

Глава 2. Детерминистические методы пространственной интерполяции

Е.А. Савельева, В.В. Демьянов, С.Ю. Чернов

Детерминистические методы традиционно широко используются в различных областях прикладной и научной деятельности: метеорологические интерполяции, включая объективный анализ ветровых полей, анализ радиологического и химического загрязнения, топографические интерполяции и другие.

При использовании детерминистических методов предполагается, что анализируемые данные описываются некоторой детерминистической функцией $V(x)$, определенной на исследуемой области (S) , где $x \in S$ – координата точки. Задача состоит в том, чтобы, базируясь на известных данных ($V_i = V(x_i)$ – значения измеренные в точках $x_i \in S$) и на другой контекстной информации об исследуемом явлении, построить функцию $V(x)$ для всей исследуемой области S . После этого значение в любой точке исследуемой области может быть просто вычислено по формуле.

К настоящему времени разработано множество различных детерминистических методов пространственной интерполяции. К ним относятся полигонные методы, линейные и полиномиальные интерполяторы и некоторые другие методы [1–11].

Триангуляция Делоне и полигоны Вороного (Тиссена, ячейки Дирихле)

Один из простейших детерминистических методов пространственной интерполяции – это интерполяция по ближайшему соседу. Этот метод состоит в том, что для любой точки исследуемой области в качестве значения функции в ней выбирается такое, которое имеет ближайшая (по Евклидовому расстоянию) точка измерения. Таким образом, чтобы полностью определить функцию в исследуемой области, для каждой точки с известным значением строится область влияния – полигон Вороного (ячейка Дирихле) [3, 4]. Область влияния точки X_i – это геометрическое место точек в пространстве координат $A(X_i)$, таких, что $\forall x \in A(X_i), \forall j \neq i: ||X_i, x|| < ||X_j, x||$. Всем точкам, принадлежащим области влияния $A(X_i)$, присваивается значение исходной заданной точки X_i .

Функция, полученная методом полигонов, является сильно разрывной. Разрыв наблюдается на границе каждой области влияния. Существуют методы, позволяющие сглаживать разрыв за счет использования еще нескольких соседей. Одним из таких методов является метод линейной интерполяции на треугольниках. Вся исследуемая область разбивается на треугольники с вершинами в точках X_i и непересекающимися ребрами. Искомая функция представляется как совокупность плоскостей, определяемых заданными значениями в вершинах треугольников. Такая функция точно воспроизводит значения в исходном множестве точек и является непрерывной во всей области.

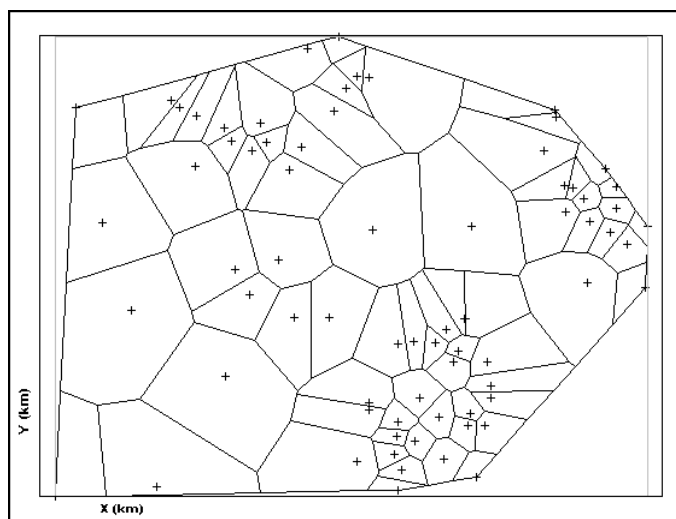


Рис. 1. Пример полигонов Вороного для множества из 73-х точек.

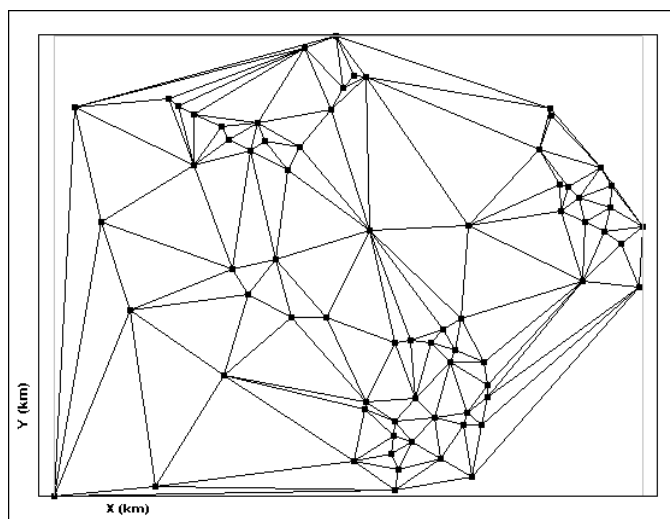


Рис. 2. Триангуляция Делоне (система ближайших соседей) для 73-х точек.

