Применение самоорганизующихся карт Кохонена для классификации и анализа пространственно распределенных неполных данных по окружающей среде

А.А. Трутце, Е.А. Савельева, В.В. Демьянов, М.Ф. Каневский, В.А. Тимонин, С.Ю. Чернов

ИНСТИТУТ ПРОБЛЕМ БЕЗОПАСНОГО РАЗВИТИЯ АТОМНОЙ ЭНЕРГЕТИКИ
113191, Москва, ул. Б. Тульская, 52
тел.: (095) 955-22-31, факс: (095) 958-11-51, эл. почта: dargot@ibrae.ac.ru,
http://www.ibrae.ac.ru/~mkanev/

Содержание

1 Введение

В настоящее время с связи с развитием информационных технологий появилась возможность сбора и хранения больших массивов информации, в том числе и по измерениям загрязнений окружающей среды (измерение большого набора различных переменных). Визуальное представление и анализ таких наборов затруднен и требует развития специальных методов. Одной из возможных альтернатив являются само организующиеся карты Кохонена. За счет классификации данных по внутреннему сходству они позволяют снизить размерность пространства и упрощают интерпретацию и понимание данных.

В данной работе делается попытка рассмотреть возможность применения метода Самоорганизующихся Карт Кохонена для классификации пространственно распределенных коррелированных данных по загрязнению окружающей среды и оценки неполных данных, используя методику ассоциативной памяти. Анализ проводился на данных по загрязнению западной части Брянской области радиоактивными изотопами ^{137}Cs и ^{90}Sr, выпавшими в результате аварии на Чернобыльской АЭС. Про изотопы ^{137}Cs и ^{90}Sr известно, что они коррелированны [1].

2 Теория Самоорганизующихся Карт Кохонена

2.1 Описание Метода Самоорганизующихся Карт Кохонена

В настоящее время для решения различных классов задач (аппроксимация функций, классификация данных, кластеризация данных, сжатие информации, восстановление утраченных данных и т. д.) широко применяются искусственные нейронные сети (ИНС). ИНС представляют структуру для параллельной обработки данных при помощи нелинейных элементов (нейронов), которые обладают локальной памятью и выполняют нелинейную локальную операцию по обработке информации. Нейроны связаны между собой связями, которым приписаны синаптические веса. Они представляют собой все знания ИНС. Процесс обучения сети – просто процесс корректировки синаптических весов. На данный момент известны различные методы обучения нейросетей: обучение с "учителем" – сеть обучается на множестве примеров с известным ответом (пар входов-выходов), с "подкреплением" – известна лишь оценка (границы) выходов сети и без "учителя" – сеть обучается только лишь на наборе входных значений.

Существуют различные методы обучения нейросетей без учителя – Хеббовское обучение, автоассоциативное обучение, соревновательное обучение. Метод Самоорганизующихся Карт Кохонена является одним из вариантов соревновательного обучения[2].

Нейронные сети Кохонена характеризуются тем, что в них нейроны представляют собой двумерный массив, каждому узлу i которого поставлен в соответствие вектор $m_i = [\mu_{i1}, \mu_{i2},..., \mu_{in}]^T$, имеющий размерность равную размерности пространства входных векторов R^n. Тип массива может быть различен – прямоугольный, гексагональный и т.д. Как отклик сети на входной вектор $x \in R^n$ понимается узел-победитель c, расстояние (которое может быть определено различным образом, обычно Евклидово $|x-m_c|$) до которого от x минимально. Для выбора узла c вектор x сравнивается со всеми узлами m_i, и $c=argmin_i\{|x-m_i|\}$, т.е.

$$|x-m_c|=min_i\{|x-m_i|\}. \quad (1)$$

Начинается обучение с задания начальных векторов $m_i(0)$. Обычно они определяются, как случайные значения с равномерным распределением в диапазоне значений соответствующей компоненты входных векторов. Если есть априорные знания о распределении значений векторов, их можно использовать при задании начальных значений векторов, и это способствует улучшению сходимости метода обучения. В процессе обучения вектора m_i, соответствующие узлам i меняют свои значения в соответствии с поданным в данный момент времени t на вход вектором $x(t)$ согласно правилу:

$$m_i(t+1)= m_i(t)+h_{ci}(t)[x(t)- m_i(t)], \quad (2)$$

где t – дискретная координата времени, а $h_{ci}(t)$– функция соседства, играющая в процессе обучения центральную роль. Для сходимости метода обучения необходимо, чтобы $h_{ci}(t)\to 0$ при $t\to\infty$. Обычно:

$$h_{ci}(t)=h(|r_c-r_i i|,t), \quad (3)$$

где $r_c \in R^2$, $r_i \in R^2$ – вектора узлов c и i соответственно. С увеличением $|r_c-r_i|$, $h_{ci}\to 0$. Чаще всего используются два простых варианта для $h_{ci}(t)$. Простейший из них – "множество соседства" точек массива вокруг узла-победителя c. Пусть множество их индексов будет обозначено как N_c, и $h_{ci}(t)=\alpha(t)$, если $i \in N_c$ и $h_{ci}(t)=0$, если $i \notin N_c$:

$$m_i(t+1)= m_i(t)+ \alpha(t)[x(t)- m_i(t)],\ i \in N_c$$

$$m_i(t+1)=0,\ i \notin N_c \quad (4)$$

(4)

Величина $\alpha(t)$ называется фактором обучения (скоростью обучения) ($0<\alpha(t)<1$). $\alpha(t)$ и радиус $N_c(t)$ обычно монотонно убывает со временем.

Другой широко распространенный вариант задания $h_{ci}(t)$ – в виде Гауссовой функции:

$$h_{ci}(t)=\alpha(t)*exp(-|r_c-r_i|^2/2\sigma^2(t)), \quad (5)$$

где $\alpha(t)$ – скалярный фактор обучения (скорость обучения) и $\sigma(t)$–параметр. И $\alpha(t)$, и $\sigma(t)$ – некоторые монотонно убывающие функции от времени.

Обучение можно разделить на два этапа – грубой и тонкой настройки координат векторов, соответствующих узлам карты. В этом случае во время первого этапа вектора упорядочиваются, во втором – вектора в каждой группе уточняются. Второй этап обычно имеет меньшую скорость обучения и большее число итераций. Оба этапа обучения длятся заданное заранее число шагов T, а $\alpha(t)$, как правило, подбирают так, чтобы $\alpha(T)=0$, например $\alpha(t)= \alpha (1-t/T)$, где α – заданныя константа.

По окончании обучения мы получаем обученную карту, представляющую собой упорядоченный двумерный массив, вектора, соответствующие узлам которого распределились в пространстве нейронов R^n в соответствии с подаваемым на вход множеством векторов X. Данный метод напоминает натягивание эластичной мембраны на данное множество векторов X, причем эластичность ее постепенно увеличивается в процессе натягивания (уменьшается $h_{ci}(t)$) для все более и более тонкой настройки.

Упорядочение векторов в виде двумерной карты выражается в том, что чем ближе координаты двух векторов на карте, тем ближе они и в пространстве R^n, но не наоборот.[3]. Значения векторов в узлах сети соответствуют характерному (возможно усредненному) значению для соответствующего класса.

В [2] приводится математическое обоснование данного подхода для одномерного массива нейронов, обоснования для двумерного случая на сегодняшний день не существует[4].

В настоящее время основными типами задач, решаемыми с помощью Карт Кохонена, являются визуализация многомерных нелинейных данных и предобработка данных. Они применяются в задачах распознавании речи, изображений, в управлении роботами, медицинской диагностике, классификации финансовых данных и многих других областях[2]. Всего по приложению самоорганизующихся карт Кохонена опубликовано более 2000 работ [5].

2.2 План Работы по анализу данных

В работе проводился анализ данных по загрязнению почкы с использованием Самоорганизующихся карт Кохонена. Для работы имеющиеся исходные данные были разбиты на два набора, первый из которых используется для обучения (обучающий), а второй для тестирования (тестовый).

Для инициализации карты (задание начальных значений узлам сети) использовались случайные числа, равномерно распределенные в диапазоне значений входных данных.

Обучение карты проводилось по алгоритму соседства ($h_{ci}(t)= \alpha(t)$, если $i \in N_{ci}$ и $h_{ci}(t)=0$, если $i \notin N_{ci}$) в два этапа – грубой и тонкой настройки координат векторов, соответствующих узлам карты. На первом этапе – вектора упорядочиваются, на втором – вектора в каждой группе уточняются до “точных” значений. На втором этапе обычно используют меньшую скорость обучения, чем на первом, меньший радиус соседства N_c и большее число итераций.

Обучение проводилось несколько раз с различными значениями параметров обучения – варьировались радиусы соседства для обоих этапов обучения и для каждого набора значений параметров обучения использовалось несколько вариантов случайной начальной инициализации.

Окончательный вариант обученной сети выбирался по наименьшей средней ошибке квантования. Средняя ошибка квантования вычисляется следующим образом: для каждого вектора из набора для тестирования обучения вычисляется ошибка его отнесения к некоторому классу, то есть расстояние от вектора до соответствующего ему вектора узла победителя, а затем проводится усреднение по всему набору. Таким образом выбирается сеть наилучшим образом классифицирующая этот набор. Так как у нас мало точек для обучения и тестирования метода в качестве набора для теста обучения использовался набор, на котором проводилось обучение. Вообще говоря, это не очень хорошо, так как очень хорошее воспроизведение набора для обучения может быть вызвано оверфитингом.

Для обученной карты проводилась визуализация, то есть процедура постановки в соответствие всем векторам, подаваемым на вход соответствующих им узлов нейронной сети (узлов победителей). Таким образом мы можем в случае полных данных классифицировать их. В случае с неполными данными (вектора с пропущенными компонентами) процедура визуализации позволяет восстановить отсутствующие компоненты по ассоциативной памяти, то есть выбирается узел победитель при условии, что в качестве пропущенной компоненты используется соответствующая компонента узла сети. При визуализации на вход обученной карте подается набор данных, а на выходе получается множество узлов карты, которым соответствуют тестовые вектора и также вычисляется ошибка квантования.

Для обученной карты Кохонена проводилось тестирование в несколько этапов:

1. Визуализировался обучающий набор данных, то есть те данные на которых проводилась тренировака сети. Это осуществлялось для определения качества обучения карты, ее способности

правильной интерпретации данных, на которых она училась. Вычислялись ошибки классификации, той же самой ошибки квантования, для каждого входного вектора, а так же ошибок по отдельным координатам, то есть соответствия значений в узлах сети исходным данным. Соответствие анализировались с помощью построения гистограмм (соответствие функции распределения) и вариограмм (сходство пространственной корреляции).

2. Визуализировался тестовый набор данных с полными входными векторами (все значения заданы). Это проводилось для более глубокого тестирования способности карты Кохонена классифицировать данные, так как тестовый набор с одной стороны, не использовался при обучении, а с другой представляет собой набор данных того же явления. Проводился полный анализ результатов на соответствие функции распределения и вариограмм тестового набора данных и соответствующих им узлов сети Кохонена.
3. Визуализировался тестовый набор данных с неполными входными векторами (с пропущенными значениями ^{137}Cs и ^{90}Sr по отднльности; ^{137}Cs и ^{90}Sr одновременно). Это выполнялось для исследования возможности дополнения пропущенных данных ^{137}Cs и ^{90}Sr по ассоциативной памяти. После получения оценок ^{137}Cs и ^{90}Sr вычислялись невязки и относительные ошибки, а для них анализировались статистические гистограммы и вариограммы.

После поведения анализа, на основании полученных результатов, делались общие выводы о возможности и целесообразности применении исследуемого алгоритма для классификации и анализа и дополнения неполных данных.

Для обучения и применения сети Кохонена использовался программный пакет SOM_PAK Version 3.1[6]. Пакет содержит все программы, необходимые для применения традиционного алгоритма Самоорганизующихся Карт, разработанного Кохоненом: инициализация карты (задание начальных векторов), обучения (модификация векторов), вычисление ошибки квантования (ошибки отнесения к классу) и визуализации (классификация).

3 Визуализация и первичная обработка данных

3.1 Набор Данных для обучения

В качестве набора данных для обучения и для тестирования были взяты данные о радиоактивном загрязнении Брянской области изотопами ^{137}Cs и ^{90}Sr. Известно, что эти данные коррелированны [1]. Следовательно должна быть возможность понизить размерность данных.

