

УДК 519.24

СВОЙСТВА КОМБИНИРОВАННОЙ ОЦЕНКИ РЕГРЕССИИ ПРИ КОНЕЧНЫХ ОБЪЕМАХ ВЫБОРОК

С.В. Скрипин

Томский государственный университет

Томский научный центр СО РАН

E-mail: skripin@ef.tsu.ru

Представлены свойства комбинированной непараметрической оценки регрессии, использующей оценки двух построенных моделей регрессии – параметрической и непараметрической. Применение бутстреп-метода для комбинированной модели позволяет получать оценки, близкие к оптимальным по критерию минимума среднеквадратического отклонения. Статистическим моделированием показано, что при выборках конечного объема комбинированная оценка предпочтительнее, чем каждая из оценок построенных моделей.

Ключевые слова:

Непараметрическая оценка регрессии, комбинированная оценка, бутстреп-оценка, статистическое моделирование.

Задача оценки регрессии при конечных объемах выборок широко применяется в технических и компьютерных системах различного назначения и может быть решена с использованием различных моделей, как из класса параметрических так и непараметрических. В условиях малого объема данных возрастают требования к качеству оценок регрессионных моделей. Для повышения качества оценок можно использовать комбинированные модели регрессии различного вида. При отсутствии априорной информации о выборке конечного объема N качество оценок регрессии комбинированных моделей может быть лучше, чем у моделей из класса параметрических или непараметрических, либо получено частное решение для конкретной выборки.

Другой подход для оценки регрессии состоит в получении оценок комбинированных моделей, использующих оценки построенных моделей регрессии и априорную информацию о некоторых функциях плотности. В настоящее время разработаны различные комбинированные модели регрессии и методы получения комбинированных оценок, в том числе [1–3].

Результаты работы [1] позволяют оперировать множеством построенных моделей из разных классов и рассматривать в виде результатов классы оценок. В данной работе рассмотрен случай реализации комбинированной модели регрессии [1], построенной по оценкам двух моделей – параметрической и непараметрической, с использованием подхода, предложенного в [2]. В этом случае матричные и векторные выражения в формулах преобразуются в скалярные, что позволяет упростить программную реализацию алгоритмов, сократить требования к вычислительным ресурсам и время обработки данных.

Пусть $(Y_1, X_1), \dots, (Y_N, X_N)$ независимые наблюдения случайного вектора (Y, X) с неизвестной плотностью вероятности $f(x, y)$; $x \in R^m$, $y \in R^1$. Для построения комбинированных оценок выберем две модели регрессии из разных классов: линейную параметрическую $J(x; \theta) = \theta_0 + \theta_1 x^{(1)} + \dots + \theta_m x^{(m)}$ и непараметрическую $j(x) = \int_{-\infty}^{\infty} y f(y|x) dy$, где $f(y|x)$ – неизвестная условная

плотность вероятности случайной величины Y при фиксированном значении $X=x$. Комбинированную оценку регрессии представим в виде:

$$J_N(x; \lambda) = J_N(x) - \lambda(J_N(x) - J(x; \hat{\theta})). \quad (1)$$

Здесь λ – коэффициент, выбираемый согласно требованиям к качеству оценки, $J(x; \theta)$ – параметрическая оценка регрессии, где θ – оценка по методу наименьших квадратов вектора $\theta = (\theta_0, \dots, \theta_m)$, $J_N(x)$ – непараметрическая оценка регрессии следующего вида

$$J_N(x^{(1)}, \dots, x^{(m)}) = \frac{\sum_{i=1}^N Y_i \prod_{j=1}^m K[(x^{(j)} - X_i^{(j)}) / h_N^{(j)}]}{\sum_{i=1}^N \prod_{j=1}^m K[(x^{(j)} - X_i^{(j)}) / h_N^{(j)}]},$$

где $K(u)$ – заданное ядро (некоторая функция плотности вероятности), $h_N^{(j)} = c_j (\frac{m}{N})^{\frac{1}{4+m}}$ – параметры масштаба, а c_j – выбираются из условия минимума суммы квадратов регрессионных остатков

$$Q(c_1, \dots, c_m) = \sum_{i=1}^N [Y_i - J_N(X_i^{(1)}, \dots, X_i^{(m)})]^2 \rightarrow \min_{c_1, \dots, c_m}.$$

