧЕНИЯ КОВАРИАЦИОННОЙ МАТРИЦЫ СРЕДНИХ НОРМИРОВАННЫХ ПОСТУПЛЕНИЙ В БЮДЖЕТ ОТ ОНР

Год	v_1	v_2	v_3	v_4	$e^2(4)$
2008	0,2527	0,0387	0,0216	0,0173	0,0941
2009	0,2202	0,0331	0,0183	0,0135	0,0843
2010	0,2127	0,0442	0,0139	0,0084	0,0484
2011	0,2092	0,0386	0,0173	0,0088	0,0414

Иными словами, ортонормированный базис информативного подпространства образован собственными векторами $\vec{u}_1$, $\vec{u}_2$, $\vec{u}_3$ и $\vec{u}_4$ матрицы $B(\lambda)$ в исходном пространстве $\left(y_1^{(AC)},\dots,y_M^{(AC)}\right)$. Соответствующее декоррелирующее преобразование формирует оценки:

$$z_{jk} = \left\{\vec{Y}_k - \vec{A}(\lambda)\right\}^T \vec{u}_j \Big/ \sqrt{v_j},$$

$j = 1, 2, 3, 4,\ k = 1, 2, \dots, K$

четырех информативных показателей эффективности различных типов ОНР по городу в целом за фиксированный год. Рис. 9 демонстрирует зависимости от времени главных компонент $z_{1k},\dots,z_{4k}$ для трех типов ОНР k = 27, 23, 10 (см. табл. 2), ранжированных по уровню поступлений в городской бюджет [3].

Ортогональная проекция ДР, полученная методом Grand Tur, на одну из гиперплоскостей в пространстве четырех главных компонент $(z_1,\dots,z_4)$ для поступлений в 2008 г. представлена на рис. 10.

Главные компоненты в силу некоррелированности содержат примерно равные доли информации об объекте анализа. Поэтому имеются все основания считать их равно важными. Иными словами, интегральный показатель эффективности различных типов ОНР по критерию поступлений в городской бюджет приобретает вид:

