ПРОГНОЗИРОВАНИЕ РЕАЛЬНЫХ ДАННЫХ СРЕДСТВАМИ ПАКЕТА R

Т. В. Цеховая, О. Н. Жданович

Белорусский государственный университет
Минск, Беларусь
e-mail: tsekhavaya@bsu.by, olenka.zhdanovich@mail.ru

Для прогнозирования реальных данных использованы методы обратных взвешенных расстояний и радиальных базовых функций. Сделан сравнительный анализ точности прогнозов.

Ключевые слова: прогнозирование; интерполяция; детерминированные методы; кроссвалидация.

REAL DATA PREDICTING WITH R

T. Tsekhavaya, V. Zhdanovich

Belarusian State University
Minsk, Belarus

In order to predict the real data were used inverse distance weighting methods and radial basis functions. A comparative analysis of the prediction accuracy was done.

Keywords: forecasting; interpolation; deterministic methods; cross-validation.

ВВЕДЕНИЕ

В настоящее время актуальны задачи построения точных прогнозов по наблюдаемым данным. В статье для прогнозирования реальных данных использованы детерминированные методы интерполяции: обратных взвешенных расстояний (Inverse Distance Weighting) и радиальных базовых функций (Radial Basis Functions) [2]. Найдены оптимальные входные параметры методов и осуществлено сравнение точности полученных прогнозов. Решение указанной задачи выполнялось в свободной среде программирования для анализа данных и графики R.

ПЕРВИЧНЫЙ АНАЛИЗ ДАННЫХ

Реальные данные представляют собой измерения процентного содержания угольной золы угольного пласта в Пенсильвании [1]. Выборка составляет 208 наблюдений.

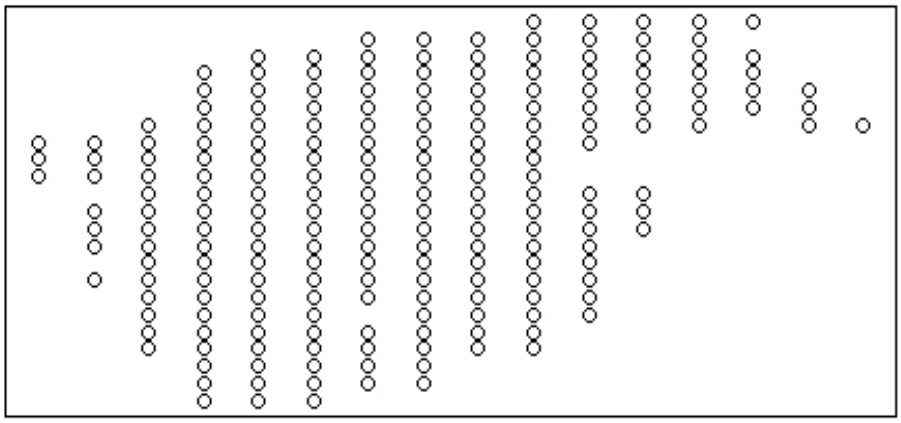

Рис. 1. Сетка наблюдений

Вычислены основные статистические характеристики исходных наблюдений. Так, например, среднее значение равно 9,78, дисперсия – 1,63, минимум и максимум составляют 7,00 и 17,61 соответственно. Коэффициент вариации равен 0,13, что свидетельствует об однородности исследуемой выборки. Осуществлена проверка гипотезы о нормальном распределении генеральной совокупности по критериям Колмогорова–Смирнова и Пирсона. Показано, что при уровне значимости 0,05 нет оснований отвергнуть гипотезу о нормальном распределении $\mathcal{N}(9{,}78,\ 1{,}63)$ генеральной совокупности. Для проверки на наличие выбросов в выборке был использован статистический критерий Граббса, согласно которому можно утверждать, что при уровне значимости 0,05 аномальных значений среди наблюдений нет. Гипотеза об отсутствии тренда в исходных данных была подтверждена при уровне значимости 0,05 проверкой критерия Манна–Кендалла.

ДЕТЕРМИНИРОВАННЫЕ МЕТОДЫ ИНТЕРПОЛЯЦИИ

Для проверки качества выбранной модели и анализа точности прогнозов была использована кросс-валидация, которая основана на сравнении реальных данных с прогнозными значениями, полученными вышеуказанными методами интерполяции [2]. Оптимальные параметры методов были найдены на основе кросс-валидации с помощью вычисления следующих характеристик качества оценки: коэффициент корреляции r между исходными и оцененными данными, сдвиг среднего значения исходных и оцененных данных Δ, сумма квадратов невязок S, среднеквадратическая ка $RSME$ и коэффициент эффективности E [2].

В среде R метод обратных взвешенных расстояний реализован с помощью функции idw() в пакете geostat. Основным параметром этой функции является Power, который можно задать вручную. По умолчанию значение этого параметра равно 2,0.