Продemonстрируем преимущества комбинированной оценки регрессии вида (1) по сравнению с оценками построенных моделей (параметрической или непараметрической). Рассмотрим два критерия качества оценок регрессии и соответствующие им комбинированные оценки. Выберем в качестве критерия выражение суммы квадратов регрессионных остатков

$$Q_1(\lambda) = \sum_{i=1}^N [Y_i - J_N(X_i^{(1)}, \dots, X_i^{(m)}; \lambda)]^2 \rightarrow \min_{\lambda}. \quad (2)$$

Из выражения (2) получим оптимальный коэффициент $\lambda = \lambda_0$:

$$\lambda_0 = \frac{\sum_{i=1}^N (J_N(X_i^{(1)}, \dots, X_i^{(m)}) - Y_i)(J_N(X_i^{(1)}, \dots, X_i^{(m)}) - J(X_i^{(1)}, \dots, X_i^{(m)}; \hat{\theta}))}{\sum_{i=1}^N (J_N(X_i^{(1)}, \dots, X_i^{(m)}) - J(X_i^{(1)}, \dots, X_i^{(m)}; \hat{\theta}))^2}. \quad (3)$$

Коэффициент λ_0 в выражении (3) не зависит от выбора точки $x \in R^m$, в которой оценивается регрессия. С учетом (3) критерий (2) примет вид

$$Q_1(\lambda_o) = \sum_{i=1}^N [Y_i - J_N(\mathbf{X}_i)]^2 - \frac{[\sum_{i=1}^N (J_N(\mathbf{X}_i) - Y_i)(J_N(\mathbf{X}_i) - J(\mathbf{X}_i; \hat{\theta}))]^2}{\sum_{i=1}^N (J_N(\mathbf{X}_i) - J(\mathbf{X}_i; \hat{\theta}))^2}.$$

Добавим внутрь скобок критерия (2) два слагаемых: $+J(\mathbf{X}_i; \hat{\theta})$ и $-J(\mathbf{X}_i; \hat{\theta})$. В результате, с учетом оптимального коэффициента λ , получим эквивалентный вид критерия (2)

$$Q_1^a(\lambda_o) = \sum_{i=1}^N [Y_i - J(\mathbf{X}_i; \hat{\theta})]^2 - \frac{[\sum_{i=1}^N (Y_i - J(\mathbf{X}_i; \hat{\theta}))(J_N(\mathbf{X}_i) - J(\mathbf{X}_i; \hat{\theta}))]^2}{\sum_{i=1}^N (J_N(\mathbf{X}_i) - J(\mathbf{X}_i; \hat{\theta}))^2}.$$

Рассмотрим другую комбинированную оценку (1) с выбором коэффициента λ на основе критерия

$$Q_2(\lambda) = E[J_N(\mathbf{x}; \lambda) - J(\mathbf{x})]^2 \rightarrow \min_{\lambda}, \quad (4)$$

где E – оператор математического ожидания, $J(\mathbf{x})$ – неизвестная истинная функция регрессии. Из выражения (4) получим оптимальный коэффициент $\lambda(\mathbf{x}) = \lambda_o(\mathbf{x})$, зависящий от выбора точки $\mathbf{x} \in R^m$,

$$\lambda_o(\mathbf{x}) = \frac{E(J_N(\mathbf{x}) - J(\mathbf{x}))(J_N(\mathbf{x}) - J(\mathbf{x}; \hat{\theta}))}{E(J_N(\mathbf{x}) - J(\mathbf{x}; \hat{\theta}))^2}, \quad (5)$$

а критерий (4) с учетом (5) примет вид

$$Q_2(\lambda_o(\mathbf{x})) = E[J_N(\mathbf{x}) - J(\mathbf{x})]^2 - \frac{[E(J_N(\mathbf{x}) - J(\mathbf{x}))(J_N(\mathbf{x}) - J(\mathbf{x}; \hat{\theta}))]^2}{E(J_N(\mathbf{x}) - J(\mathbf{x}; \hat{\theta}))^2}.$$