Для выделения обучающего набора использовался метод декластеризации, позволяющий выбирать точки равномерно по всей области координат. Это важно для того чтобы в обучающем наборе были представители из всех частей исследуемой области. Оставшиеся после декластеризации данные использовались в тестовом наборе данных. Таким образом получилось 2 набора данных: обучающий – 164 точки и тестовый – 209 точек, каждый из которых состоял из четырехмерных векторов (1-я и 2-я координаты – ламбертовские координаты точки замера, 3-я и 4-я – концентрация изотопов ^{137}Cs и ^{90}Sr соответственно). Данные из обучающего набора представлены на Рис. 3.1.1-3.1.2

Рис. 3.1.1 Распределение концентрации ^{137}Cs в обучающем наборе данных

Реальные значения всех компонент входных векторов были преобразованы в отрезок [0,1] с использованием равномерного линейного преобразования. Ламбертовские координаты были преобразованы: по X – из диапазона [-150;-50] в [0;1]; по Y – из диапазона [-100;25] в [0;1]. Значения загрязнения изотопами ^{137}Cs были преобразованы из диапазона [0;100] в [0;1], а ^{90}Sr – из диапазона [0;1.5] в [0;1]. Преобразование проводилось для приведения значений всех полей к общему диапазону значений [0;1], чтобы предотвратить доминирования одних компонент векторов над другими при вычислении Эвклидового расстояния, как следствие при обучении[7]. Преобразованные данные представлены на Рис. 3.1.3-3.1.4

Так как в нашем случае ожидается стремление к концентрации в узлах сети усредненных значений как координат, так и ^{137}Cs и ^{90}Sr, для последующего анализа и сравнения результатов проведено вычисление локальных средних преобразованных значений ^{137}Cs и ^{90}Sr (для обучающего и для тестового наборов данных). Результаты представлены на Рис. 3.1.5-3.1.6

Как мы можем видеть из Рис. 3.1.1 и 3.1.2 в обучающем наборе данных присутствует явная координатная зависимость (тренд) значений ^{137}Cs и ^{90}Sr. Кроме того, наблюдаются локальные максимумы, причем в некоторых случаях они совпадают по местоположению для ^{137}Cs и ^{90}Sr.

Рис. 3.1.2 Распределение концентрации ^{90}Sr в обучающем наборе данных

Рис. 3.1.3 Преобразованные значения концентрации ^{137}Cs в обучающем наборе данных

Вычисление и анализ локальных средних проводился, так как исходя из характера обучения (тип метрики), обученная карта по значениям в узлах должна быть похожа именно на локальные средние.

Плюсами на карте помечены координаты реальных данных для последующего изучения корреляции координат узлов обученной карты с координатами исходных данных.

Рис. 3.1.4 Преобразованные значения концентрации ^{90}Sr в обучающем наборе данных

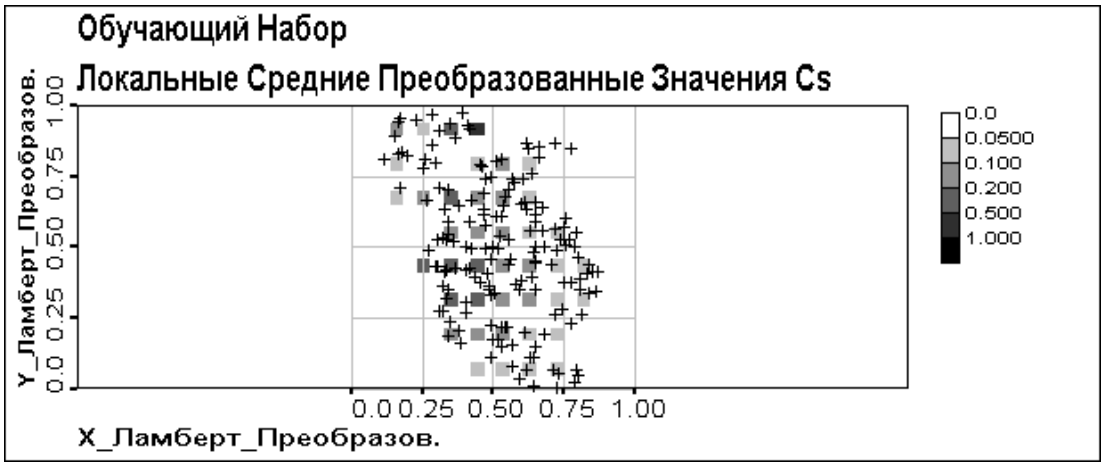

Рис. 3.1.5 Локальные средние значения ^{137}Cs для обучающего набора данных

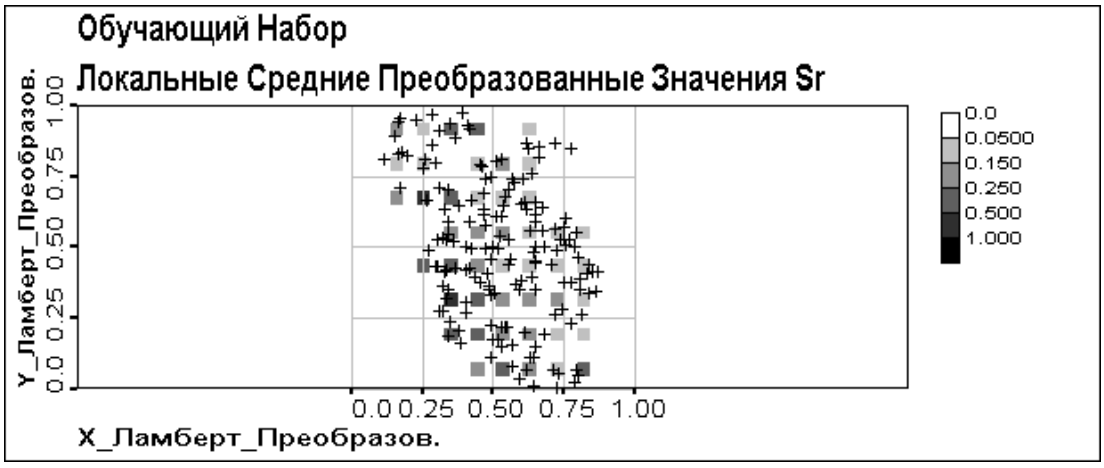

Рис. 3.1.6 Локальные средние значения ^{90}Sr для обучающего набора данных

3.2 Тестовый Набор Данных

На Рис. 3.2.1-3.2.2 можно видеть распределение концентраций ^{137}Cs и ^{90}Sr в тестовом наборе данных, на Рис. 3.2.3-3.2.4 – распределение преобразованных значений концентраций, а на Рис. 3.2.5 - 3.2.6 – локальные средние.

Рис. 3.2.1 Распределение концентрации ^{137}Cs в тестовом наборе данных

Рис. 3.2.2 Распределение концентрации ^{90}Sr в тестовом наборе данных

Рис. 3.2.3 Преобразованные значения концентрации ^{137}Cs в тестовом наборе данных

Как можно видеть из этих рисунков в тестовом наборе данных наблюдается еще более четкая координатная зависимость значений ^{137}Cs и ^{90}Sr, чем в обучающем наборе данных. Кроме того, наблюдается заметно меньшее количество точек с низкими значениями, они были вынуты процедурой декластеризации в обучающий набор. Это также приводит к различиям в статистических и пространственно корреляционных характеристиках обучающего и тестового наборов, которые будут показаны ниже.

Рис. 3.2.4 Преобразованные значения концентрации ^{90}Sr в тестовом наборе данных

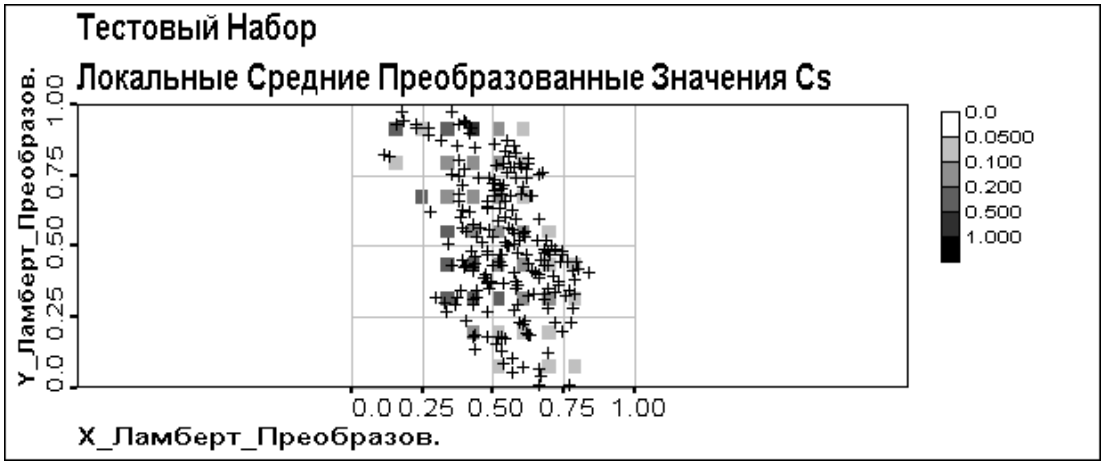

Рис. 3.2.5 Локальные средние значения ^{137}Cs для тестового набора данных

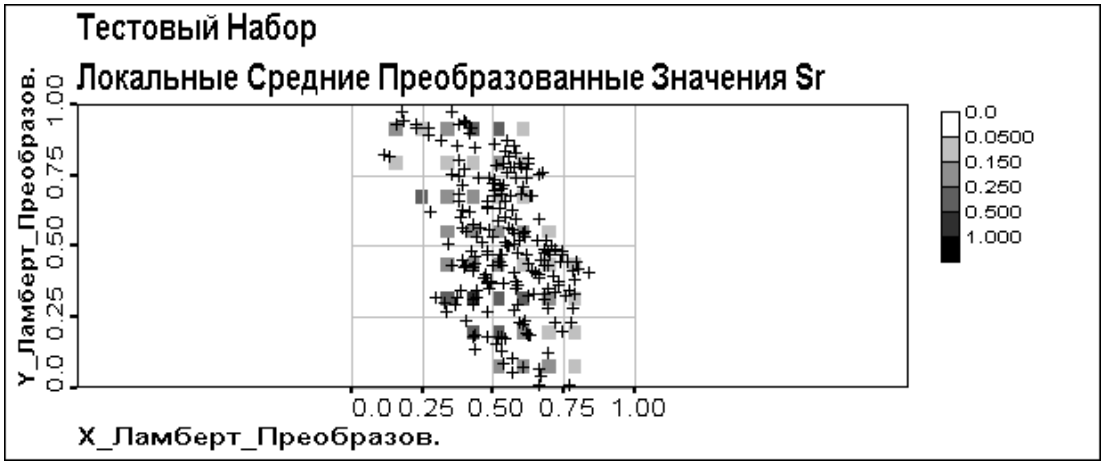

Рис. 3.2.6 Локальные средние значения ^{90}Sr для тестового набора данных

3.3 Сравнение Обучающего и Тестового Наборов

Для корректного объяснения результатов работы Самоорганизующейся Карты Кохонена в данном разделе проводится сравнительный анализ статистического распределения и пространственной корреляции обучающего и тестового наборов и их сравнение с полным набором данных. Понимание расхождения между тестовым и обучающим наборами позволит правильно проинтерпретировать полученные результаты. Анализ проводится на основе сравнения вариограмм и статистических гистограмм преобразованных значений концентраций ^{137}Cs и ^{90}Sr. (см. Рис. 3.3.1-3. 3.4)

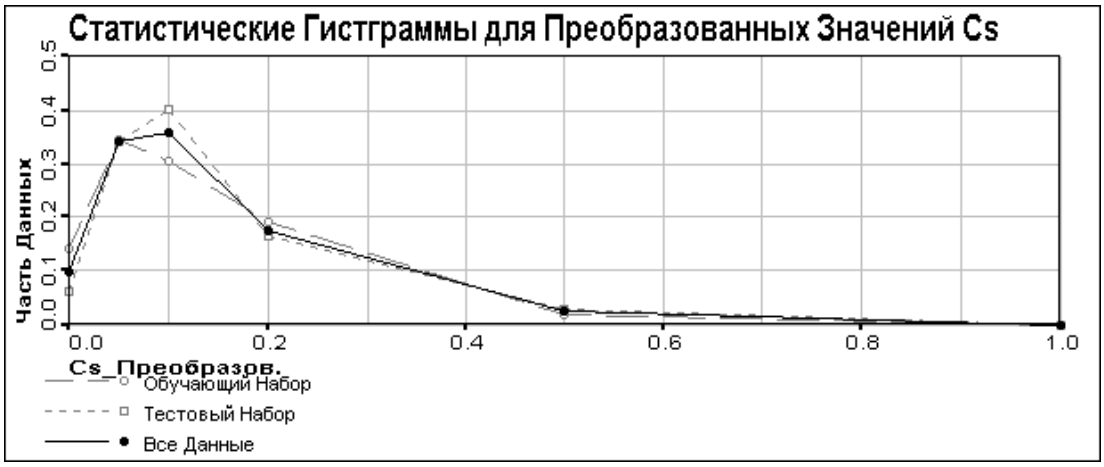

Рис. 3.3.1 Расхождения в распределении ^{137}Cs в обучающем и тестовом наборах данных

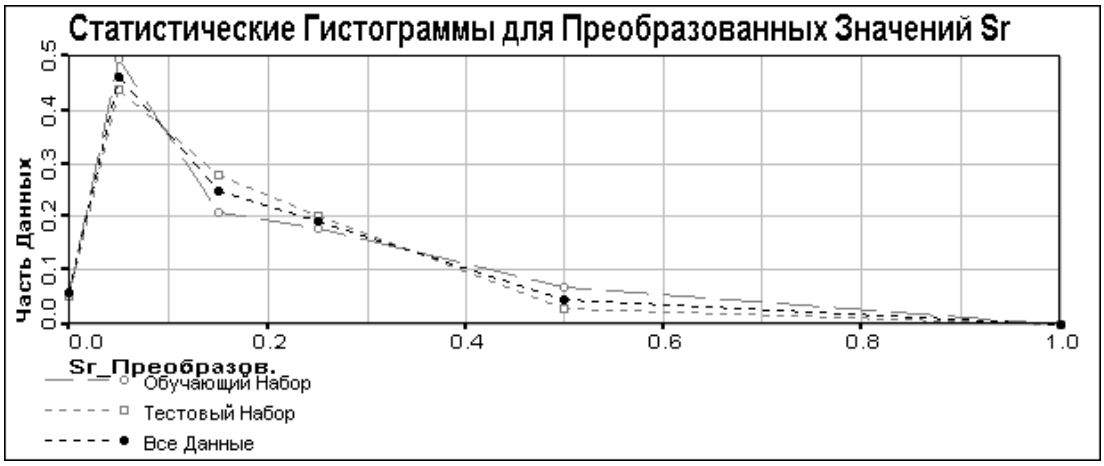

Рис. 3.3.2 Расхождения в распределении ^{90}Sr в обучающем и тестовом наборах данных

Как можно видеть из гистограмм, статистическое распределение ^{137}Cs и ^{90}Sr в обучающем и тестовом наборах данных примерно одинаково, особенно ярко это видно на данных по ^{90}Sr. По ^{137}Cs наблюдается некоторое смещение пика тестового набора вправо относительно обучающего набора, но, так как у полного набора данных пик там же, где и у тестового, этот факт объясняется ошибками декластеризации. (см. п. 3.1)

Из представленных вариограмм можно сделать следующие выводы о пространственной корреляционной структуры обучающего и тестового наборов данных:

- по Cs – плато (sill) и эффективный радиус корреляции (range) – у обучающего и у тестового наборов данных одинаковы и совпадают с таковыми у всех данных. Наггет (nugget) – разный. Априорная вариация примерно одинакова.