Как видно из рис. 2, наилучшая интерполяция будет при Power = 4.

В программной среде R для реализации метода радиальных базовых функций существует функция rbf() из пакета geostat. В качестве главного параметра этот метод использует различные базисные функции. В данном случае наилучший результат получен при использовании обратной мультиквадратичной базисной функции.

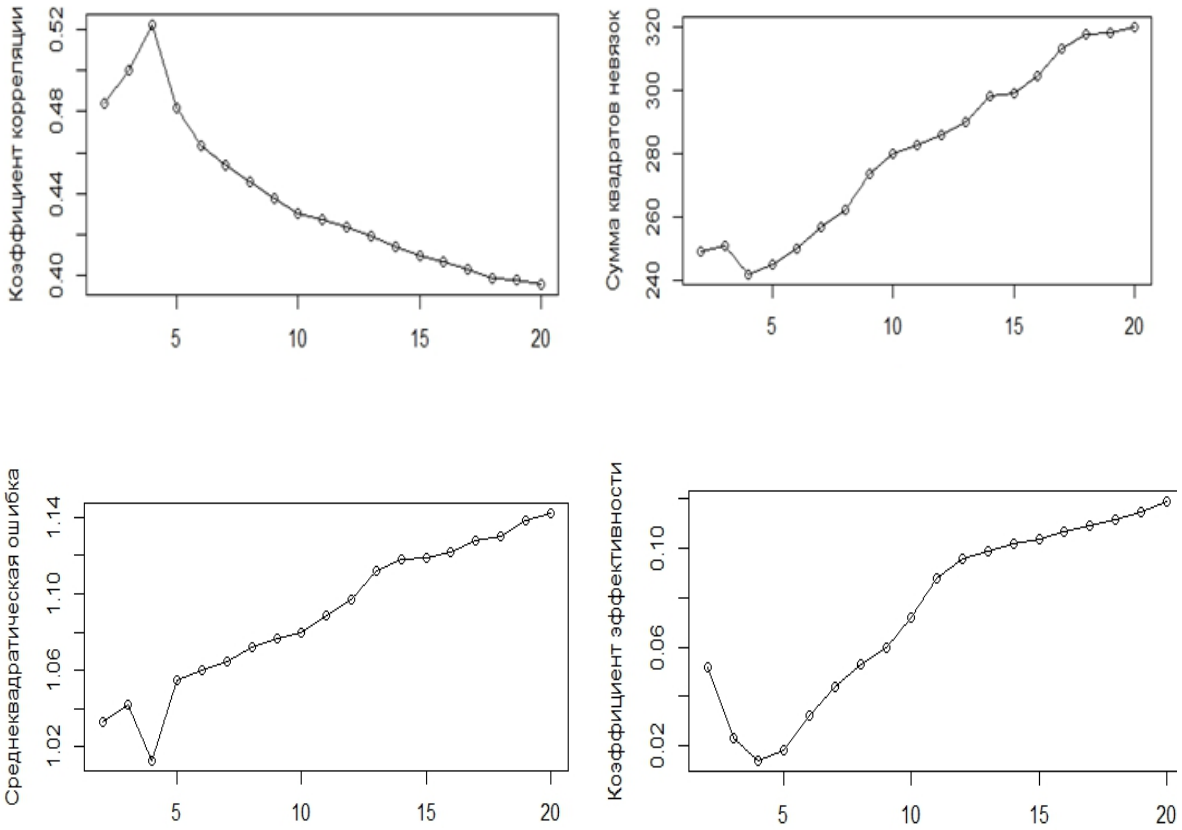

Рис. 2. Зависимость характеристик качества оценки, полученной методом обратных взвешенных расстояний, от параметра Power

