УДК 51 (075.8)

МЕТОДЫ И ИНСТРУМЕНТЫ МОДЕЛИРОВАНИЯ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА И ИНТЕРПРЕТАЦИИ ЦИФРОВЫХ ПОЛЕЙ

Ю. О. Ходос

Белорусский государственный университет, Беларусь, Минск, fpm.hodos@bsu.by

Рассматриваются методы и инструменты моделирования, интеллектуального анализа и интерпретации цифровых полей. Описан процесс подготовки на примере модели рельефа задаваемой формы эталонных площадных распределений, имитации съемки данных высоты поверхности на рассеянном множестве точек в системе компьютерной алгебры Wolfram Mathematica. Для сравнения приводятся два метода предобработки исходного набора точек с замерами высоты: среднее арифметическое и среднее геометрическое значений координат. Рассматривается кластерный метод анализа с различным числом кластеров, используя различные методы.

Ключевые слова: Wolfram Mathematica; цифровые поля; 3D визуализация; карты изолиний; кластеризация.

METHODS AND TOOLS FOR MODELING, INTELLECTUAL ANALYSIS AND INTERPRETING DIGITAL FIELDS

Y. O. Khodos

Belarussian state university, Belarus, Minsk, fpm.hodos@bsu.by

Method and tools of modeling, intellectual analysis and interpretation of digital fields are considered. The process of preparation is described using the example of a relief model of a given shape of reference area distributions, simulation of shooting surface height data on a scattered set of points in the Wolfram Mathematica computer algebra system. For comparison, two methods of preprocessing the initial set of points are given: the arithmetic mean and the geometric mean of coordinate values. A cluster analysis method with a different number of clusters using different methods is considered.

Keywords: Wolfram Mathematica; 3D visualization; contour maps; digital field; clustering.

Введение

Математическое моделирование процессов, явлений в разных областях естествознания, экономики приобретает всё более широкое распространение. При этом одной из ключевых является задача цифрового опи-

сания пространственных объектов, их структуры и свойств. Например, при решении задач математического моделирования объектов геологии, подземной гидродинамики, экологии развивается концепция, следуя которой ядром и теоретической основой для построения компьютерных моделей является цифровое описание распределений изучаемых параметров на выбранном пространственном слое. Таковым параметром может быть, например, концентрация загрязняющего вещества в слое почвы, температура в зоне лесного пожара, насыщенность нефтью пласта. Считается, что исходными данными при этом являются значения наблюдаемого параметра в точках с известными геометрическими координатами, а сами точки с замерами могут быть размещены на площади нерегулярно. В цифровом описании значения параметра восстанавливаются на равномерной прямоугольной сетке. Подобные цифровые поля – не что иное, как сеточные функции, а с ними можно работать средствами численного анализа.

Моделирование и подготовка данных

На примере эталонной функции

$$fXY(x,y) = e^{-(x-1)^2-(y-2)^2} - 2/3 * e^{-(x-7)^2-(3y-3)^2}$$

рассмотрим визуализацию результатов и моделирование подготовки данных.

Подготовим цифровые данные для анализа. Для начала сгенерируем трёхмерные массивы, каждая строка которых будет являться координатами одной точки. Таких массивов будет 4: для 50, 100, 150 и 250 точек. Генерировать значения по переменной *x* будем в диапазоне от 5 до 15, а по переменной *y* – от 0 до 4. Для систематизации упорядочения по плотности размещения точек будем рассматривать различные разбиения этих двух отрезков, то есть будем рассматривать различные равномерные сетки. Причем, в случаях для нескольких точек, попавших в ячейку такого разбиения, будем оставлять одну точку в ячейке сетки, используя формулы расчета: среднее арифметическое и среднее геометрическое координат точек, с последующим уточнением вычисленных значений уровня. Для удобства воспроизведения данных проинтерполируем ([1]) полученные массивы точек с помощью встроенной функции ListInterpolation [2, 3]. Для вычисления границ разбиений сеток по координатам *x* и *y* будем использовать следующие формулы:

$$arX(i,n) = aX + i * ((bX - aX)/n),$$

$$arY(j,m) = aY + j * ((bY - aY)/m),$$

где *aX*, *aY* – минимальные значения координат *x* и *y*; *bX*, *bY* – максимальные значения; *n*, *m* – количества разбиений по *x* и *y*; *i*, *j* – номер разбиения.