$$z_k = \frac{1}{4}\left(z_{1k} + \dots + z_{4k}\right),\ k = 1, 2, \dots, K.$$

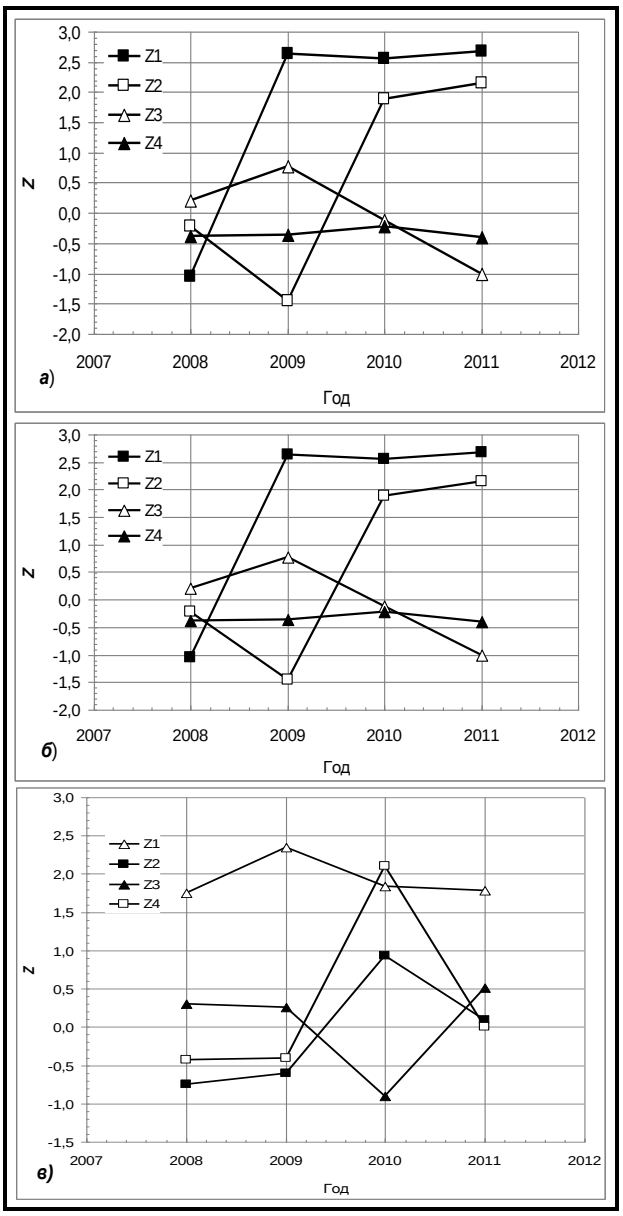

Рис. 9. Зависимости от времени информативных показателей эффективности различных типов ОНР: а) щит отдельно стоящий; б) транспарант-перетяжка; в) настенное панно (больше 2 кв. м)

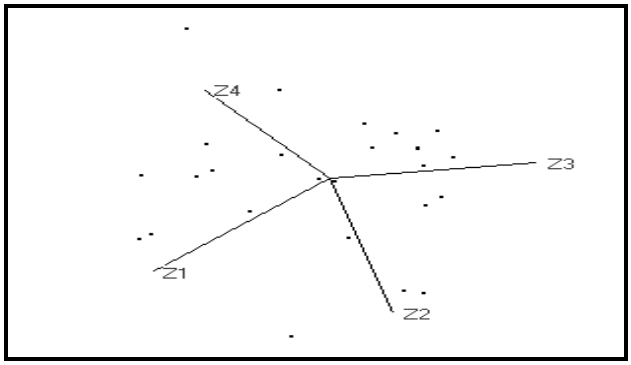

Рис. 10. Ортогональная проекция ДР в пространстве главных компонент

Коррелирующее преобразование в рамках МГК позволяет реконструировать вектор:

$$\hat{\vec{Y}}_k = \left\{ \hat{y}_{1k}^{(AC)}, \dots, \hat{y}_{Mk}^{(AC)} \right\}^T$$

комплексных показатели $y_{mk}^{(AC)}$ эффективности ОНР по административным округам:

$$\hat{\vec{Y}}_k = \vec{A}(\lambda) + \sum_{j=1}^{4} z_{jk} \vec{u}_j \sqrt{v_j}, \quad k = 1, 2, \dots, K$$

с достаточно малыми ошибками:

$$\left\| \vec{Y}_k - \hat{\vec{Y}}_k \right\|^2 \le e^2(4) \left\| \vec{Y}_k \right\|^2.$$

Рис. 11 иллюстрирует точность реконструкции комплексных показателей эффективности $y_{mk}^{(AC)}$ различных типов ОНР $k = 1, 2, \dots, K$ в 2011 г. для общегородского заказа (m = 1) и в Центральном административном округе (m = 8) г. Москвы (см. табл. 1).