Или, аналогично критерию (2), в эквивалентном виде

$$Q_2^a(\lambda_o(\mathbf{x})) = E[J(\mathbf{x}; \hat{\theta}) - J(\mathbf{x})]^2 - \frac{[E(J(\mathbf{x}) - J(\mathbf{x}; \hat{\theta}))(J_N(\mathbf{x}) - J(\mathbf{x}; \hat{\theta}))]^2}{E(J_N(\mathbf{x}) - J(\mathbf{x}; \hat{\theta}))^2}.$$

Из представленных выражений для квадратов остатков комбинированной оценки (1) – $Q_1(\lambda_o)$, $Q_1^a(\lambda_o)$, $Q_2(\lambda_o(\mathbf{x}))$, $Q_2^a(\lambda_o(\mathbf{x}))$ следует, что они меньше, чем квадраты остатков построенных моделей (первые слагаемые выражений). Следовательно, комбинированная оценка (1) эффективнее оценок построенных моделей (параметрической или непараметрической) как в случае конечного объема выборок N (критерий (2)), так и в асимптотическом случае (критерий (4)). Но проблема заключается в получении действительно оптимальных коэффициентов λ . На качество оценок влияет и объем выборок N (особенно когда он мал), и ошибки измерений, другие шумы в исходных данных. Далее, в отличие от (3) выражение (5) не может быть оцене-

но непосредственно по исходной выборке, поскольку $J(\mathbf{x})$ – неизвестна. Но можно предложить несколько способов оценки коэффициента $\lambda_o(\mathbf{x})$. Выделив в (5) главную часть и оценив ее по исходной выборке, получим оценку $\lambda_o(\mathbf{x})$ в точке $\mathbf{x}$ [2]:

$$\hat{\lambda}_o^g(\mathbf{x}) = \frac{C(\mathbf{x})}{V(\mathbf{x})} \left(1 - \frac{NH(\Delta_1(\mathbf{x}))^2 / V(\mathbf{x})}{1 + NH(\Delta_1(\mathbf{x}))^2 / V(\mathbf{x})} \right). \quad (6)$$

Здесь $C(\mathbf{x}) = g_2(\mathbf{x}) - J_N(\mathbf{x})g_1(\mathbf{x})$,

$$\Delta_1(\mathbf{x}) = \Delta_0(\mathbf{x}) \left(1 - \frac{1}{1 + (NH)^{\frac{1}{3}}(\Delta_0(\mathbf{x}))^2} \right), \quad H = \prod_{j=1}^m h_N^{(j)},$$

$$\Delta_0(\mathbf{x}) = g_1(\mathbf{x}) - J(\mathbf{x}; \hat{\theta})g_0(\mathbf{x})$$

$g_0(\mathbf{x})$, $g_1(\mathbf{x})$, $g_2(\mathbf{x})$ имеют вид:

$$g_0(\mathbf{x}) = \frac{1}{NH} \sum_{i=1}^N \prod_{j=1}^m K[(x^{(j)} - X_i^{(j)}) / h_N^{(j)}],$$

$$g_1(\mathbf{x}) = \frac{1}{NH} \sum_{i=1}^N Y_i \prod_{j=1}^m K[(x^{(j)} - X_i^{(j)}) / h_N^{(j)}],$$

$$g_2(\mathbf{x}) = \frac{1}{NH} \sum_{i=1}^N Y_i^2 \prod_{j=1}^m K[(x^{(j)} - X_i^{(j)}) / h_N^{(j)}].$$

Оценку коэффициента $\lambda_o(\mathbf{x})$ в выражении (5) можно выполнить и бутстреп-методом [4]. При этом оценку коэффициента можно представить в двух вариантах.

Первый вариант предусматривает получение для каждой бутстреп-выборки оценок регрессии двух моделей – параметрической и непараметрической. Обозначим оценки, получаемые по бутстреп-методу символом звездочка, а по исходной выборке – без звездочки. В соответствии с (5) оценка оптимального коэффициента $\lambda_o(\mathbf{x})$ по бутстреп-выборкам примет вид [3]

$$\hat{\lambda}_o^*(\mathbf{x}) = \frac{\sum_{j=1}^B (J_{N,j}^*(\mathbf{x}) - J_N(\mathbf{x}))(J_{N,j}^*(\mathbf{x}) - J(\mathbf{x}; \hat{\theta}^*))}{\sum_{j=1}^B (J_{N,j}^*(\mathbf{x}) - J(\mathbf{x}; \hat{\theta}^*))^2}, \quad (7)$$

где B – размер серии сформированных бутстреп-выборок, соответствующих одной исходной выборке, для оценки коэффициента $\lambda_o^*(\mathbf{x})$ в точке $\mathbf{x} \in R^m$.