Линейные интерполяторы

Линейные интерполяторы представляют искомую функцию в виде линейной комбинации известных значений:

$$V^*(x) = \sum_{i=1}^{N(x)} W_i(x) V(x_i) \quad (1)$$

где $V^*(x)$ – оцениваемое значение в точке x , $V(x_i)$ – известные значения в точках x_i , $N(x)$ – количество исходных точек, принимающих участие в оценке для координаты x , $W_i(x)$ – весовые коэффициенты.

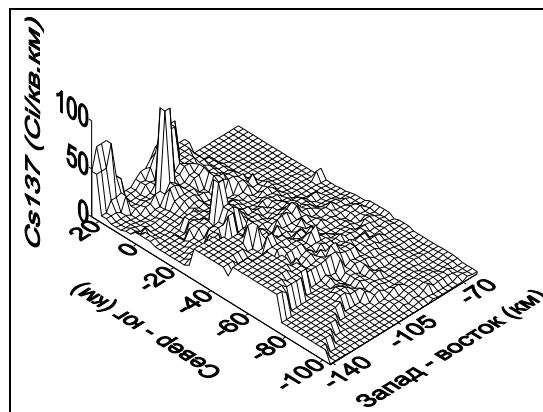


Рис. 3. Результат интерполяции по методу ближайшего соседа.

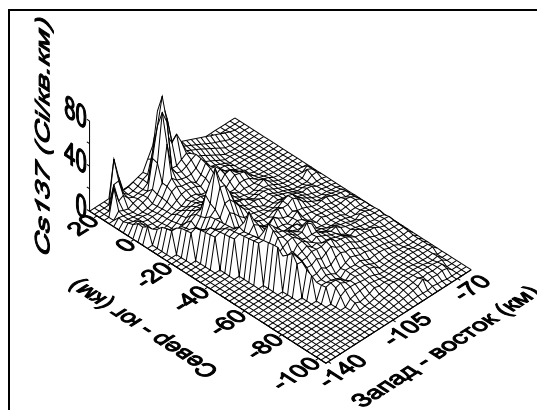


Рис. 4. Результат линейной интерполяции по триангуляции.

Линейные интерполяторы могут быть глобальными — все точки с известными значениями используются при интерполяции, или локальными — только часть значений в точках, ближайших к оцениваемой, используются для интерполяции. Глобальные интерполяторы делают искомую функцию более гладкой, локальные позволяют выявлять некоторые локальные особенности. Возможны различные подходы к применению локальных интерполяторов. Может быть фиксировано число ближайших к оцениваемой точке соседей, которые разрешается использовать при интерполяции: $N(x)=N=const$. Тогда размер зоны, влияющей на значение в точке x , зависит от плотности точек измерения в ее окрестности. Возможно, наоборот, задание размера области (радиус поиска D), точки из которой используются при оценивании значения в x . В этом случае $N(x)$ будет меняться и в области с редкими измерениями при малом значении D могут возникнуть не оцененные зоны (в D -окрестности точки x нет вообще никаких измерений).

Линейные интерполяторы различаются формой весовых коэффициентов. Например, описанный выше полигонный метод Тиссена можно свести к линейному интерполятору со следующими весовыми коэффициентами:

$$W_i(x) = \begin{cases} 1, & x \in A(x_i) \\ 0, & x \notin A(x_i) \end{cases} \quad (2)$$

Широко используются весовые коэффициенты, обратно пропорциональные степени расстояния между точками ($p > 0$):

$$W_i(x) = \frac{1}{R_i^p} \left(\sum_{i=1}^{N(x)} \frac{1}{R_i^p} \right)^{-1}, \quad (3)$$

где $R_i = \|x - x_i\|$ – расстояние от оцениваемой точки до i -ой точки исходного набора.