Предобработка, прореживание исходных данных, алгоритмы среднего арифметического и среднего геометрического значений координат

Визуально сопоставлялись контурные графики смоделированных массивов данных для различного числа точек и различных разбиений. С учётом вывода, сделанного в ходе работы, что представительным разбиением из рассмотренных (на 20 частей по *x* и на 10 частей по *y*, на 50 частей по *x* и на 25 частей по *y*, на 100 частей по *x* и на 50 частей по *y*) является разбиение 100x50, продемонстрируем только его (рис. 1). По тем же причинам рассмотрим только случаи 50 и 250 точек.

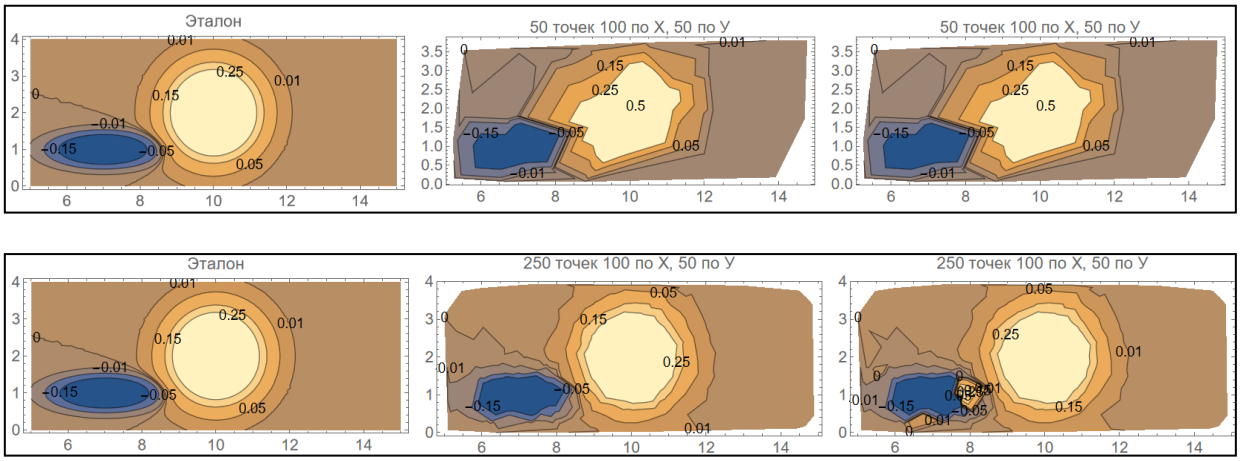

Рис.1

Слева отображён эталон, по центру – контурный график массива, обработанного алгоритмом среднего арифметического значения точек, справа – контурный график массива, обработанного алгоритмом среднего геометрического значения точек. В двух случаях получили сходимость к эталону, но в случае со средним арифметическим не возникает лишних возмущений поверхности, как в случае с 250 точками. В случае малого количества точек получили идентичные результаты для двух методов.

Применение кластеризации для интерпретации цифровых полей

В данном пункте рассмотрим различные способы кластеризации цифровых полей. Кластеризация (кластерный анализ) это задача группировки объектов таким образом, что более похожие друг на друга (по некоторым признакам) объекты окажутся в одном кластере, а различные – в разных. Кластеризация применяется в том случае, когда мы не знаем о том, как организованы данные.

В качестве эталона будем использовать контурный график той же функции, только уберём заливку фрагментов. Рассмотрим три вида кластеризации (по двум параметрам, по трём параметрам и только по третьему параметру) четырьмя методами (метод по умолчанию, KMeans, Spectral, KMedoids) [2, 3] для различного числа кластеров. Исходя из вывода, сделанного в ходе работы, отобразим только вариант 5 кластеров. Начнём с двух параметров (рис.2).