Второй вариант предусматривает получение для каждой бутстреп-выборки оценок регрессии только одной модели – непараметрической. Тогда оценка оптимального коэффициента $\lambda_o(\mathbf{x})$ по бутстреп-выборкам будет иметь вид

$$\hat{\lambda}_o^+(\mathbf{x}) = \frac{\sum_{j=1}^B (J_{N,j}^*(\mathbf{x}) - J_N(\mathbf{x}))(J_{N,j}^*(\mathbf{x}) - J(\mathbf{x}; \hat{\theta}))}{\sum_{j=1}^B (J_{N,j}^*(\mathbf{x}) - J(\mathbf{x}; \hat{\theta}))^2}. \quad (8)$$

Оценка (8) позволяет экономить вычислительные ресурсы и избегать проблем с получением оценок параметрической модели регрессии в случае вырожденной матрицы ковариаций у сформированных бутстреп-выборок.

Для сравнения качества оценок регрессии в условиях малых объемов выборок был проведен имитационный эксперимент, а также проверка качества оценок на нескольких выборках реальных статистических данных. Сравнение свойств комбинированных оценок является нетривиальной задачей, поскольку оценки коэффициентов λ в выражениях (3), (6)–(8) получены с использованием разных критериев оптимальности и требуют корректных условий сравнения. Для простоты интерпретации результатов имитационный эксперимент был проведен с одномерной величиной $x \in R^1$. Поскольку оценки (6)–(8) адаптированы к выбору точки x , вначале были рассмотрены свойства оценок регрессии в фиксированных точках выборки x_i ($i=1, \dots, n$), выбранных с одинаковым шагом в заданном интервале моделирования. Затем выполнено сравнение свойств оценок со случайным выбором точек X_i . Для проверки качества комбинированных оценок в различных условиях в эксперименте использованы две функции «истинной» регрессии – линейная $P_1(x)=2,0+x$ и нелинейная $P_2(x)=x(10,0-x)$. Для добавления шумов к переменной отклика выбраны соответственно функции $Y_1(x)=P_1(x)+\xi$ и $Y_2(x)=P_2(x)+\xi$, где ξ – независимая случайная величина с нормальным законом распределения, нулевым математическим ожиданием и одинаковым значением дисперсии σ для всех точек заданного диапазона.

Для проведения имитационного эксперимента были выбраны следующие условия моделирования:

- для сравнения качества оценок в условиях малых объемов выборки выбран фиксированный объем $n=10$;
- интервал моделирования величин x_i и X_i выбран в диапазоне $[0,9;9,0]$;
- шаг между фиксированными точками x_i выбран 0,9;
- случайная величина X_i выбрана из равномерного закона в диапазоне $[0,9;9,0]$;
- величина дисперсии случайной величины ξ равна $\sigma=3,0$;
- числовые результаты эксперимента были получены по серии исходных выборок количеством $K=1000$ (с одинаковым объемом наблюдений n в каждой выборке). При этом для каждого наблюдения каждой исходной выборки из серии K формировалась серия B бутстреп-выборок размером $B=1000$.

Сравнение качества оценок регрессии у шести моделей ($J_1=J(x;\hat{\theta})$, $J_2=J_N(x)$, $J_3=J_N(x;\lambda_0)$, $J_4=J_N(x;\lambda_0^*(x))$, $J_5=J_N(x;\lambda_0^s(x))$, $J_6=J_N(x;\lambda_0^+(x))$) можно проводить по многим критериям. Поскольку преимущества комбинированной оценки регрессии вида (1) для применения к выборочным данным очевидны, ограничимся критериями, вычисляющими близость оценок моделей к истинной регрессии. Укажем критерии, вычисленные для каждой исходной выборки.

1. Средняя квадратичная погрешность прогнозирования в фиксированных точках

$$S_j^p(P(x)) = \frac{1}{n} \sum_{i=1}^n [P_p(x_i) - J_j(x_i)]^2, \quad j = \overline{1,6}, \quad p = \overline{1,2},$$

где $P_p(x_i)$ – истинное значение регрессии вида p в фиксированной точке x_i , $J_j(x_i)$ – оценка регрессии по модели j в фиксированной точке x_i .