Радиусы поиска могут быть явно введены в весовые коэффициенты. Например, это сделано в *весовых коэффициентах Крессмана*:

$$W_i(x) = \frac{D_i^2 - R_i^2}{D_i^2 + R_i^2}, \quad (4)$$

где D_i – радиус влияния i -ой точки. Радиус влияния может быть задан постоянным по всей исследуемой области, а может меняться в зависимости от локальной плотности точек измерения. При использовании весовых коэффициентов Крессмана их рекомендуется нормализовать: $\sum_{i=1}^{N(x)} W_i(x) = 1$.

Очень широко распространен в метеорологии *метод Барнса/Ахтмейера*. При использовании этого метода делается предположение, что искомая поверхность может быть представлена в виде суммы бесконечного числа гармонических волн. Весовые коэффициенты вычисляются как экспоненты от параметров, связанных с плотностью исходного набора точек.

Для вычисления весовых коэффициентов при оценке значения в точке x множество точек измерения предварительно упорядочивается по возрастанию величины расстояния до оцениваемой точки. Весовой коэффициент для i -ой в этом списке точки задается следующей формулой:

$$W_i(x) = \exp\left(-\left(\frac{a_i \delta}{\beta}\right)^2\right), \quad (5)$$

где δ – радиус поиска, нормированный на среднее расстояние между точками измерений, а a_i и β определяются следующим образом:

$$a_i = \left(\frac{i}{\pi}\right)^{\frac{1}{2}} \delta, \quad \beta = \left(\frac{c}{2.3}\right)^{\frac{1}{2}} \frac{1}{N} \sum_{i=1}^N \left(\frac{1}{\pi}\right)^{\frac{1}{2}},$$

где c – параметр, определяющий форму волны. Для получения несмещенной оценки весовые коэффициенты рекомендуется нормализовать.

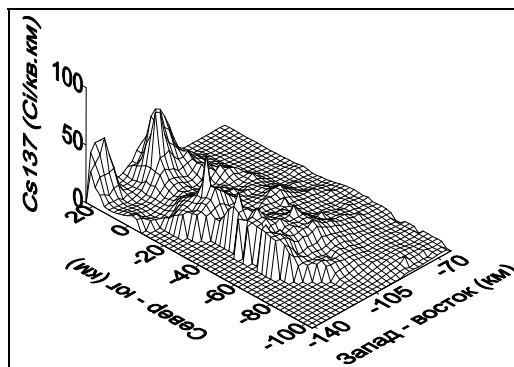


Рис. 5. Результат интерполяции по методу обратных квадратов.

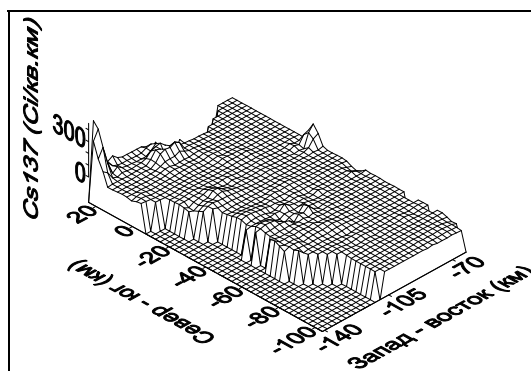


Рис. 6. Результат интерполяции методом кубического сплайна.

Полиномиальные методы

Полиномиальные интерполяторы представляют значение в точке в виде полинома от координат. В двумерном случае для точки x с координатами (x, y) : $V^*(x, y) = P_n(x, y)$, где P_n – полином n -ой степени.

Обычно на практике для двумерного случая при полиномиальной интерполяции используется полином 2-ой степени:

$$P_2(x, y) = a_1 + a_2x + a_3y + a_4xy + a_5x^2 + a_6y^2. \quad (6)$$

Теоретически можно использовать и полиномы более высокого порядка, но это вызывает необоснованное усложнение задачи.

Задача полиномиальной интерполяции сводится к тому, чтобы определить неизвестные коэффициенты a_i , так, чтобы полиномы максимально хорошо соответствовали данным в заданных точках. Для этого находят минимум по всем коэффициентам a_i функции χ^2 , задающей интегральную ошибку интерполяции и определяемой следующим образом:

$$\chi^2 = \sum_{i=1}^N (V(x_i, y_i) - P_n(x_i, y_i))^2. \quad (7)$$

Минимизация сводится к решению системы 6-ти линейных уравнений с 6-ю неизвестными.