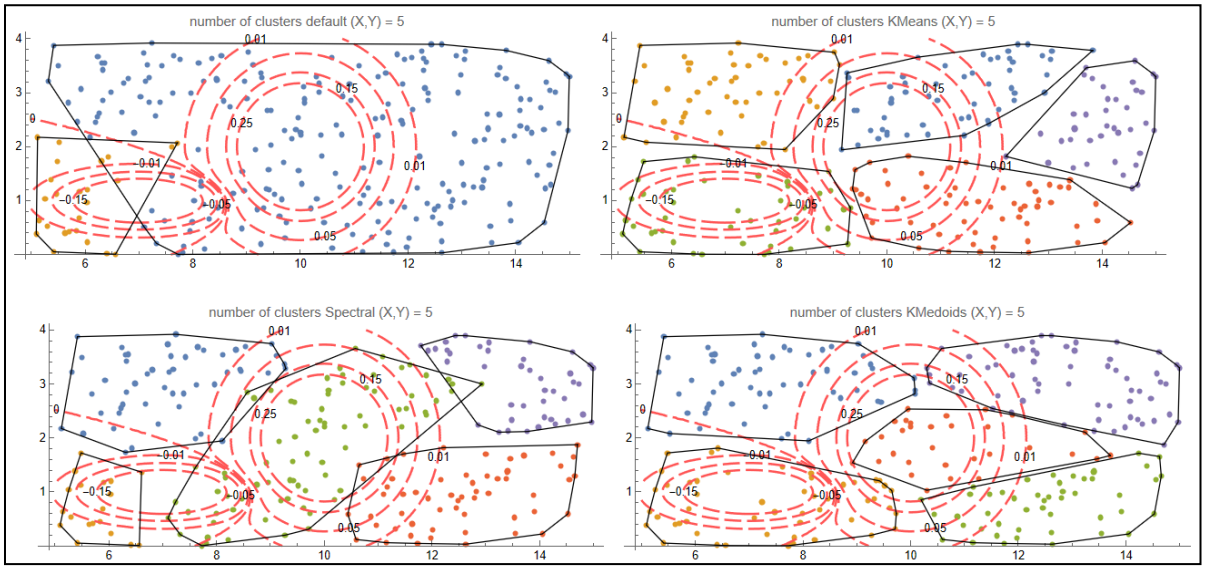

Рис. 2

Далее покажем результаты для трёх параметров (рис. 3).

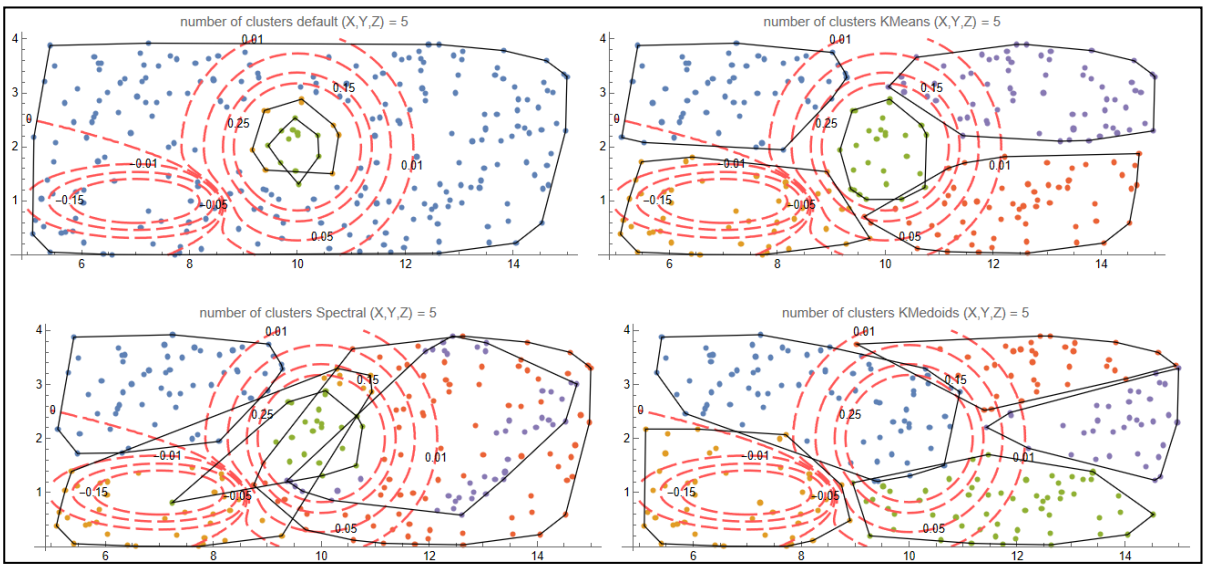

Рис. 3

И, наконец, только для третьего параметра (рис.4).