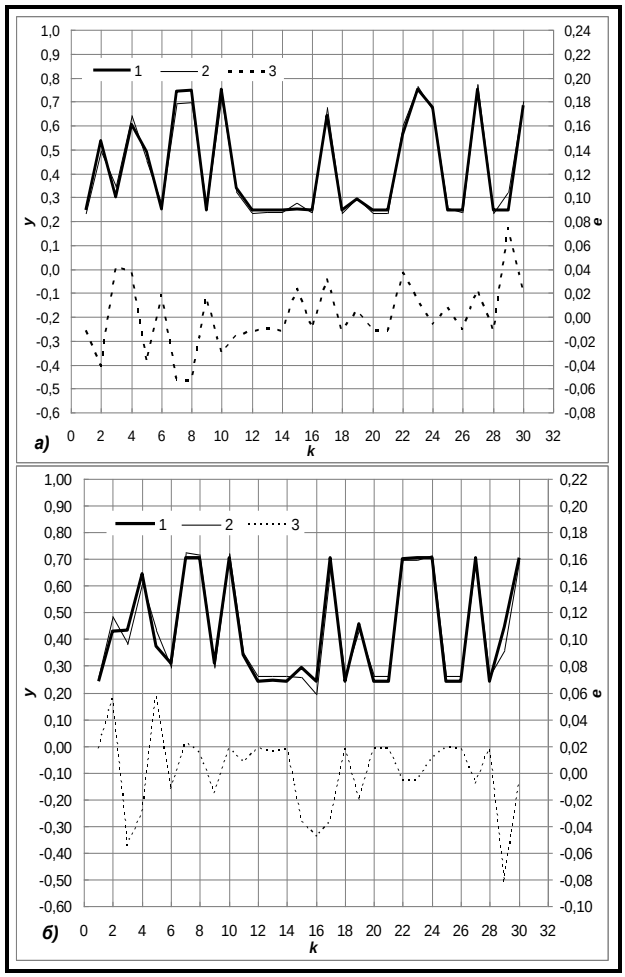

Рис. 11. Реконструкция комплексных показателей эффективности ОНР: а) общегородской заказ; б) центральный административный округ; 1 – показатель $y_{mk}^{(AC)}$; 2 – реконструкция; 3 – ошибка реконструкции

Такого рода оценки позволяют в принципе решать задачи прогнозирования показателей $y_{mk}^{(AC)}$ эффективности ОНР в ортонормированном базисе $(\vec{u}_1, \vec{u}_2, \vec{u}_3, \vec{u}_4)$ информативного подпространства на основе экстраполяции зависимостей главных компонент z_{jk} от времени. Один из вариантов решения указанной выше задачи представлен в работе [4].

ЗАКЛЮЧЕНИЕ

В работе проанализирована тесная взаимосвязь вероятностного и нечетко-множественного подходов к анализу и моделированию эффективности различных типов ОНР по критерию поступлений в московский бюджет. Представлена методика учета фактора неопределенности, обусловленного малым объемом выборки на основе применения фундаментальной процедуры сглаживания экспериментальных данных. В частности, проиллюстрирована эффективность лингвистического анализа распределений исходных признаков в результате комбинации непараметрической модели, в виде гистограммы, сглаженной сдвигом, и параметрического описания в виде конечной смеси стандартных плотностей. Сочетание этих моделей позволило сформулировать этапы нечеткого логического вывода в терминах байесовской процедуры агрегирования исходных показателей эффективности ОНР и формирования интегрального критерия методом главных компонент. Такого рода описания позволяют в принципе синтезировать динамические адаптивные модели нечетких множеств.

Литература

1. Барсегян А.А. и др. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP [Текст] / А.А. Барсегян, М.С. Куприянов, В.В. Степаненко, И.И. Холод. – 2-е изд., перераб. и доп. – СПб : БХВ-Петербург, 2007. – 384 с.
2. Ведерников В.В. Нечетко-множественное моделирование в анализе и прогнозировании экономических явлений и процессов: исторический аспект [Электронный ресурс] / В.В. Ведерников // Проблемы современной экономики. – 2006. – №1. Режим доступа: http://www.m-economy.ru
3. Лабунец Л.В. и др. Анализ динамики поступлений в бюджет от московского рынка наружной рекламы [Текст] / Л.В. Лабунец, Н.Л. Лебедева, М.Ю. Чижов // Вестн. Российского нового ун-та. – 2013. – №4. – С. 66-81.
4. Лабунец Л.В. Прогнозирование объемов продаж компании методами структурного анализа данных [Текст] / Л.В. Лабунец, Н.Л. Лабунец // Труды Десятой Междунар. науч. конф.: в 2 ч. Ч. 2: тез. докл. – М., 2009. – С. 368-372.
5. Лабунец Л.В. Рандомизация многомерных распределений в метрике Махаланобиса [Текст] / Л.В. Лабунец // Радиотехника и электроника. – 2000. – Т. 45 ; №10. – С. 1214-1225.
6. Лабунец Л.В. и др. Реконструкция отражательных характеристик 3D-объектов в однопозиционной системе оптической локации [Текст] / Л.В. Лабунец, Д.С. Лукин, А.А. Червяков // Радиотехника и электроника. – 2012. – Т. 57. – №12. – С. 1289-1300.
7. Лукашин Ю.П. Современные направления статистического анализа взаимосвязей и зависимостей [Текст] / Ю.П. Лукашин, Л.И. Рахлин ; отв. ред. Ю.П. Лукашин. – М. : ИМЭМО РАН, 2012. – 54 с.
8. Недосекин А.О. Лингвистический анализ гистограмм экономических факторов [Текст] / А.О. Недосекин, С.Н. Фролова // Вестник ВГУ, сер. Экономика и управление. – 2008. – №2. – С. 48-55.
9. Недосекин А.О. Нечетко-множественный анализ рисков фондовых инвестиций [Текст] / А.О. Недосекин. – СПб. : Сезам, 2002. – 181 с.
10. Орлов А.И. Прикладная статистика [Текст] / А.И. Орлов. – М. : Экзамен, 2006. – 671 с.
11. Сидельников Ю.В. Системный анализ экспертного прогнозирования [Текст] / Ю.В. Сидельников. – М. : МАИ, 2007. – 453 с.