- по Sr – эффективный радиус корреляции одинаков, но плато различно. Наггет также различен. Кроме того, априорная вариация обучающего набора данных больше, чем у тестового и у полного наборов данных. Вероятно этот факт имеет место быть вследствие того, что в обучающий набор данных попали большинство локальных максимумов. Это же подтверждается средним положением априорной вариации для всех данных. При моделировании вариограммы его можно использовать плато для обоих наборов.

Отсюда можно делать выводы, что оба набора данных можно использовать в качестве обучающего и тестового наборов, но необходимо помнить, что имеются некоторые расхождения, и эти расхождения по ^{137}Cs меньше, чем по ^{90}Sr.

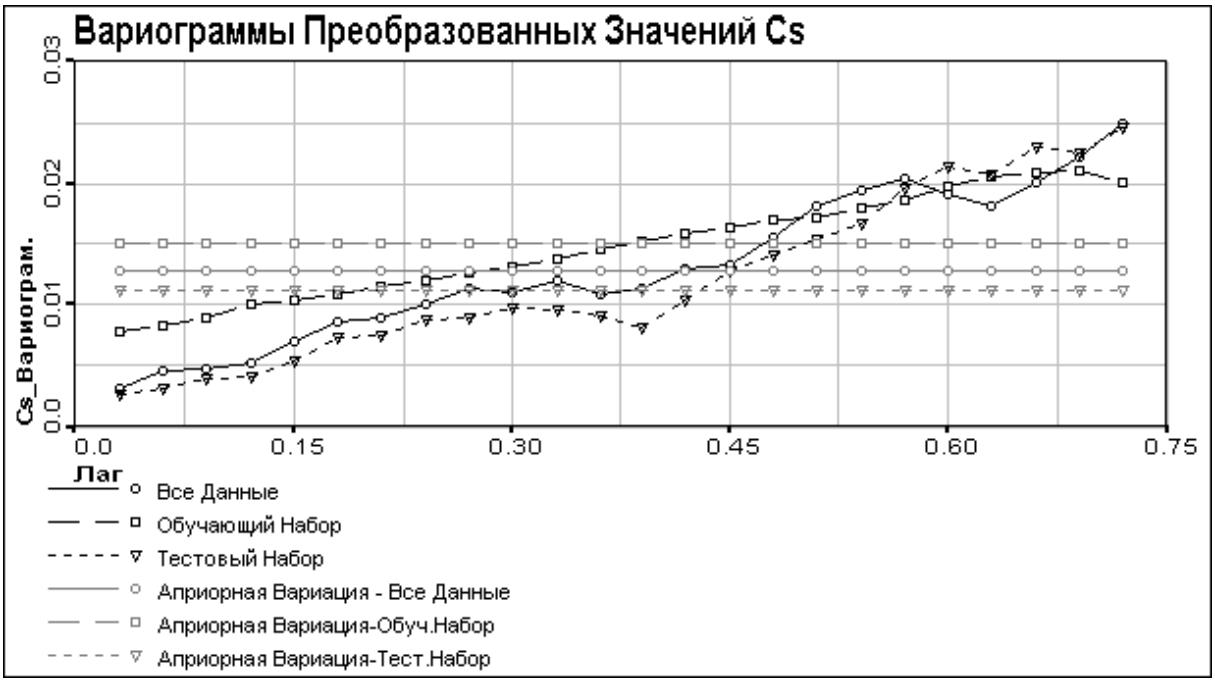

Рис. 3.3.3 Сравнение вариограмм ^{137}Cs для обучающего и тестового наборов данных

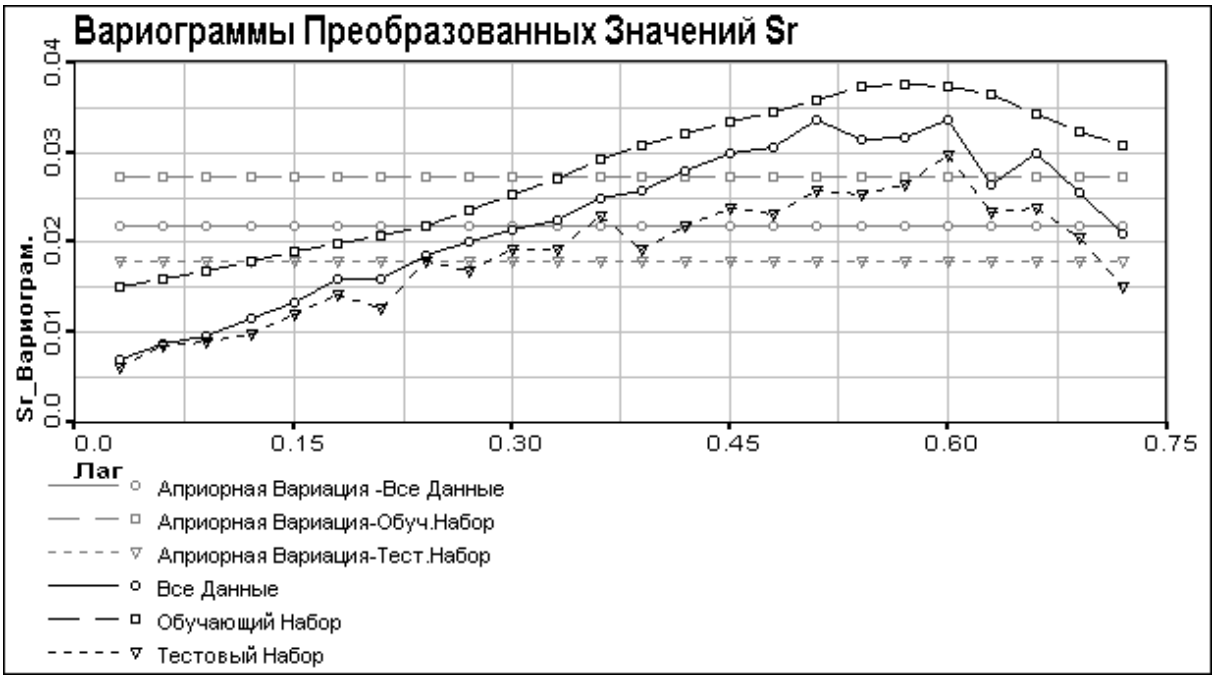

Рис. 3.3.4 Сравнение вариограмм ^{90}Sr для обучающего и тестового наборов данных

4 Анализ реальных данных

4.1 Обучение Самоорганизующейся Карты Кохонена

4.1.1 Схема обучения и тестирования

Была проведена серия экспериментов. В каждом из них обучение проводилось на обучающем наборе с различными параметрами для обучения карты Кохонена. В результате этих экспериментов предполагалось выбрать параметры обучения, позволяющие получить наилучшим образом обученную карту. Критерием качества обучения служила наименьшая средняя ошибка квантования (см. часть. 2.2).

Во всех экспериментах обучалась сеть в 800 узлов (40×20) с прямоугольной топологией (каждый узел имеет 4 соседних). Данный размер сети определялся структурой и количеством входных данных – они имеют большую протяженность в направлении север-юг (см. Рис. 3.1.5-3.1.6), вследствие чего сеть тоже лучше делать вытянутую в этом направлении.

Начальные значения векторам, соответствующим узлам карты задавались случайным образом в диапазоне [0;1] для каждого поля данных.

Обучение сети проводилось в 2 этапа – грубой и тонкой модификации координат узлов. Все обучение осуществлялось на наборе данных для обучения. Для всех экспериментов при обучении использовались следующие параметры:

на первом этапе обучения–

- 5000 циклов обучения
- $\alpha(0)$=0.05 и линейно убывает к 0

на втором этапе обучения –

- 20000 циклов обучения
- $\alpha(0)$=0.01

Так как данных для обучения меньше, чем циклов обучения, то одни и те же данные подавались на вход несколько раз. Во время каждого цикла обучения входной вектор выбирался случайным образом из всего набора данных для обучения.

Всего было проведено 9 экспериментов. В каждом брались различные значения r: при грубой и тонкой модификации координат. Для грубой модификации использовались значения r 100, 50 и 20. Для тонкой – 10, 5 и 3. Для каждой пары параметров проводилось 3 сеанса обучения и вычислялась средняя ошибка квантования в каждом сеансе. За основу для дальнейших экспериментов использовалось 50 соседей при грубом и 3 при тонком обучении, как более оптимальный, то есть имеющий наименьшую среднюю ошибку квантования, постоянную для всех 3-х сеансов. Ее значения колеблются в диапазоне от 0.043058 (наилучший результат) до 0.068446.

После обучения сети с использованием оптимальных параметров обучения строились карты сети и проводилась визуализация сети. При этом компоненты 4-мерных векторов узлов сети рассматриваются по смысловому содержанию, как соответствующие компоненты входных векторов, то есть 1-ая и 2-ая координаты, 3-ья–^{137}Cs, 4-ая–^{90}Sr. Такая интерпретация позволяет представлять полученные Карты Кохонена в реальных координатах, используя свои первые колонки, как координаты.

Визуализация проводилась по обучающему(см. часть 4.2) и по тестовому (см. часть 4.3) наборам данных. Визуализация тестового набора данных проводилась в 4-х вариантах – со всеми известными полями (см. часть 4.3.1), с неизвестным ^{137}Cs(см. часть 4.3.2), с неизвестным ^{90}Sr(см. часть 4.3.3), с неизвестными ^{137}Cs и ^{90}Sr(см. часть 4.3.4).

Визуализация обучающего набора данных выполнялась с целью проверки качества обучения. Визуализация тестового набора данных проводилась со следующими целями: со всеми известными полями – валидации обученной карты и для последующего сравнения с визуализацией неполных данных; неполных данных – собственно для получения оценок (см. часть 2.2) отсутствующих данных.

По оценкам, полученным в результате визуализации вычислялись невязки (разницы между исходными и оценочными значениями) и относительные ошибки, строились вариограммы оценок для последующего сравнения их с вариограммами исходных данных и вариограммы невязок для определения их пространственной корреляции.