Мультикватерничные уравнения

Метод мультиквадратичных уравнений основан на попытке представить оцениваемую поверхность, как сумму квадратичных поверхностей:

$$V^*(x, y) = \sum_{i=1}^{N(x)} c_i [q(x_i, y_i, x, y)] \quad (8)$$

$q(x_i, y_i, x, y)$ – некоторая квадратичная форма, например круговой гиперboloид, тогда

$$q(x_i, y_i, x, y) = \sqrt{(x_i - x)^2 + (y_i - y)^2} + C, \quad (9)$$

где C – произвольный параметр, при $C=0$ получается конус. Вертикальные оси симметрии каждого квадратичного члена расположены в точках (x_i, y_i) . Коэффициент c_i определяет алгебраический знак вхождения соответствующего квадратичного члена и степень его влияния.

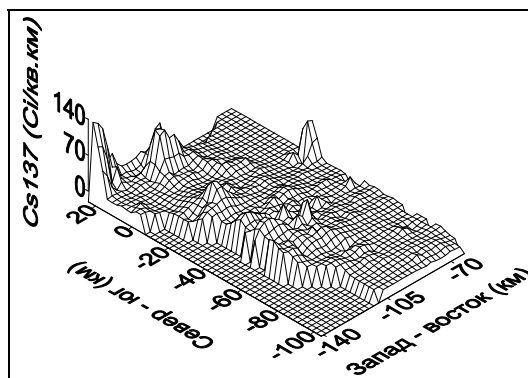


Рис. 7. Результат интерполяции методом мульти-квадратичных уравнений.

Коэффициенты c_i определяются из условия точности оценки в заданных точках, то есть во всех заданных точках (x_i, y_i) модель интерполяции должна давать оценку равную заданным значениям $V(x_i, y_i)$. Таким образом, для того, чтобы найти коэффициенты c_i необходимо решить систему из N линейных уравнений с N неизвестными:

[illegible]

Изначально метод был разработан, как глобальный, но его исследования показали, что он может быть использован и как локальный, т.е. $N=N(x)$ – количество точек исходных данных, участвующих в оценке функции в отдельных точках области, может быть меньше полного их числа. Отметим, что метод мультиквадратичных уравнений относится к классу интерполяторов, основанных на теории радиальных базисных функций. Это обеспечивает его универсальность и эффективность.

Кросс-валидация и сравнение методов

Разнообразные детерминистические методы с различными значениями параметров были применены к одним и тем же исходным данным по поверхностному загрязнению почвы ^{137}Cs в западной части Брянской области. Вычисления проводились на регулярной прямоугольной сетке 31×50 (размеры ячеек 2.61 км и 2.62 км по осям X и Y соответственно), с помощью пакета SURFER. Результаты вычислений представлены на рисунках 3-7. Из рисунков видно, что результаты интерполяции качественно отличаются для разных методов и различных параметров одного метода. Возникает вопрос – каким же результатам верить? Т.е. возникает проблема сравнения методов (и их параметров) и выбора наилучшего, причем не глобально, а для конкретных данных. Кроме того, анализ методов должен позволить оценить ожидаемые ошибки интерполяции.

Поскольку значения оцениваемой функции известны нам лишь для заданного набора точек, то сравнение степени соответствия прогноза реальным данным для различных методов приходится проводить на основе только этой информации. Одна из наиболее простых процедур для такого анализа носит название **кросс-валидация** и выполняется следующим образом:

1. Из базы данных временно изымается одна точка и проводится оценка значения в ней с использованием всех остальных данных.
2. Полученное значение сравнивается с известным и рассчитываются относительная ошибка и отклонение (невязка) – разница между измеренными и оцененными значениями ($V - V^*$).
3. Первые два шага проводятся для всех точек базы данных.

Относительная ошибка вычисляется по формуле:

$$error(x_i, y_i) = 100 \frac{V(x_i, y_i) - V^*(x_i, y_i)}{V(x_i, y_i)}. \quad (11)$$

где $V^*(x, y)$ – оцененное значение в точке (x, y) , $V(x, y)$ – известное значение в этой же точке.

По результатам кросс-валидации могут быть построены карты невязок, которые дают общее представление в каких областях метод дает более точную оценку. Для примера на рисунке 8 приведена карта невязок для метода мультиквадратичных уравнений. По ней видно, что для рассмотренных данных метод мультиквадратичных уравнений с вычислениями по 5-ти ближайшим соседям дает большие ошибки по краям исследуемой области.