2. Средняя квадратичная погрешность прогнозирования в произвольных точках

$$S_j^p(P(X)) = \frac{1}{n} \sum_{i=1}^n [P_p(X_i) - J_j(X_i)]^2, \quad j = \overline{1,6}, \quad p = \overline{1,2},$$

где $P_p(X_i)$ – истинное значение регрессии вида p в произвольной точке X_i , $J_j(X_i)$ – оценка регрессии по модели j в произвольной точке X_i .

Укажем критерии, вычисленные по серии K исходных выборок.

3. Среднее квадратичных погрешностей прогнозирования по серии K в фиксированных точках

$$Q_j^p(P(x)) = \frac{1}{K} \sum_{s=1}^K S_j^p(P(x)), \quad j = \overline{1,6}, \quad p = \overline{1,2}. \quad (9)$$

4. Среднее квадратичных погрешностей прогнозирования по серии K в произвольных точках

$$Q_j^p(P(X)) = \frac{1}{K} \sum_{s=1}^K S_j^p(P(X)), \quad j = \overline{1,6}, \quad p = \overline{1,2}. \quad (10)$$

5. Средняя квадратичная погрешность прогнозирования по серии K в каждой фиксированной точке

$$Q_j^p(P(x_i)) = \frac{1}{K} \sum_{s=1}^K [P_p(x_i) - J_j(x_i)]^2, \quad i = \overline{1,n}, \quad j = \overline{1,6}, \quad p = \overline{1,2}. \quad (11)$$

6. Среднее оценок регрессии по серии выборок K в каждой фиксированной точке

$$\bar{J}_j^p(x_i) = \frac{1}{K} \sum_{s=1}^K J_j^p(x_i), \quad i = \overline{1,n}, \quad j = \overline{1,6}, \quad p = \overline{1,2}.$$

7. Квадраты отклонений прогнозирования по серии K в каждой фиксированной точке

$$I_j^p(x_i) = [P_p(x_i) - \bar{J}_j^p(x_i)]^2, \quad i = \overline{1,n}, \quad j = \overline{1,6}, \quad p = \overline{1,2}.$$

8. Среднее квадратов отклонений прогнозирования по серии K в фиксированных точках

$$I_j^p(x) = \frac{1}{n} \sum_{i=1}^n I_j^p(x_i), \quad j = \overline{1,6}, \quad p = \overline{1,2}. \quad (12)$$

Приведем результаты эксперимента. Средние квадратичных погрешностей оценок регрессии по критериям (9), (10) и (12) представлены в таблице.

Средние погрешностей оценок регрессии по критериям (11) представлены на рисунке.

По результатам сравнения при указанных условиях моделирования можно сделать следующие выводы.

1. Оценки модели регрессии:

- параметрической J_1 дают наилучший результат среди представленных моделей в случае, когда истинна линейная регрессия. При этом качество оценок комбинированных моделей J_5 и J_6 почти равно J_1 (рисунок, а). В случае

нелинейной регрессии оценки параметрической модели наихудшие (рисунок, б);

Таблица. Средние квадратичных погрешностей и квадратов отклонений прогнозирования в точках

Вид истинной регрессии	Модели регрессии					
	Линейная параметрич. J_1	Непараметрическая J_2	Комбинированная J_3	Комбинированная J_4	Комбинированная J_5	Комбинированная J_6
Среднее квадратичных погрешностей прогнозирования в фиксированных точках $Q^0(P(x))$						
Линейная P_1	1,252	2,515	2,318	3,189	1,450	1,437
Нелинейная P_2	44,645	6,881	4,924	17,577	18,206	17,247
Среднее квадратичных погрешностей прогнозирования в произвольных точках $Q^0(P(X))$						
Линейная P_1	1,395	5,315	2,612	4,195	2,057	2,236
Нелинейная P_2	44,801	19,524	19,164	20,991	24,267	22,709
Среднее квадратов отклонений прогнозирования в фиксированных точках $P(x)$						
Линейная P_1	0,00064	0,13873	0,07389	0,09404	0,00116	0,05292
Нелинейная P_2	43,429	4,618	0,510	14,685	15,711	14,769