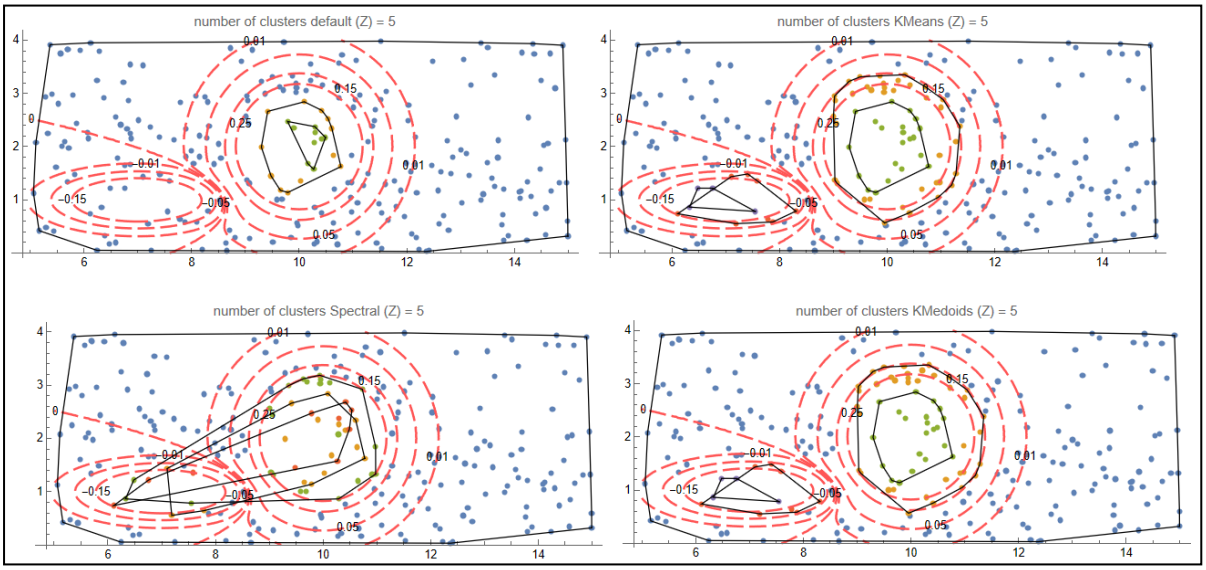

Рис. 4

Можно видеть, что лучше справляются с задачей методы KMeans и KMedoids. Окаймление в нескольких приведенных вариантах требует уточнений.

Получаемые результаты кластеризации являются базой для алгоритма интерполяции итогового цифрового распределения путем избирательного интерполирования по узлам в кластерах и последующего синтеза составных полей на всю площадь.

Заключение

Приведён пример генерирования массива числовых данных замеров уровня поверхности, его обработка и графическая интерпретация. При предобработке массива исходных точек использованы два метода: среднее арифметическое и среднее геометрическое значений координат с дальнейшей интерполяцией. Для анализа сходимости рассмотрены случаи с различным количеством точек и различными разбиениями отрезков. Показана сходимость к эталону при двух методах обработки. Приведены результаты кластерного анализа цифровых данных замеров в вариантах с разным числом кластеров и использованием различных методов, метрик.

Библиографические ссылки

1. *Морозов, А. А.* Программирование задач численного анализа в системе Mathematica: учеб. Пособие / А.А. Морозов, В.Б. Таранчук – Мн.: БГПУ, 2005. – 145 с.

2. *Таранчук, В. Б.* Методы и примеры интеллектуальной обработки данных для геологических моделей / В.Б. Таранчук // «Научные ведомости Белгородского государственного университета. Серия: Экономика. Информатика»: – 2019. – № 3 том 46. – С. 511–522. DOI 10.18413/2411-3808-2019-46-3-511-522.

3. Каталог «WOLFRAM Demonstrations Project» – https://demonstratioms.wolfram.com/