Сравнивая карты невязок для различных методов, можно оценить, какой метод лучше и для какой области.

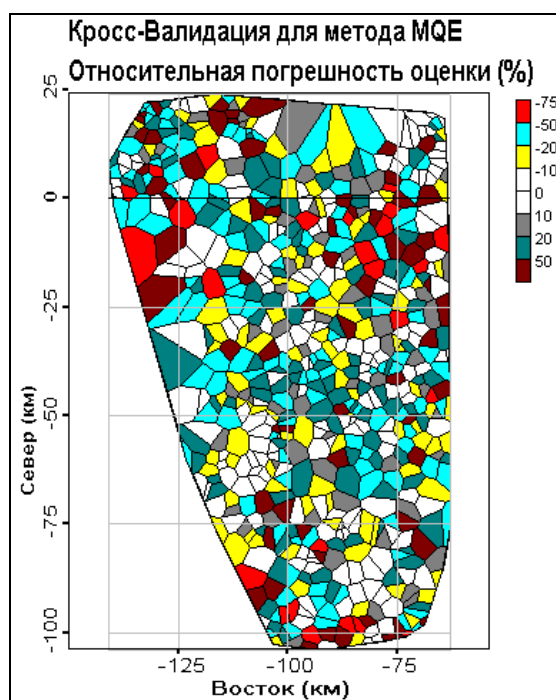


Рис. 8. Обратные невязки кросс-валидации (V^*-V) для метода мультиквадратичных уравнений по 5 ближайшим соседям, представление в виде полигонов Вороного.

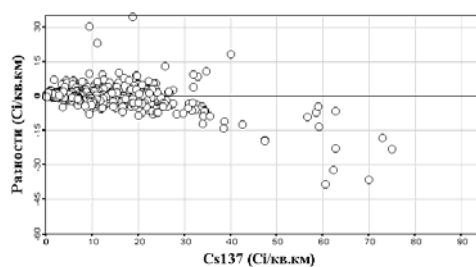


Рис. 9. Разброс обратных невязок кросс-валидации (V^*-V) в зависимости от измеренного значения для метода обратных квадратов с радиусом поиска 8.75 км.

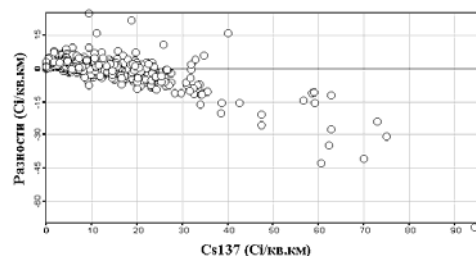


Рис. 10. Разброс обратные невязок кросс-валидации (V^*-V) в зависимости от измеренного значения для метода обратных квадратов с радиусом поиска 50 км.

Полезно при выборе параметров представить результаты кросс-валидации в виде графика – $Y(V(x,y))=V^*(x,y)$ или аналогичного ему – $Y(V(x,y))=\Delta(x,y)$. Проведение на графике биссектрисы (или соответственно прямой $F=0$), соответствующей равенству оценки и исходного значения, позволяет видеть характер отклонения: большее отклонение для больших или для маленьких значений V , какие-либо тренды в поведении оценки и т.п. Легко видеть, как влияют параметры (радиус поиска или число точек, использованных для оценки) на результаты различных методов. Вместе с тем, на всех графиках с невязками прослеживается эффект сглаживания – область низких значений в среднем переоценивается, а область повышенных значений недооценивается.

По рисункам 9 и 10 видно, что для исследуемых данных методу обратных квадратов больше подходит радиус поиска 8.75 км, чем 50 км. Для больших значений функции, которых не так много, невязки между оцененными и измеренными величинами существенно отличаются от нуля при использовании обоих значений параметров. Для маленьких и средних значений функции невязки, полученные с использованием радиуса поиска 8.75 км, расположены равномерно вокруг нуля. Для радиуса поиска 50 км, невязки расположены вокруг прямой, имеющей явный наклон.

Кроме вычисления характеристик в каждой точке можно вычислять и стараться минимизировать (с целью определения оптимальных параметров) глобальные характеристики метода по результатам кросс-валидации:

1. Сдвиг среднего значения $\Delta_m = m - m^*$, где m – среднее, оцененное по исходным данным, m^* – среднее, оцененное по полученным результатам.
2. Сумма квадратов невязок:

$$S = \sum_{i=1}^N (V(x_i, y_i) - V^*(x_i, y_i))^2 + R \quad (12)$$

Здесь R – слагаемое, которое играет роль штрафа (эквивалент регуляризующего члена). Причина его появления состоит в том, что при некоторых параметрах метода значения оценки $V^*(x_i, y_i)$ не могут быть получены (например, если задан слишком маленький радиус поиска и т.п.).