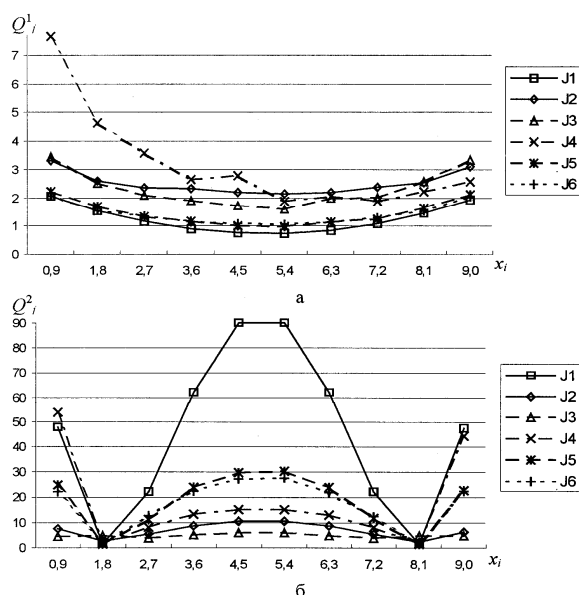


Рисунок. Средняя квадратичная погрешность прогнозирования $Q^0(P(x_i))$ в каждой точке x_i для случая регрессии: а) линейной P_1 , б) нелинейной P_2

- непараметрической J_2 дают невысокий результат среди представленных моделей в случае истинной линейной регрессии. В случае нелинейной регрессии оценки модели дают второй по качеству результат (рисунок, б);
- комбинированной J_3 в случае, когда истинна линейная регрессия, дают результат лучше, чем у непараметрической модели J_2 и комбинированной

ной J_4 . В случае нелинейной регрессии модель J_3 дает наилучшие результаты (рисунок). Это свидетельствует о хороших свойствах модели J_3 ;

- комбинированной модели регрессии J_4 в случае истинной нелинейной регрессии дают результат, уступающий лишь оценкам непараметрической модели J_2 и комбинированной модели J_3 (рисунок, б);
 - комбинированной J_5 в случае, когда истинна линейная регрессия, дают третий по качеству результат среди представленных моделей, уступая лишь модели J_1 и модели J_6 (рисунок, а). В случае, когда истинна нелинейная регрессия оценки модели J_5 дают удовлетворительный результат (рисунок, б);
 - комбинированной J_6 дают результат, почти одинаковый с оценками модели J_5 по указанным критериям, несколько выигрывая у модели J_5 . Это справедливо в случаях линейной и нелинейной регрессии (рисунок).
2. Проигрыш оценок комбинированной модели J_4 в случае истинной линейной регрессии можно объяснить малым объемом данных исходных выборок n . Применение бутстреп-метода для моделей J_5 и J_6 требует больших вычислительных затрат, растущих с увеличением N .
 3. При переходе к произвольному выбору точек X свойства оценок регрессии у представленных моделей почти не меняются, а величины погрешностей для случаев линейной и нелинейной регрессии – сближаются.
 4. При невысоком качестве оценок одной из построенных моделей (параметрической или непараметрической) комбинированная оценка (1) позволяет достигать наилучших результатов в условиях малых объемов выборок.
 5. Погрешность оценок регрессии на краях диапазона моделирования (рисунок) у всех представленных моделей регрессии увеличивается (в большей или меньшей степени).

На реальных данных оценки комбинированных моделей регрессии (J_3 , J_4 , J_5) показали лучшие результаты по сравнению с оценками построенных моделей J_1 и J_2 .

Результаты эксперимента позволяют утверждать, что предложенные комбинированные оценки регрессии J_3 , J_4 , J_5 и J_6 предпочтительнее оценок построенных моделей J_1 и J_2 при конечных объемах выборок. Получение оценок моделей J_3 и J_4 требует значительно меньших вычислительных ресурсов по сравнению с оценками моделей J_5 и J_6 .

Представленные комбинированные оценки могут найти широкое применение для получения качественных оценок регрессии в условиях малых объемов выборок и отсутствии априорной информации о данных.

Результаты моделирования, представленные на рисунках и в таблицах, получены с помощью кластера Межрегионального Вычислительного Центра ТГУ СК ИФ Cyberia (skif.tsu.ru). Автор выражает благодарность сотрудникам Центра за оказанную помощь.