4.1.2 Результат обучения

После обучения с оптимальными выбранными параметрами была получена карта Кохонена. На рис. 4.1.2.1-4.1.2.6 представлены распределения значений векторов узлов сети:

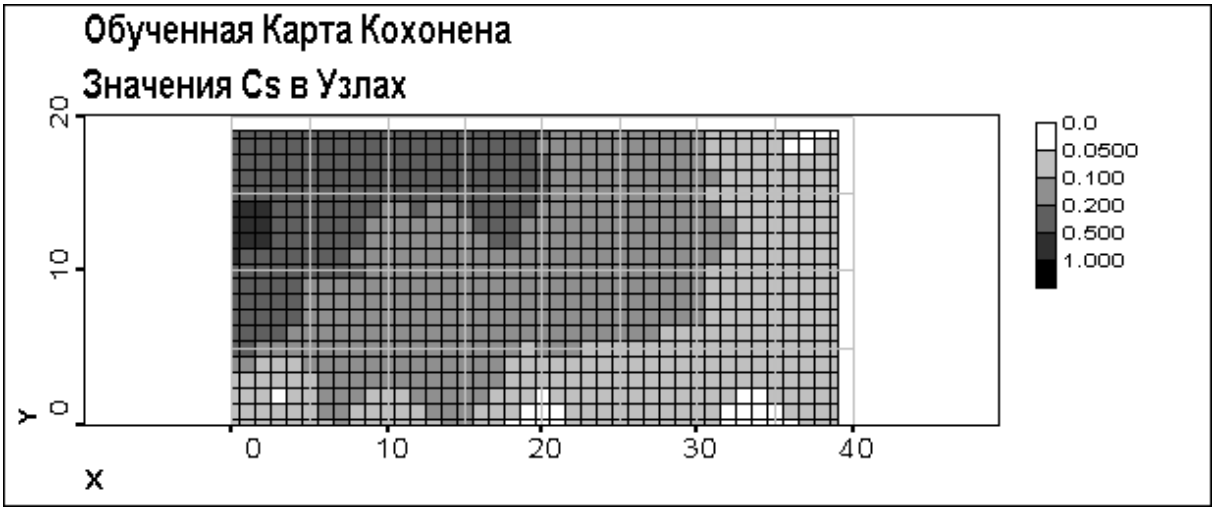

Рис. 3.1.2.1 Значения ^{137}Cs в узлах обученной карты

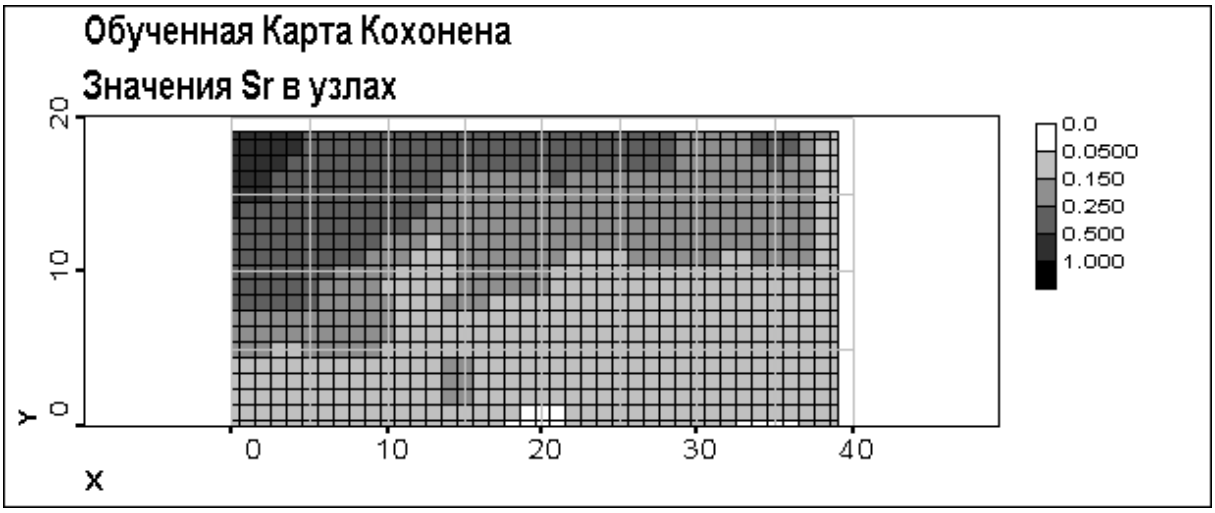

Рис. 3.1.2.2 Значения ^{90}Sr в узлах обученной карты

Из Рис. 3.1.2.1-3.1.2.4 видно, что в процессе обучения значения в узлах карты упорядочились по всем компонентам и в соответствии между собой. Можно выделить несколько классов узлов по значениям компонент и направлениям их изменения. Очевидна связь между значениями различных компонент вектора, соответствующего узлу карты (например, при X_Ламбертовское > 0.8, ^{137}Cs < 0.1, можно привести и другие связки).

На Рис. 4.1.2.5-4.1.2.6 представлена карта Кохонена, если 1-й и 2-й компоненты векторов, соответствующих узлам сети считать координатами. Плюсами помечены координаты векторов узлов сети. Из этих рисунков видно воспроизведение картой Кохонена структуры исходной сети мониторинга – группировка векторов узлов сети вокруг исходных данных, и областей высоких и низких значений и направления их изменения – выделение области более высоких и более низких значений ^{137}Cs и ^{90}Sr.

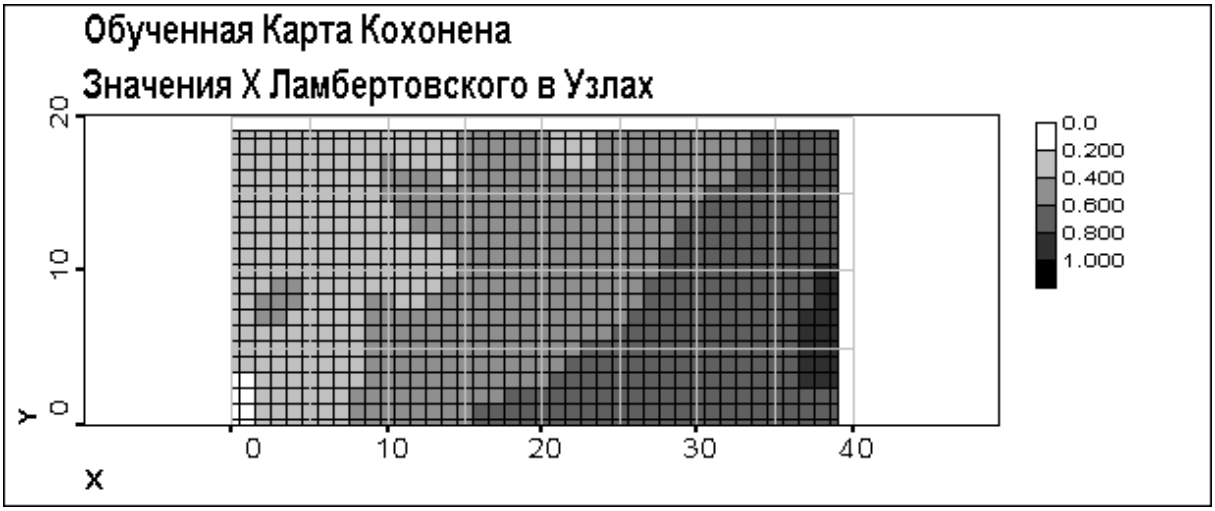

Рис. 3.1.2.3 Значения X Ламбертовского в узлах обученной карты

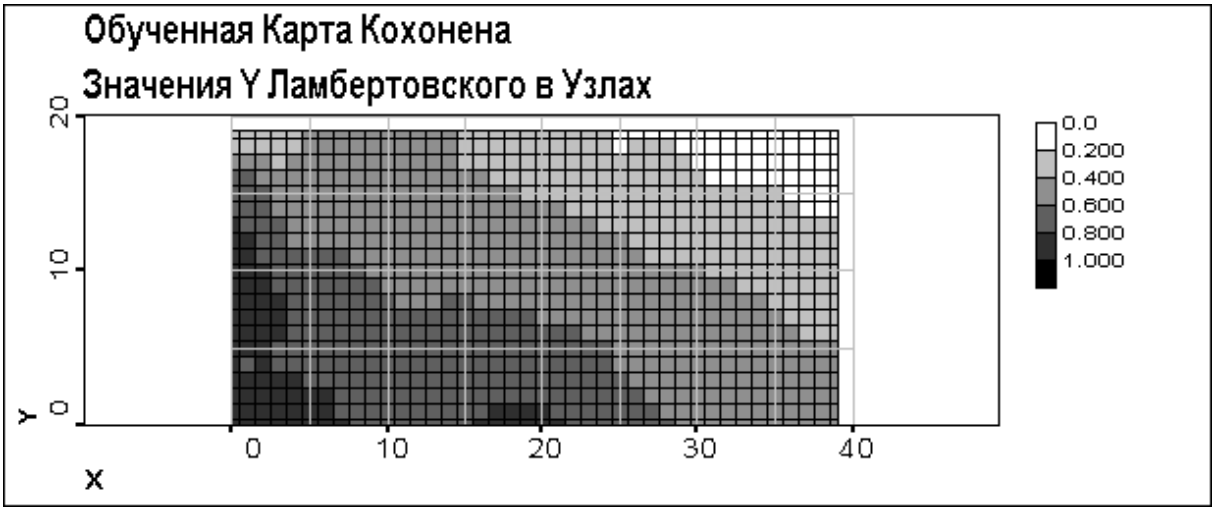

Рис. 3.1.2.4 Значения Y Ламбертовского в узлах обученной карты

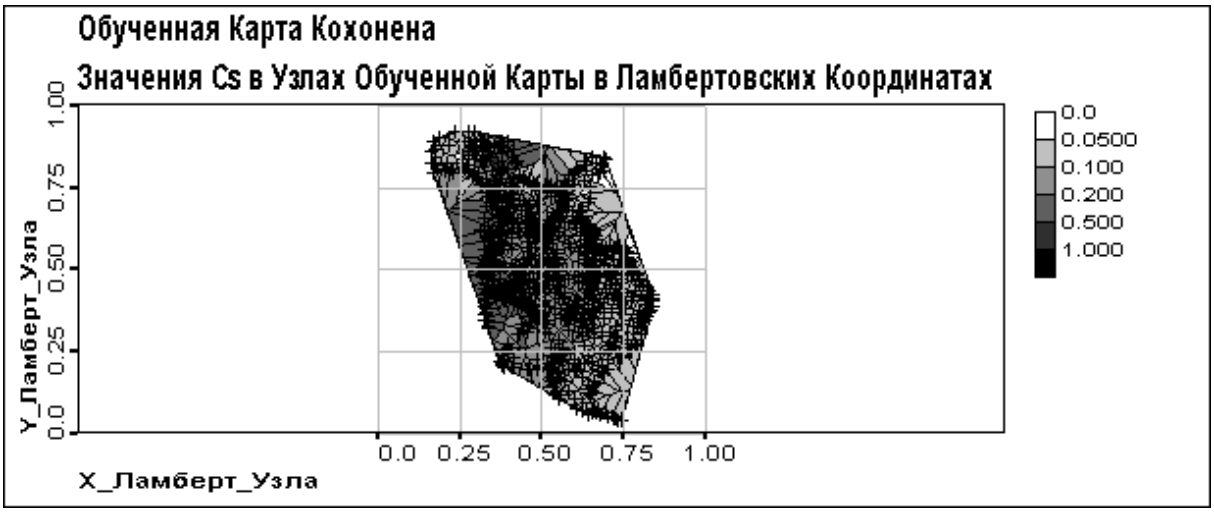

Рис. 4.1.2.5 Значения ^{137}Cs в узлах обученной карты в ламбертовских координатах

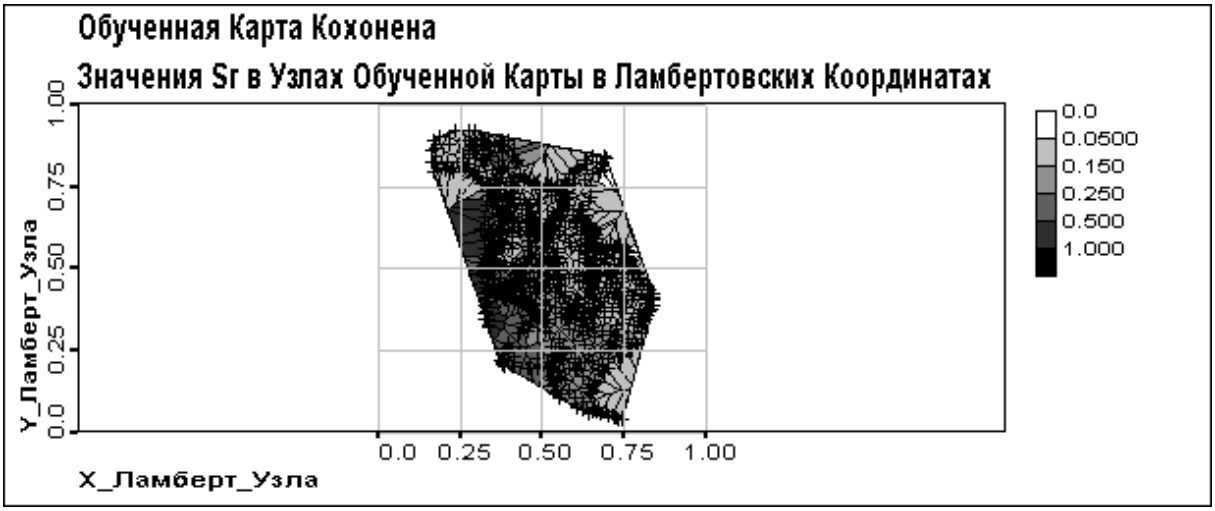

Рис. 4.1.2.6 Значения ^{90}Sr в узлах обученной карты в ламбертовских координатах

4.1.3 Обучение Самоорганизующихся Карт на некоррелированных Данных

В качестве эксперимента была проверена работа метода Самоорганизующихся Карт на случайных пространственно некореллированных данных (см. Рис. 4.1.3.1-4.1.3.2). Целью этого было продемонстрировать невозможность использования метода при отсутствии корреляции. Невозможность понижения размерности при отсутствии корреляции.