3. Средняя квадратичная ошибка (root mean square error – RMSE):

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (V(x_i, y_i) - V^*(x_i, y_i))^2} \quad (13)$$

4. Коэффициент эффективности $E = \frac{S}{S_0}$, где $S_0 = \sum (V(x_i, y_i) - m)^2$.

Для примера приведена таблица глобальных характеристик, полученных по результатам кросс-валидации, для метода мультиквадратичных уравнений с различными значениями параметра $N(x)$ (число ближайших соседей, учитывающихся при оценке) (см. Таблица 1).

Среднее значение оценок становится ближе к среднему значению исходных данных по мере роста числа ближайших соседей, учитывающихся при оценке (больше 15), но параметр эффективности растет. Это означает, что совпадение средних значений происходит не из-за улучшения оценки, а из-за простого сглаживания функции. Для лучшего выбора параметра рекомендуется провести анализ невязок.

Таблица 1. Глобальные параметры качества оценки кросс-валидации.

Число соседей, учитывающихся при оценке	Сдвиг среднего значения $m_{mes}-m_{est}$	Сумма квадратов отклонений S	Коэффициент эффективности E
5	-0.25	23776	0.72
10	-0.41	16783	0.8
15	-0.26	14865	0.83
20	-0.17	14795	0.83
40	-0.06	14660	0.82

Метод кросс-валидации дает возможность провести такой анализ. Известно, что метод интерполяции можно использовать тогда, когда его ошибки независимы, то есть не влияют одна на другую. Иначе при вычислении на сетке может возникнуть их неконтролируемый рост. Кроме того, ошибки должны быть нормально распределены со средним равным 0. Для выяснения этих вопросов можно провести статистический анализ невязок кросс-валидации..

Результаты статистического анализа невязок кросс-валидации между оцененным и заданным значениями, полученных в результате применения метода мультиквадратичных уравнений с различными параметрами, представлены в таблице. Они подтверждают сделанные выше выводы для рассматриваемых данных. Соответствие нормальному распределению невязок ухудшается с ростом числа использующихся ближайших соседей. Следовательно, лучшим параметром в рассмотренном примере следует считать число ближайших соседей N=5. Хотя всегда встает вопрос – чему отдать предпочтение – меньшему значению вариации или несмещенности оценки.

Таблица 2. Статистика невязок кросс-валидации для метода мультиквадратичных уравнений.

Число соседей учитывающихся при оценке	Среднее оценки (среднее исх.данных 11.61)	Среднее невязок	Вариация невязок	Симметрия невязок	Экссесс невязок
5	10.75	0.25	35.69	-2.23	57.3
10	10.91	0.41	25.69	2.77	57.32
15	10.89	0.26	22.28	4.75	66.0
20	10.99	0.18	22.23	4.15	59.49
40	11.1	0.06	22.04	4.02	59.41

Основываясь на алгоритме кросс-валидации, можно предложить процедуру автоматического подбора параметров конкретного метода интерполяции (АвтоКартирование), которые были бы наиболее оптимальными для данного набора исходных данных. В частности, для метода обратных степеней можно пытаться найти более оптимальный показатель степени. Далее можно ставить задачу выбора радиуса поиска (либо числа ближайших соседей), наиболее оптимального для заданного расположения точек и свойств оцениваемой

функции. Для выполнения такой процедуры Авто-Картирования необходимо сформулировать какой-либо критерий оптимальности оценки. Самым простым вариантом является интегральная квадратичная ошибка (сумма квадратов невязок) (12). Необходимо минимизировать эту величину как функцию параметров метода интерполяции. Опять же самый примитивный способ добиться этого – провести перебор по параметрам. На рисунке 11 представлен график зависимости интегральной невязки кросс-валидации от радиуса поиска для метода интерполяции по обратным квадратам расстояний для тех же данных по загрязнению поверхности земли ^{137}Cs в западной части Брянской области. На графике видно, что зависимость имеет минимум при значении радиуса около $D \approx 7\text{ км}$. При этом было наложено условие, что в D -окрестности точки оценивания должно находиться не менее 3-х точек исходных данных. Большие значения критерия при малых радиусах зоны поиска объясняются наличием штрафного члена в формуле (12). Этот штраф накладывается в ситуациях, когда из-за малости зоны поиска остаются неоцененные точки.