СПИСОК ЛИТЕРАТУРЫ

1. Дмитриев Ю.Г. Непараметрическое условное оценивание функционалов плотности распределения // Математическое моделирование и теория вероятностей. — Томск: Изд-во «Пеленг», 1998. — С. 169–177.
2. Скрипин С.В. Комбинированная непараметрическая оценка регрессии // Проблемы компьютерной безопасности и криптография: Докл. IV Сибирской научной школы-семинара с международным участием. — Sibecrypt'05 // Вестник Томского государственного университета. — 2005. — Прил. № 14. — С. 311–313.
3. Скрипин С.В. О свойствах комбинированной оценки регрессии // Проблемы компьютерной безопасности и криптография: Докл. VI Сибирской научной школы-семинара с международным участием. — Sibecrypt'07 // Вестник Томского государственного университета. — 2007. — Прил. № 23. — С. 326–329.
4. Эфрон Б. Нетрадиционные методы многомерного статистического анализа. — М.: Финансы и статистика, 1988. — 261 с.

Поступила 09.06.2008 г.

УДК 504.064(4)

ПРИМЕНЕНИЕ ПРОСТРАНСТВЕННЫХ ХАРАКТЕРИСТИК ПРИ МОДЕЛИРОВАНИИ ИЗМЕНЕНИЙ ЛАНДШАФТНОГО ПОКРОВА С ИСПОЛЬЗОВАНИЕМ КЛЕТОЧНЫХ АВТОМАТОВ

А.В. Замятин

Томский политехнический университет
E-mail: zamyatin@tpu.ru

Предложен подход к моделированию изменений ландшафтного покрова с использованием клеточных автоматов и 7-ми пространственных характеристик, наиболее часто встречающихся при решении различных задач геоинформатики, позволяющий повысить адекватность моделирования. Проведен поиск информативности каждой из характеристик в различных размерах скользящего окна. Предложена методика оценки значимости характеристик в различных сочетаниях при определении правил функционирования клеточных автоматов, а также выполнена соответствующая оценка значимости характеристик. Результаты экспериментов получены с использованием типовых разновременных данных растровой геоинформационной системы Idrisi Kilimanjaro.

Ключевые слова:

Моделирование, клеточные автоматы, пространственные характеристики, численные эксперименты, экосистемный мониторинг.

Введение

Моделирование изменений ландшафтного покрова является одним из наиболее перспективных подходов к прогнозированию таких явлений как распространение лесных пожаров, обезлесение, эрозия почвы, изменение границ снежного покрова [1, 2]. Решают задачи моделирования с использованием как детерминированных, так и стохастических моделей [3]. С учетом случайной природы моделируемых процессов ландшафтного покрова и возрастающими возможностями компьютерной техники наиболее перспективными следует считать стохастические модели, учитывающие пространственное взаимодействие элементов с помощью клеточных автоматов (КА).

Одним из ключевых факторов, влияющих на адекватность моделирования с использованием КА, является определение в каждом конкретном случае правил взаимодействия элементов (правил функционирования) КА. В настоящее время все большее распространение получают подходы к использованию пространственных характеристик при решении различных задач геоинформатики [4–6], а предварительные исследования в [7] показывают перспективность использования таких характеристик для определения правил функционирования КА.

При моделировании практическое применение пространственных характеристик, позволяющих более полно учитывать особенности ландшафта, обычно затруднено в связи с их различной информативностью в различных размерах анализируемой окрестности, а также неопределенностью вариантов возможного сочетания характеристик и их необходимого количества в каждом конкретном случае. Так, в [2, 4] использована лишь одна, а в [7] — четыре пространственные характеристики при решении задач моделирования изменений ландшафтного покрова.

Данная работа направлена на повышение эффективности моделирования изменений ландшафтного покрова с использованием увеличенного набора из 7-ми пространственных характеристик, наиболее часто встречающихся при решении различных задач геоинформатики, анализ этих характеристик в варьируемом размере анализируемой окрестности и определение их результативной сочетаемости при совместном применении.

Правила функционирования КА при моделировании

В [7] автором правила функционирования КА в алгоритме моделирования изменений ландшафтного покрова предлагается определять в каждой