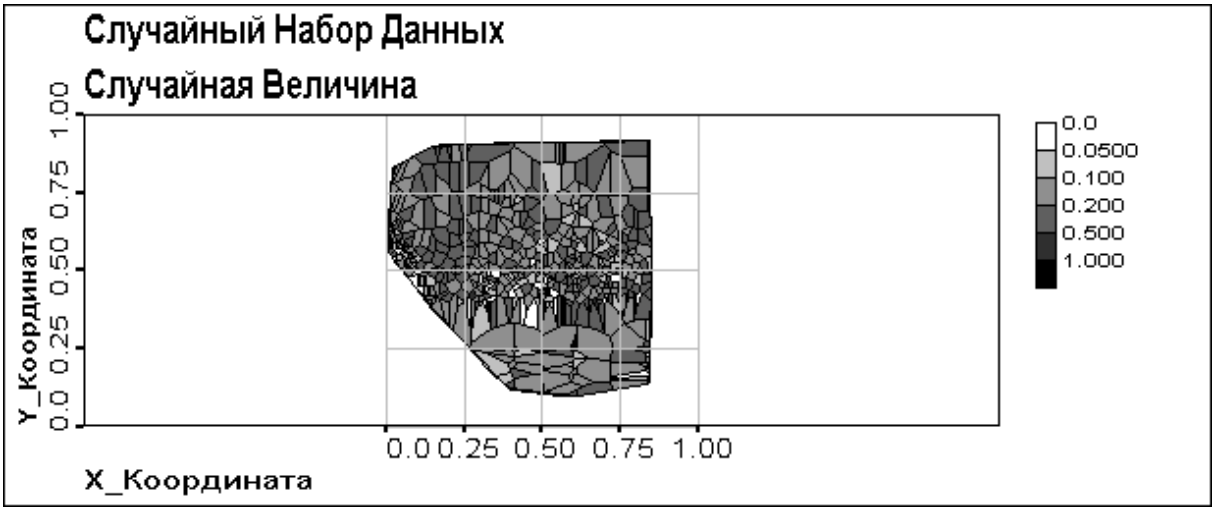

Рис. 4.1.3.1 Распределение случайной величины

Из Рис. 4.1.3.1 и 4.1.3.2 видно, что распределение данных представляет собой очень рябую картину, а вариограмма колеблется вокруг априорной вариации – случай, когда нет пространственной корреляции и данные распределены случайно – чистый наггет (pure nugget).

Из Рис.4.1.3.3-4.1.3.5 видно, что в карте Случайной Величины нет четкой упорядоченности, присутствуют множественные пятна неправильной формы, очень много классов ввиду отсутствия пространственной корреляции. Упорядоченность наблюдается только в координатах, то есть сеть мониторинга воспроизвелась.

Далее проводилась визуализация исходных данных по карте, обученной на них, представляющая собой, по сути, их классификацию (см. часть 2.2).

На Рис. 4.1.3.6 видно сильное усреднение (сравните с Рис. 4.1.3.1, особенно нижнюю часть) – алгоритм пытается как-то организовать пространственно некореллированные данные.

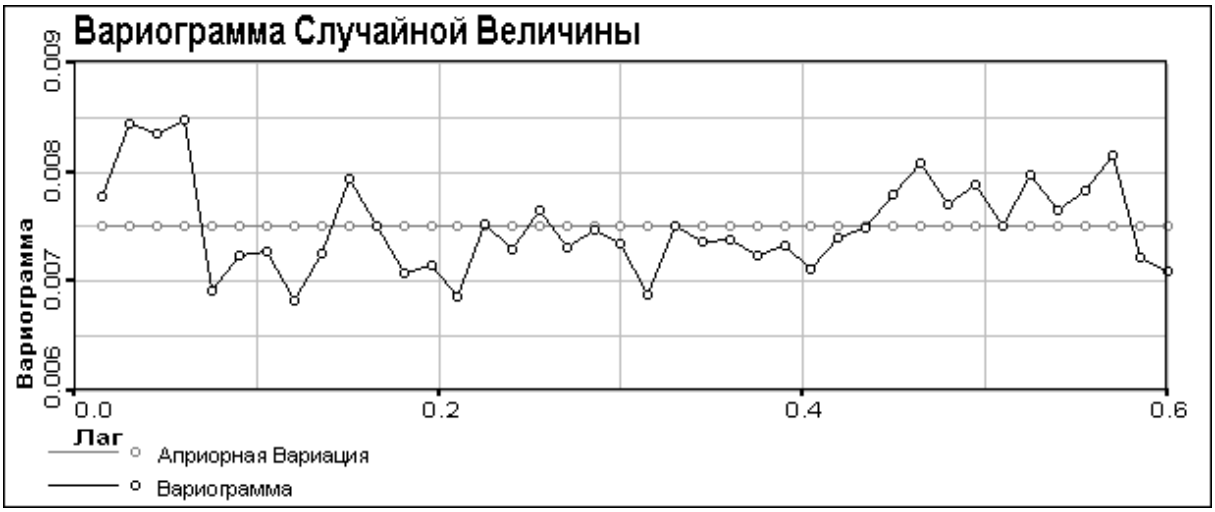

Рис. 4.1.3.2 Вариограмма случайной величины

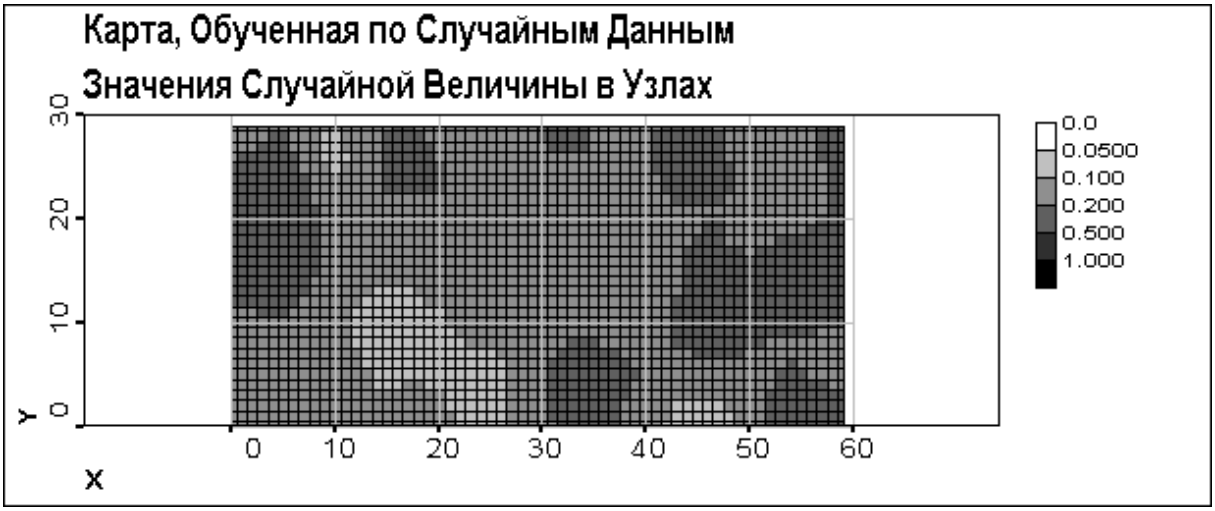

Рис. 4.1.3.3 Значения Случайной Величины в узлах карты, обученной по случайным данным

Рис. 4.1.3.4 Значения X_ Координаты в узлах карты, обученной по случайным данным

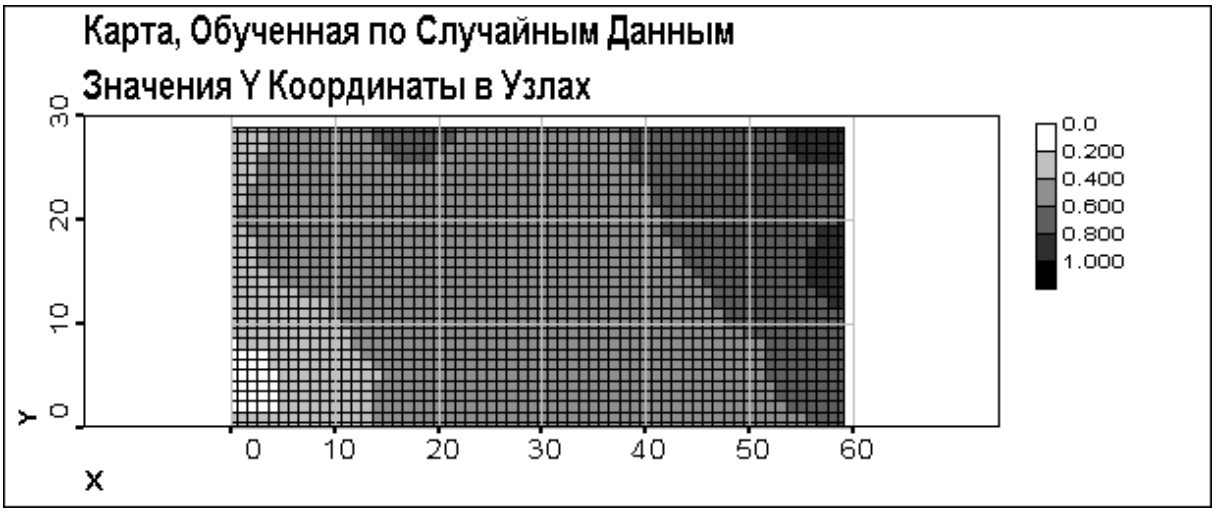

Рис. 4.1.3.5 Значения Y_ Координаты в узлах карты, обученной по случайным данным

Рис. 4.1.3.6 Классификация Случайной Величины по карте, обученной на случайных координатах

Рис. 4.1.3.7 Вариограмма Классификации Случайной Величины

Из Рис. 4.1.3.7 видно, что значения вариограммы уменьшились за счет усреднения и сглаживания данных с большими значениями, но чистый наггет (pure nugget) как был, так и остался.

Из полученных результатов можно сделать вывод, что исследуемый алгоритм бесполезно использовать при отсутствии корреляции.

4.2 Тестирование на Данных для Обучения

С целью проверки качества обучения проводилась классификация данных для обучения по обученной карте. При этом под оценками значений ^{137}Cs или ^{90}Sr понимались значения соответственно 3-ей и 4-ой компонент узла сети, к которому отнесен данный, подаваемый на вход вектор. Невязки вычислялись, как расхождения между оценкой, описанной выше и реальным значением. Для вычисленных невязок проводился статистический и корреляционный анализ (см. Части 2.2 и 4.1).

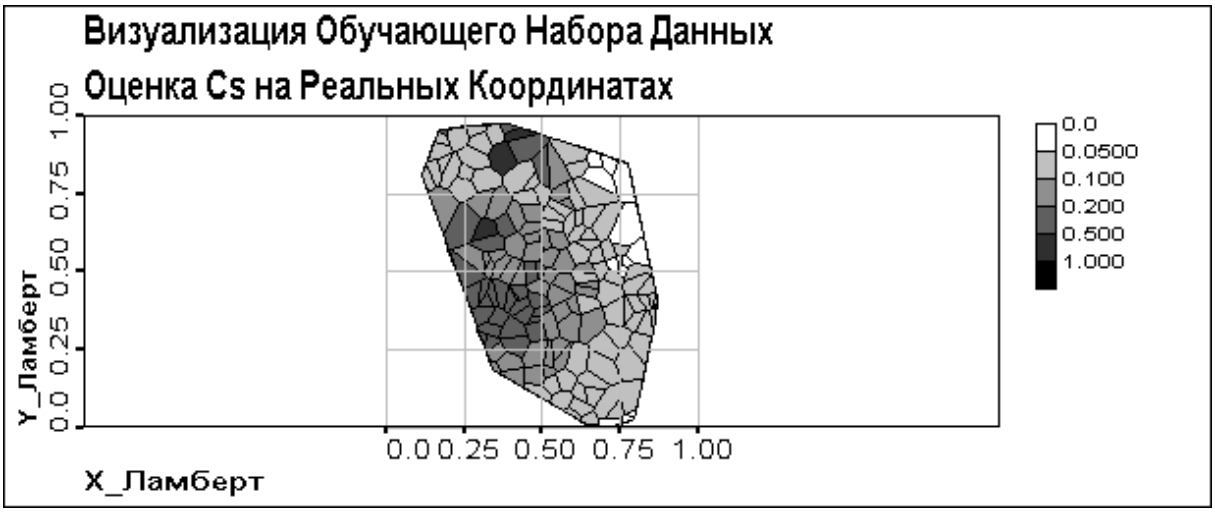

Рис. 4.2.1 Визуализация данных. Оценка ^{137}Cs на реальных координатах

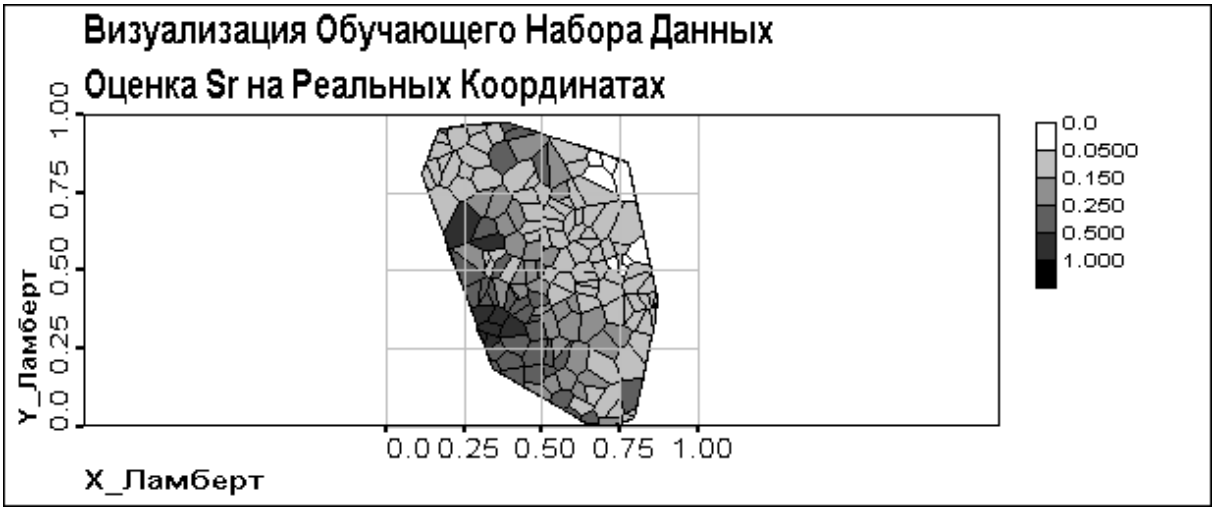

Рис. 4.2.2 Визуализация данных. Оценка ^{90}Sr на реальных координатах

Как можно увидеть, сравнивая Рис.4.2.1 и 4.2.2 с Рис. 3.1.3 и 3.1.4, в оценке сохраняется общая структура данных, но частично сглаживаются локальные минимумы и максимумы вследствие тяготения метода к усреднению значений.