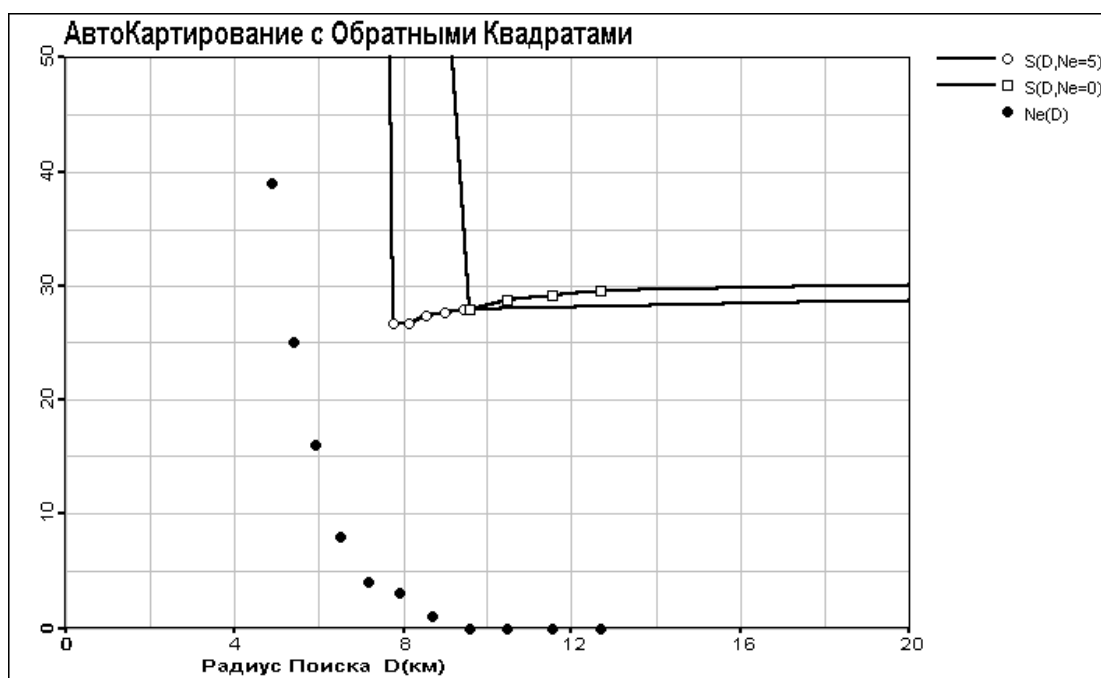


Рис.11. Зависимость среднеквадратичной ошибки кросс-валидации $S(D, Ne)$ от радиуса поиска D для метода обратных квадратов расстояний и количество неоцененных точек в зависимости от радиуса поиска, для данных по Чернобыльским выпадениям в Брянской области.

При таком подходе может случиться, что из-за особенности сети мониторинга оптимальным будет считаться радиус поиска, равный размеру всей оцениваемой области. В этом случае одна (или несколько) сильно удаленная от остальных точка множества исходных данных может экранировать своим штрафом влияние всех остальных точек сети мониторинга. На рисунке 11 черными кружками изображен график зависимости числа неоцененных точек $Ne(D)$ от радиуса поиска D . Видно, что для рассматриваемых данных критическим является радиус поиска $D \approx 9\text{ км}$.

Начиная с этого значения, данные покрывают своим влиянием всю область исследования. Однако, если модифицировать штрафную часть (R) критерия (12) в части (например, считать штраф равным нулю, если число неоцененных точек области меньше чем $M=1,2,...,n$), то можно еще уменьшить радиус поиска, и тем самым повысить локальность метода. В этом случае надо иметь ввиду, что могут остаться неоцененные зоны, дооценку которых можно выполнить отдельно. Для этих областей можно провести специальные вычисления с увеличенным радиусом поиска, разрешить оценивать такие точки не по 3-м (минимум) соседям, а по меньшему числу, в том числе и по одному единственному соседу и т.д. На рисунке 11 кривая, помеченная квадратами, соответствует случаю, когда наличие неоцененных точек не допускается ($N_e=0$). Кривая с белыми кружками соответствует $N_e=5$, т.е. при критерия (12) оценке разрешается оставить неоцененными до 5-ти точек. Это будет означать, что оптимальный радиус поиска будет выбран по оставшимся данным без оглядки на эти пять неоцененных точек. В зоне расположения этих точек для корректной оценки необходимо локально увеличить радиус поиска. К сожалению, пока трудно предложить универсальный рецепт проведения процедуры АвтоКартирования (т.е. полностью ее автоматизировать). Проведение визуального анализа сети мониторинга остается в большинстве случаев обязательным этапом такой процедуры.