12. Шурыгин А.М. Прикладная стохастика: робастность, оценивание, прогноз [Текст] / А.М. Шурыгин. – М. : Финансы и статистика, 2000. – 224 с.
13. Box G.E.P., Cox D.R. An analysis of transformation // Journal of the royal statistical society. Series B (methodological). 1964. Vol. 26. №2. Pp. 211-252.
14. Buja A., Cook D., Asimov D., Hurley C. Theory and computational methods for dynamic projections in high-dimensional data visualizations // Journal of computational and graphical statistics. 1999. Vol. 8. №3. Pp. 1-24.
15. Scott D.W. Multivariate density estimation: theory, practice, and visualization. N.-Y.: John Wiley & Sons, Inc, 1992. 317 p.

Ключевые слова

Data Mining; преобразования Бокса-Кокса; диаграмма рассеяния; квазистатистики; экспоненциально взвешенные оценки Мешалкина; лингвистический анализ гистограммы; ***EM***-алгоритм; байесовские оценки; нечеткий логический вывод; метод главных компонент.

Лабунец Леонид Витальевич

Лебедева Наталья Леонидовна

Чижов Михаил Юрьевич

РЕЦЕНЗИЯ

Актуальность темы статьи Лабунца Л.В., Лебедевой Н.Л., Чижова М.Ю. «Нечетко-множественный анализ московского рынка наружной рекламы» обусловлена, во-первых, необходимостью разработки инновационных управленческих методов и моделей в составе современных систем поддержки принятия решений в экономике. Во-вторых, статья содержит достаточно аргументированный ответ на развернувшуюся в научном сообществе дискуссию о соотношении вероятностно-статистической и нечетко-множественной методологий в процедурах формирования решений в условиях неполной и нечеткой исходной информации.

Практическая значимость работы состоит в том, что представлено оригинальное решение для задачи оценки эффективности различных типов наружной рекламы по критерию поступлений в городской бюджет на примере Москвы. Изложенный в статье анализ позволил авторам получить практически важные результаты относительно кластеризации объектов наружной рекламы по критерию их эффективности, как в пределах административных округов, так и по городу в целом. Результаты расчетов представлены в динамике по годам в период с 2008 г. по 2011 г., что позволяет формировать адекватные прогнозы поступлений в городской бюджет.

Научная новизна работы связана с разработкой методики ранжирования и нечеткой кластеризации по выбранному критерию качества объектов анализа. Методика в значительной мере инвариантна к содержанию предметной области. Ее основой является рациональное сочетание вероятностно-статистической и нечетко-множественной методологий. В частности, следует отметить, что жесткое ограничение малого объема выборки данных в значительной мере ослабляется за счет применения таких непараметрических и параметрических методов и моделей прикладной статистики (в терминологии авторов – интеллектуального анализа данных), как гистограмма, сглаженная сдвигом, конечная смесь стандартных распределений и ***EM***-алгоритм, байесовское описание лингвистической переменной и процедуры агрегирования частных показателей.

Заключение: рецензируемая статья отвечает требованиям, предъявляемым к научным публикациям, и может быть рекомендована к опубликованию.

Орлов А.И., д.т.н., д.э.н., к.ф.-м.н., проф. кафедры «Экономика и организация производства» (ИБМ-2) научно-учебного комплекса «Инженерный бизнес и менеджмент» Московского государственного технического университета им. Н.Э. Баумана, зав. лабораторией экономико-математических методов в контроллинге Научно-образовательного центра «Контроллинг и управленческие инновации» Московского государственного технического университета им. Н.Э. Баумана

Перейти на Главное МЕНЮ
Вернуться к СОДЕРЖАНИЮ