Как видно из Рис. 4.2.3-4.2.4, невязки как по ^{137}Cs, так и по ^{90}Sr принимают сколько-нибудь отличное от 0 значение исключительно в местах сглаживаемых локальных экстремумов.

Рис. 4.2.3 Невязки по ^{137}Cs

Рис. 4.2.4 Невязки по ^{90}Sr

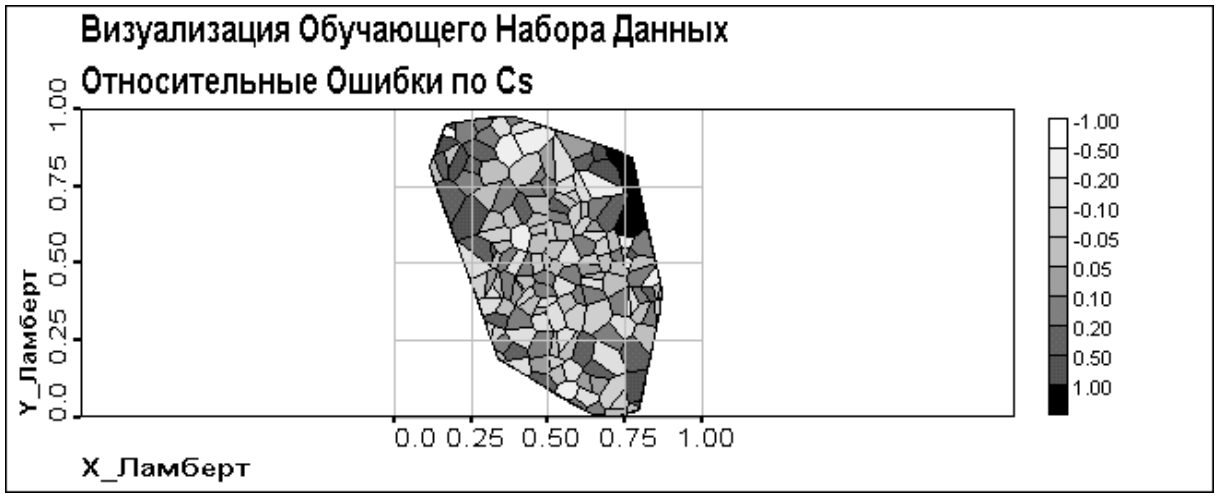

Рис. 4.2.5 Относительные ошибки по ^{137}Cs

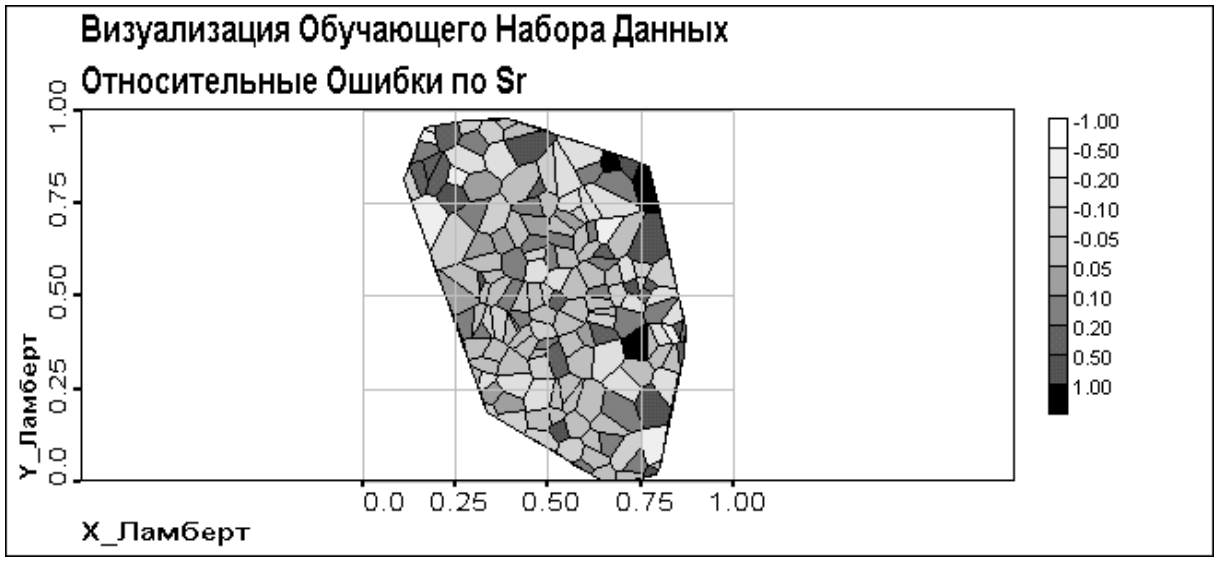

Рис. 4.2.6 Относительные ошибки по ^{90}Sr

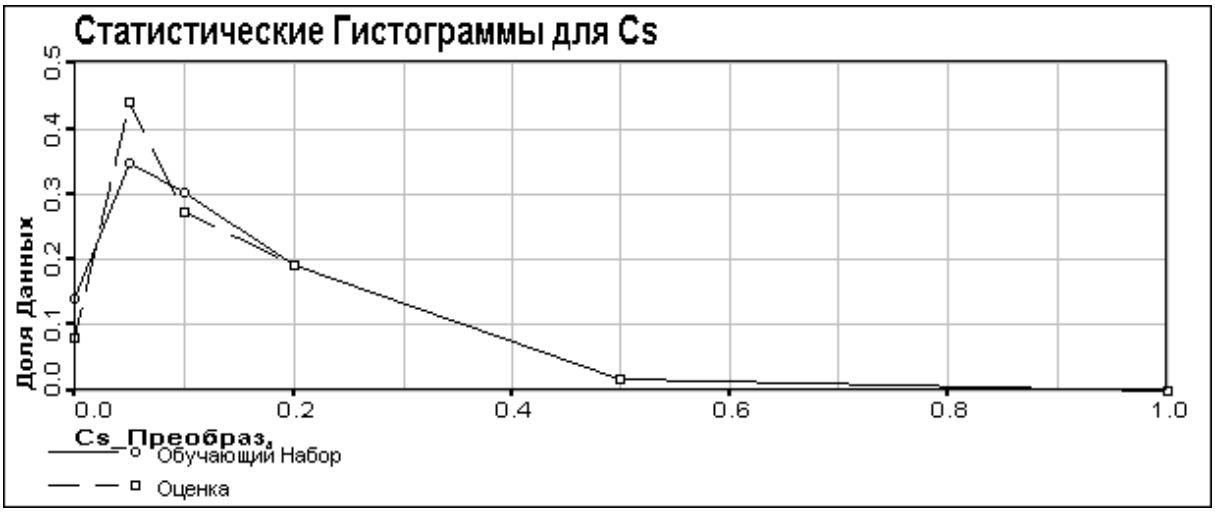

Рис. 4.2.7 Статистические гистограммы для ^{137}Cs

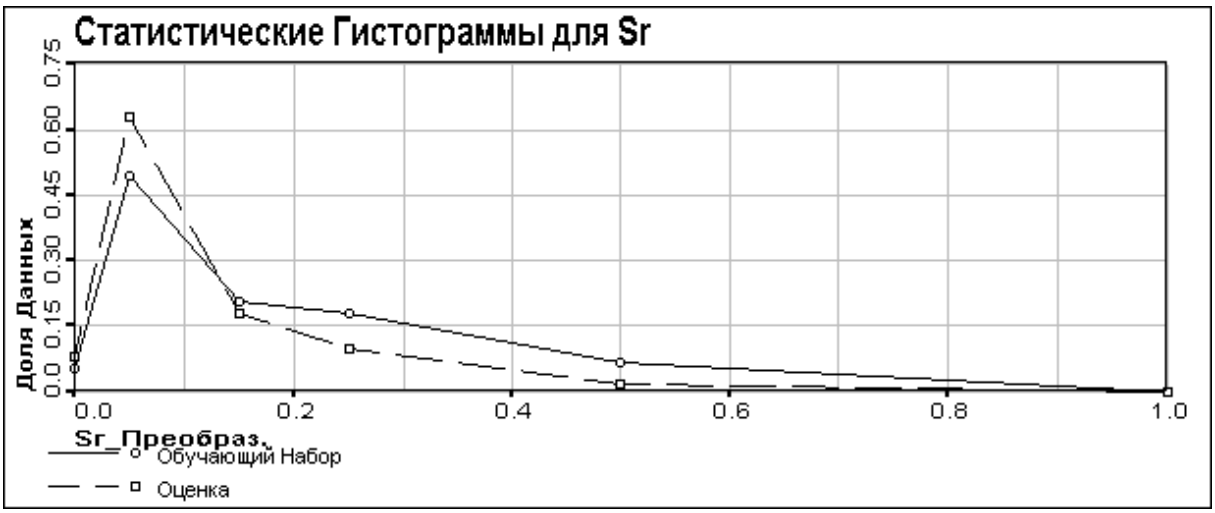

Рис. 4.2.8 Статистические гистограммы для ^{90}Sr

Как можно видеть из Рис. 4.2.5 и 4.2.6 относительная ошибка как по ^{137}Cs, так и по ^{90}Sr, становится значительной только в области малых значений этих компонент.

Из Рис. 4.2.7 и 4.2.8 мы видим, что распределение значений оценки хорошо совпадает со значениями исходных данных, за исключением того, что, вследствие усреднения, увеличивается доля данных со средними значениями и уменьшается с экстремальными.