В заключение отметим, что несмотря на эффективность и мощность метода кросс-валидации, пользоваться им необходимо с известной осторожностью. Являясь прекрасным качественным инструментом и индикатором качества модели, при количественной интерпретации результатов могут возникнуть сложности, особенно в случае сильной неоднородности или кластерности сети мониторинга. Эти проблемы возникают независимо от типа используемого и оптимизируемого метода.

В случае большого количества данных метод кросс-валидации может быть достаточно дорогим в вычислительном плане. В этом случае можно использовать метод “складного ножа” (jack-knife), когда N исходных данных разбиваются на M групп, в каждой из которых находится $L=N/M>1$ точек. При этом “кросс-валидация” проводится над M группами. Это позволяет значительно уменьшить вычислительное время. Обычная кросс-валидация определяется при $L=1$.

И, наконец, возможно проведение настоящей валидации, когда реальная оценка качества модели выполняется с использованием независимых представительных данных, не использовавшихся в процессе ее построения и настройки. Обычно такие независимые данные предоставляет заказчик разработок на этапе приемки в эксплуатацию готовых моделей.

Заключение

Широкий спектр детерминистических моделей интерполяции часто применяется в различных приложениях в силу их простоты в использовании и компьютеризации. Однако, детерминистические интерполяции дают только значение самой оценки и могут характеризовать ее качество. Детерминистические модели имеют ряд серьезных недостатков: игнорирование пространственной корреляции, необходимость подбора оптимальных

параметров метода. Кросс-валидация позволяет решить проблему оптимизации параметров модели и даже оптимизировать их подбор.

Литература

1. R. Franke. Scattered Data Interpolation: Test of Some Methods. *Mathematics of Computation*. V.38, No.157, 1982, pp.181-200.
2. Macidonio G., Pareschi M.T. An algorithm for the triangulation of arbitrarily distributed points: applications to the volume estimate and terrain fitting. *Computers & Geosciences*, V.17, No.7, 1991, pp.859-874.
3. G.Q. Tabois, J.D.Salas. A Comparative Analysis, for Spatial Interpolation of Precipitation. *Water Resources Bulletin*, V.21, No.3, 1985, pp.365-380.
4. R.Lattuada, J.Raper. Applications of 3D Delaunay triangulation algorithms in geoscientific modelling. <http://www.iah.bbscr.ac.uk/phd/gisruk95.html> .
- A. Okabe, B. Boot, K. Sugihara. *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*. J. Wiley & Sons, 1992, New York, ISBN 0 471 93430 5, 532p., A thorough and comprehensive survey and analysis of the theoretical and applications literature.
5. SURFER (Win32) Version 6.01, 1995, Surface Mapping System. Golden Software, Inc.
6. J.W.Grimm, J.A. Lynch. Statistical analysis of error in estimating wet deposition using five surface estimation algorithms. *Atmospheric Environment*, 1991, V.25A, p.317-127.
7. R.L.Hardy. Multiquadric-equations of topology and other irregular surfaces. *Journal of GeoPhysical Research*, V.76, 1971, pp.1905-1915.
8. D. Weber, E. Englund. Evaluation and Comparison of Spatial Interpolators. *Mathematical Geology*, V.24, No.4, 1992, pp.381-391.
9. R. L. Hardy. Theory and applications of the multiquadric-biharmonic equations method. *Computers Math. Appl.*, 1990, V.19, p.163-208.
10. H.C. Saunderson. Multiquadric surfaces in C. *Computers and Geosciences*, Vol.20, No.7/8, pp.1103-1122, 1994
11. P.M. Bartier, C.P. Keller. Multivariate interpolation to incorporate thetic surface data using inverse distance weighting. *Computers and Geosciences*, Vol.22, No.7, pp.795-799, 1996.
12. A.W. Moore, M.S. Lee. Efficient Algorithms for Minimizing Cross Validation Error. *Advances in Neural Information Processing Systems*, 7, Morgan Kaufmann, 1995.