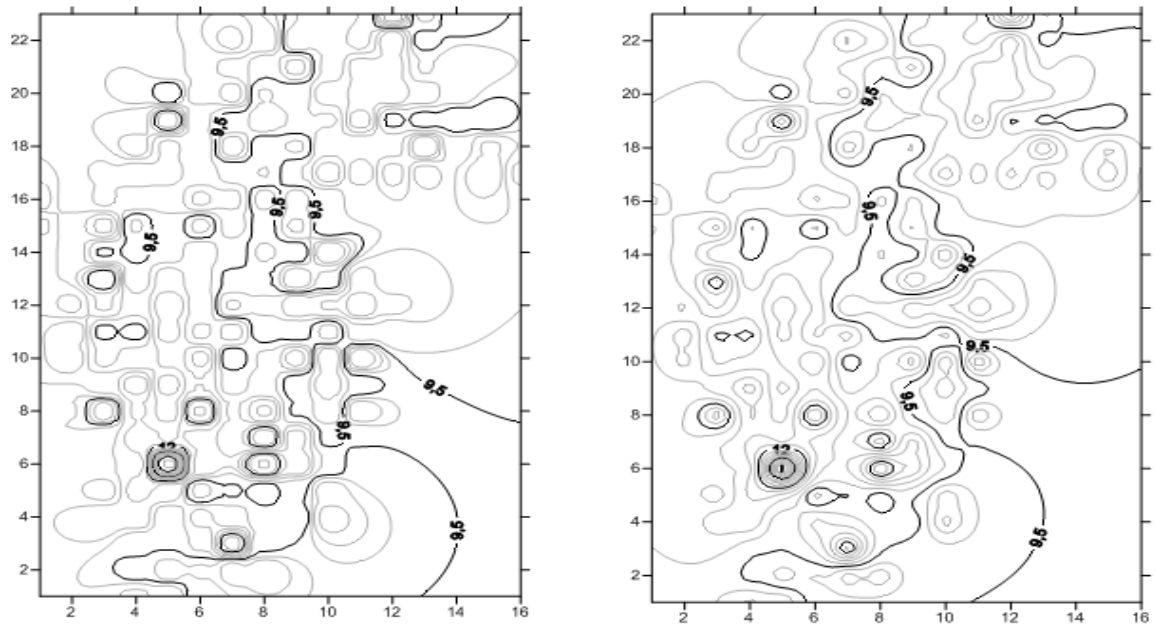

Рис. 3. Изолинии, полученные методом обратных взвешенных расстояний с параметром Power = 4 и методом радиальных базисных функций с параметром Inverse Multiquadric

Следует отметить, что метод обратных взвешенных расстояний работает лучше в областях с низкой, средней и высокой обеспеченностью исходными данными.

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ

Применение различных детерминированных методов интерполяции дает различную степень точности полученных результатов. В таблице представлены значения характеристик качества оценок, построенных методами обратных взвешенных расстояний и радиальных базисных функций с их оптимальными параметрами.

Таблица

Характеристики качества оценок, полученные детерминированными методами с оптимальными параметрами

	r	Δ	S	$RSME$	E
Inverse Distance Weighting (Power = 4)	**0,522**	**0,001**	**242,112**	**1,013**	**0,014**
Radial Basis Functions (Inverse Multiquadric)	0,482	0,010	260,276	1,118	0,102

Из таблицы следует, что наилучшим по всем характеристикам является метод обратных взвешенных расстояний.

В R существует множество функций для построения 3D-графиков. Одна из них plot_ly() из библиотеки plotly. Результат ее использования представлен ниже.

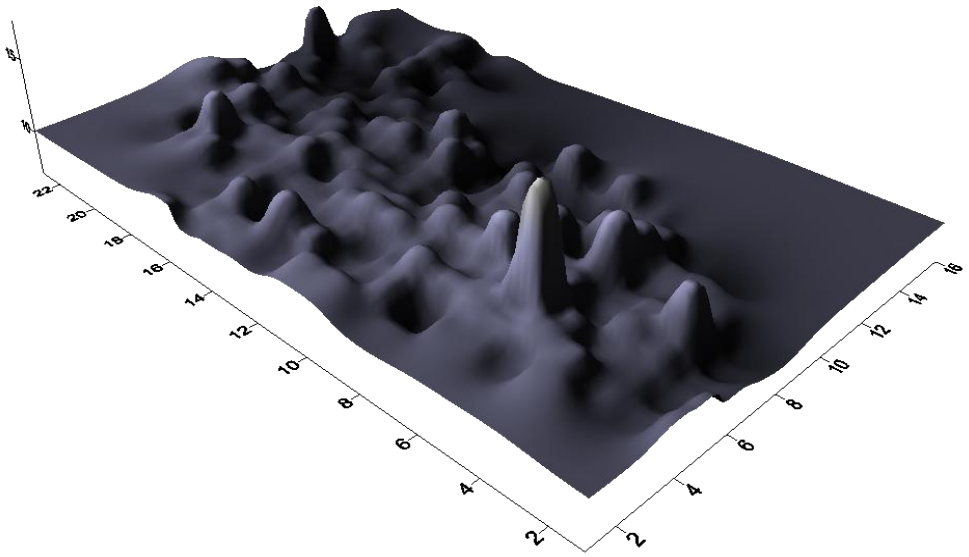

Рис. 4. Результат интерполяции методом обратных взвешенных расстояний с параметром Power = 4

В дальнейших исследованиях планируется осуществить прогнозирование как рядом других детерминированных методов интерполяции, так и геостатистическими методами, а также провести их сравнительный анализ.

БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. Noel A. C. Cressie. Statistics for spatial data. Cressie. Canada, Jonh Wiley & Sons, 1991. P. 32–34.

2. Демьянов В. В., Савельева Е. А. Геостатистика: теория и практика. М. : Наука, 2010. С. 54–64.