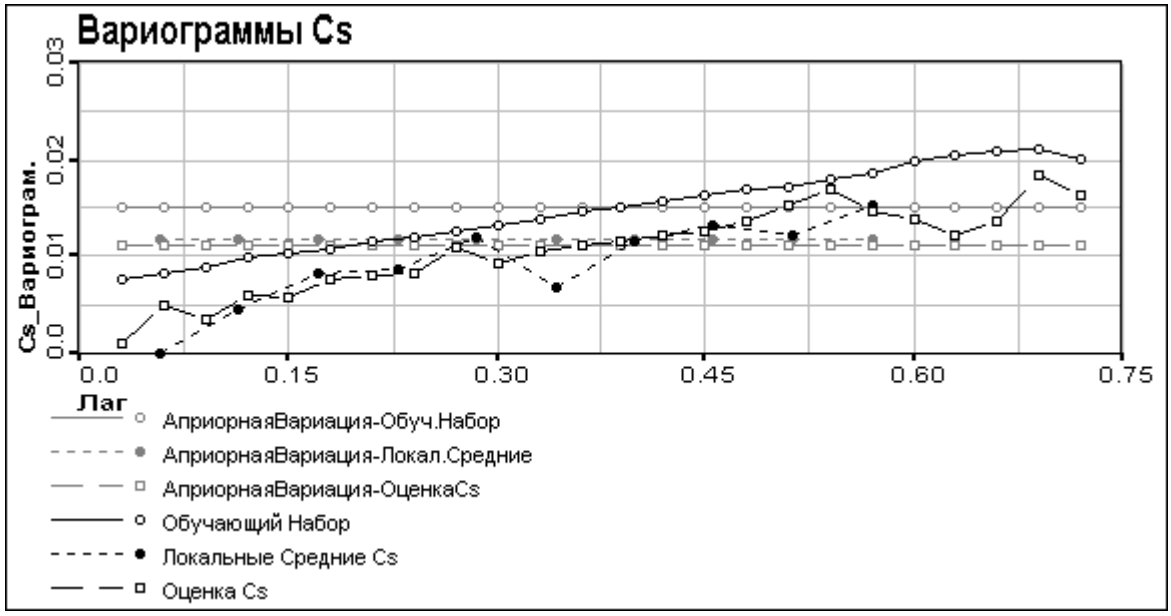

Рис. 4.2.9 Вариограммы ^{137}Cs: исходные данные, оценка и локальные средние

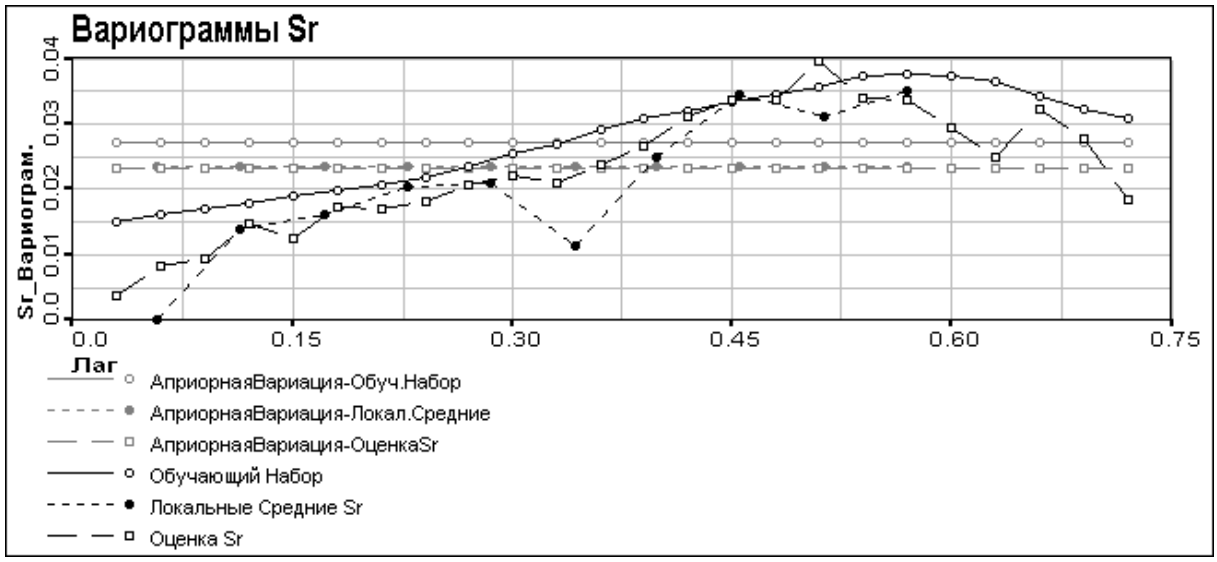

Рис. 4.2.10 Вариограммы ^{90}Sr: исходные данные, оценка и локальные средние

Из вариограмм, представленных на Рис. 4.2.9 и 4.2.10 видно, что, вследствие усреднения данных априорная вариация оценки совпадает с априорной вариацией локальных средних, как и ожидалось из теории. Также мы можем видеть, что значения вариограммы оценок по ^{137}Cs меньше, чем для обучающих данных, но эффективный радиус корреляции практически совпадает. По ^{90}Sr же мы можем наблюдать практически полное совпадение вариограмм.

Как видно из Рис. 4.2.11 и 4.2.12, невязки по ^{137}Cs и по ^{90}Sr коррелированны, особенно по ^{137}Cs, по ^{90}Sr в следствие того, что, при достаточном числе пар с малым лагом, с ростом лага значения падают, а не растут, можно предполагать отсутствие корреляции. Корреляции невязок возникают из-за того, что метод

проводит усреднение, следовательно выявляет только тренд данных. Для улучшения окончательной оценки при коррелированных невязках можно использовать методы геостатистики, например, кригинг.

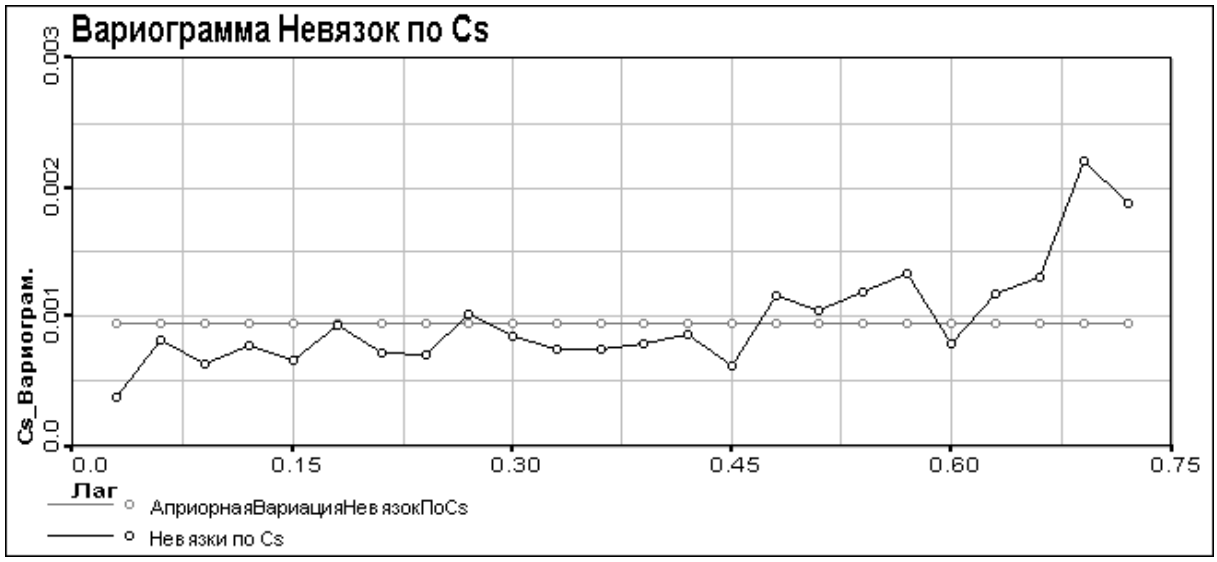

Рис. 4.2.11 Вариограмма невязок по ^{137}Cs

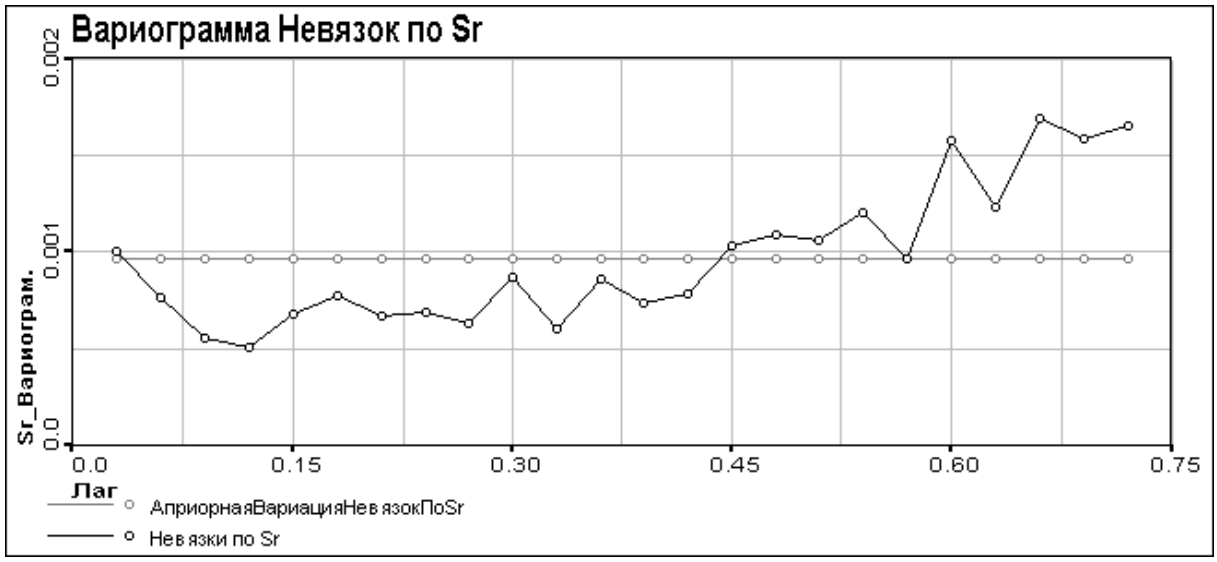

Рис. 4.2.12 Вариограмма невязок по ^{90}Sr

Из представленных в этой части результатов можно сделать вывод, что метод Самоорганизующихся Карт может достаточно хорошо воспроизводить данные, на которых учится. Следовательно, можно продолжить тестирование.

4.3 Использование Самоорганизующейся карты Кохонена на Тестовом Наборе Данных

4.3.1 Визуализация Тестового Набора Данных со Всеми Известными Полями

Следующим шагом тестирования карты Кохонена была визуализация тестового набора со всеми известными полями. Это делалось для проверки классификации на тестовом наборе и последующего сравнения с результатами визуализации с неполными данными. Как и в п. 4.2, оценками значения ^{137}Cs и

^{90}Sr в данной точке считались соответствующие значения узла, ближайшего к данному вектору. По полученным оценкам вычислялись невязки и строились вариограммы оценок и невязок.

Рис. 4.3.1.1 Визуализация Данных. Оценка ^{137}Cs на реальных координатах

Рис. 4.3.1.2 Визуализация Данных. Оценка ^{90}Sr на реальных координатах

Как видно на Рис. 4.3.1.1 и 4.3.1.2, расположение зон повышенной и пониженной концентрации в оценках ^{137}Cs и ^{90}Sr практически совпадает (особенно по ^{137}Cs) с исходными данными, т.е., как и в случае визуализации обучающего набора данных, , в оценке сохраняется общая структура данных, но частично сглаживаются локальные минимумы и максимумы вследствие стремления метода к усреднению значений.

Как видно из Рис. 4.3.1.3-4.3.1.6, несмотря на то, что значения невязок по ^{137}Cs и ^{90}Sr малы, значения относительных ошибок довольно велики, особенно по ^{90}Sr.

На Рис. 4.3.1.7 и 4.3.1.8 хорошо видно, что статистические гистограммы оценок ^{137}Cs и ^{90}Sr практически совпадают с гистограммами исходных данных (см. Рис. 3.1.3-Рис.3.1.6).

Из вариограмм, представленных на Рис. 4.3.1.9-4.3.1.10 можно видеть, что по ^{137}Cs наблюдается практически полное совпадение построенных вариограмм, а по ^{90}Sr при совпадении эффективного радиуса корреляции значения вариограммы оценок меньше, чем исходных данных, но в обоих случаях вариограммы оценок практически полностью совпадают с вариограммами локальных средних, что объясняется общим стремлением метода к усреднению.

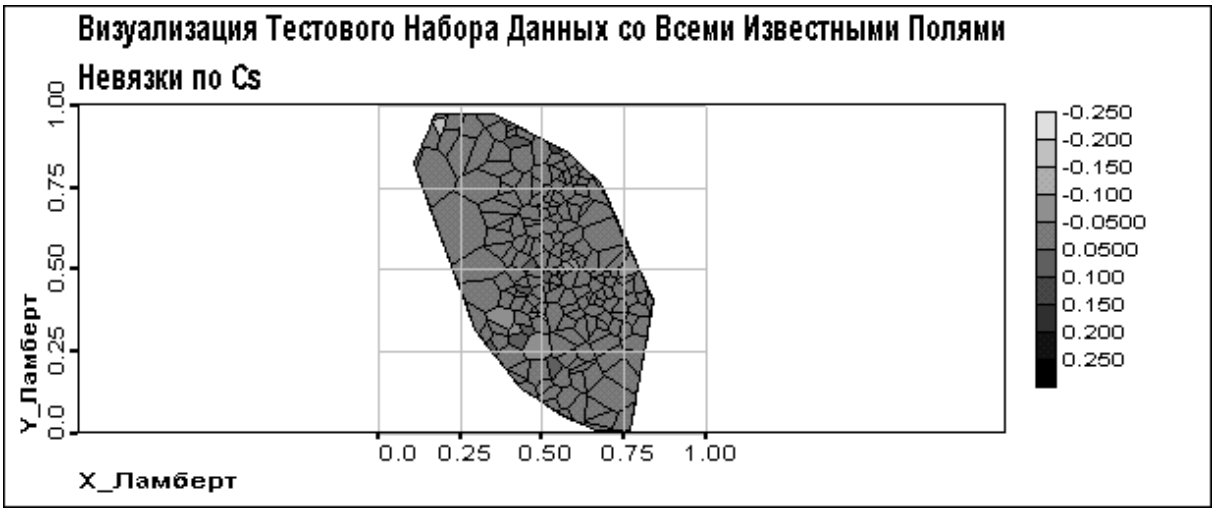

Рис. 4.3.1.3 Невязки по ^{137}Cs

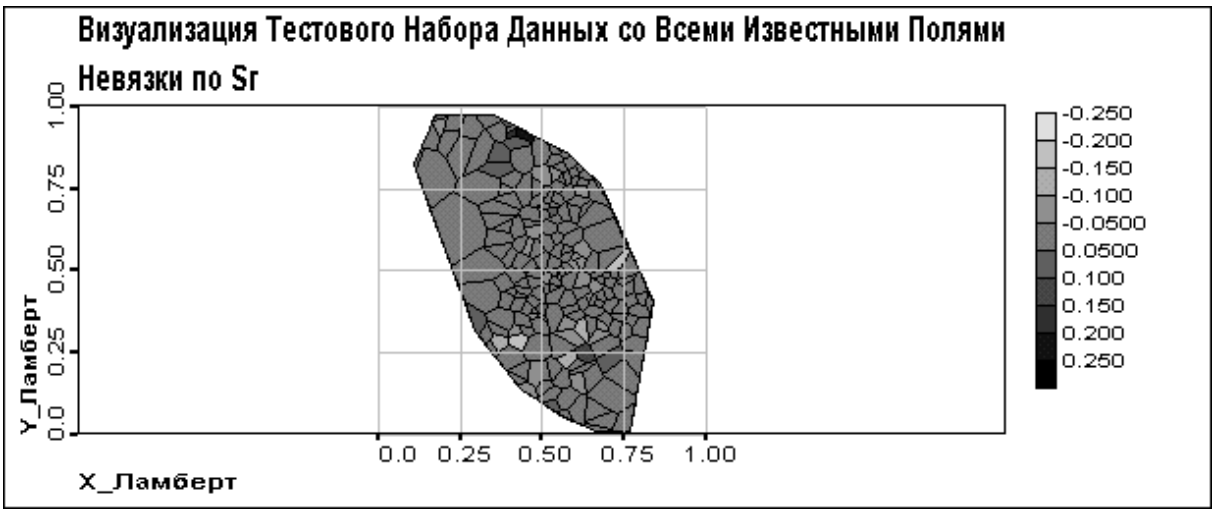

Рис. 4.3.1.4 Невязки по ^{90}Sr

Рис. 4.3.1.5 Относительные Ошибки по ^{137}Cs

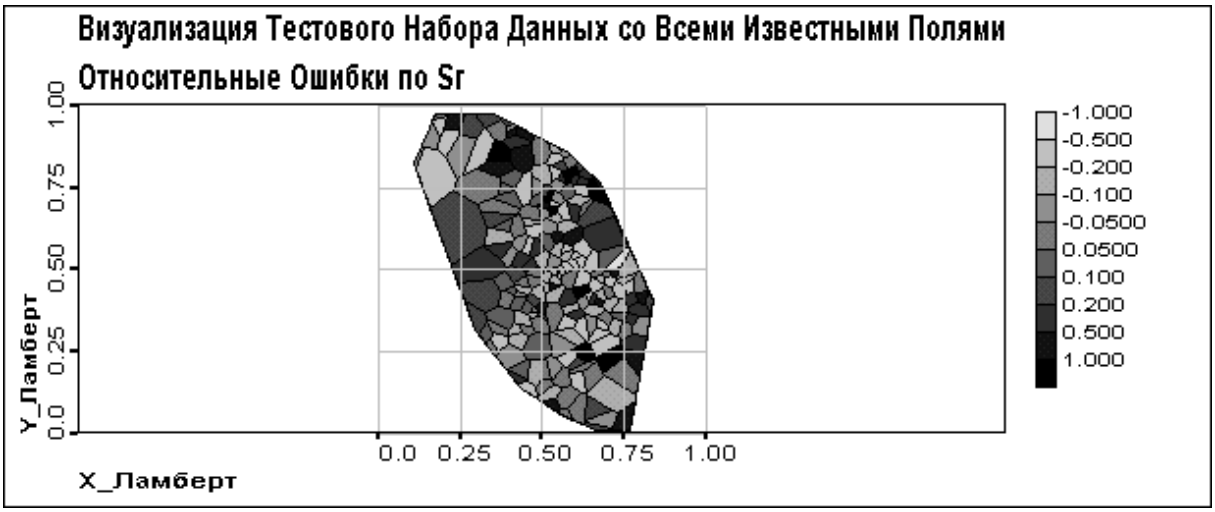

Рис. 4.3.1.6 Относительные Ошибки по ^{90}Sr

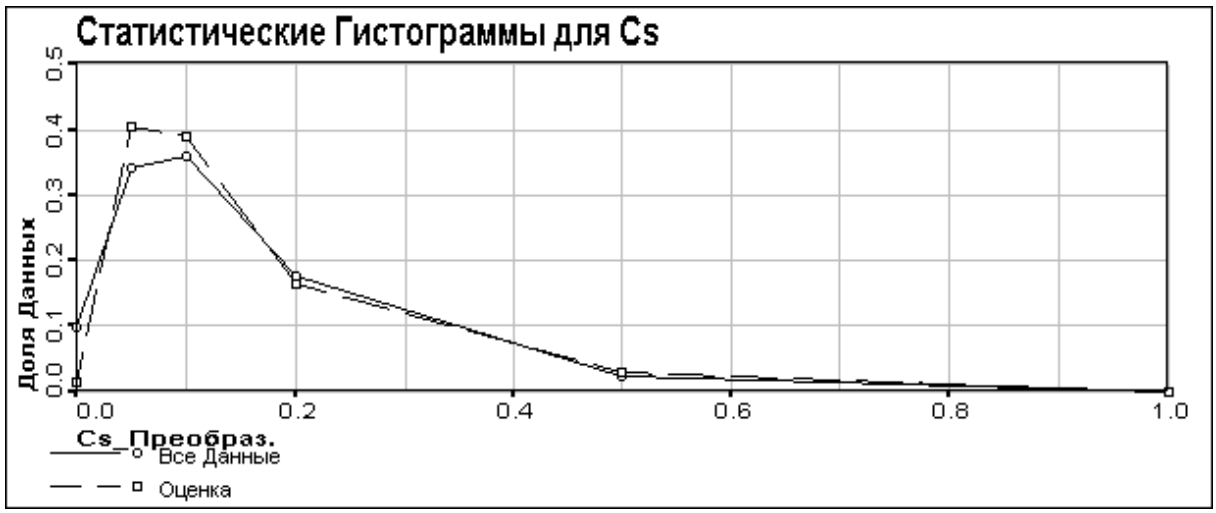

Рис. 4.3.1.7 Статистические гистограммы для ^{137}Cs

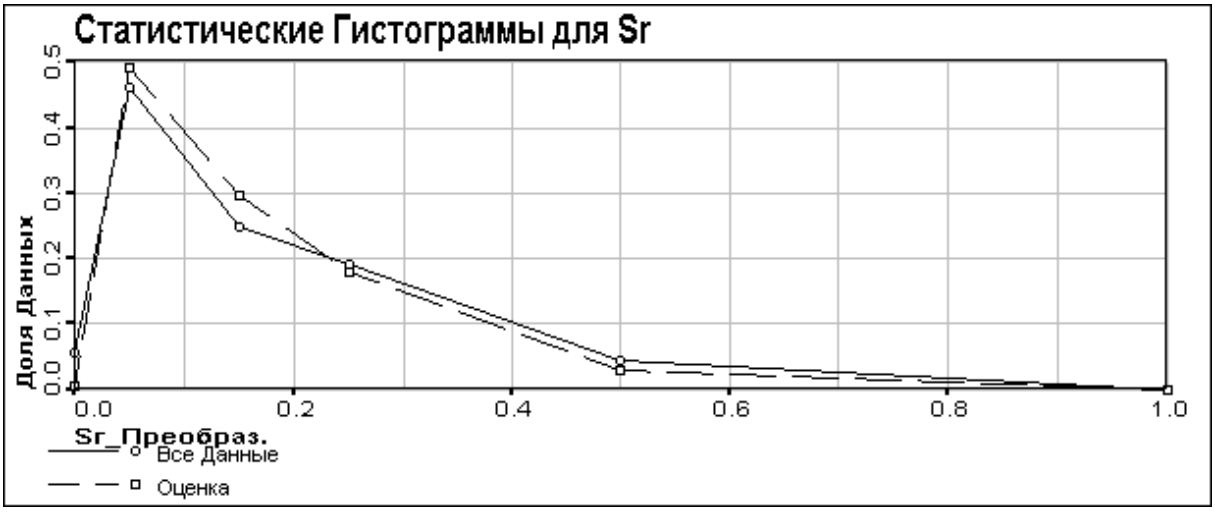

Рис. 4.3.1.8 Статистические гистограммы для ^{90}Sr

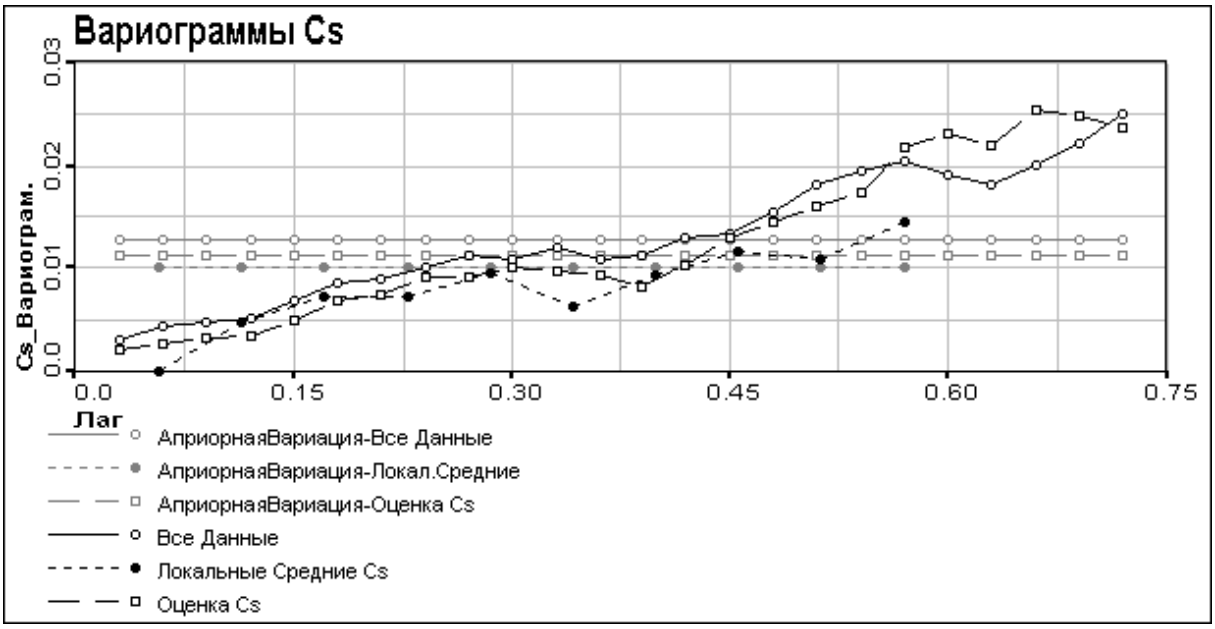

Рис. 4.3.1.9 Вариограммы ^{137}Cs: исходные данные, оценка и локальные средние

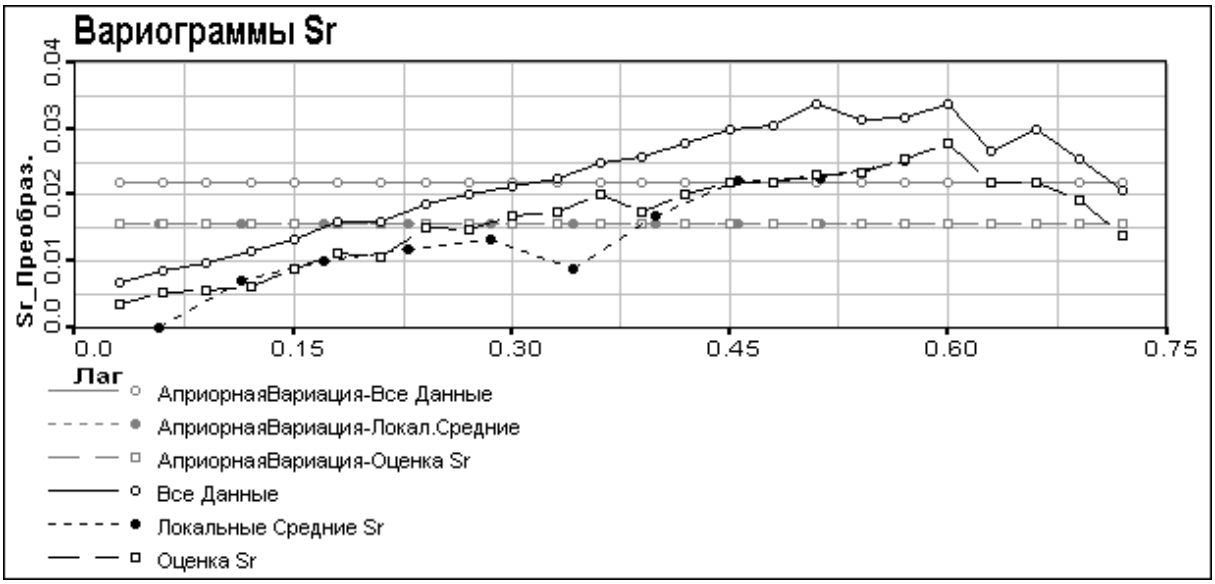

Рис. 4.3.1.10 Вариограммы ^{90}Sr: исходные данные, оценка и локальные средние

На Рис. 4.3.1.11-4.3.1.12 представлены вариограммы невязок, из которых видно, что, как и в случае визуализации обучающего набора данных, невязки пространственно коррелированны.

Из этой части можно сделать вывод, что обученная сеть успешно классифицирует тестовый набор.

4.3.2 Дополнение пропущенных данных ^{137}Cs в Тестовом Наборе Данных

После получения результатов визуализации тестового набора данных со всеми известными полями стало возможным перейти к следующему этапу работы – визуализации по обученной карте тестового набора данных с неполными данными. В этой части будет рассмотрен вариант, когда на вход карты подается набор данных с неизвестным ^{137}Cs – значение этого поля во входных данных было пропущено и эта компонента не участвует в вычислении расстояния от входного вектора до узлов сети с целью нахождения узла-победителя, т.е. растояние определяется только по 3-м компонентам входного вектора. По узлу-победителю оценивается значение ^{137}Cs в данной точке, полученные оценки анализируются с помощью гистограмм и вариограмм.

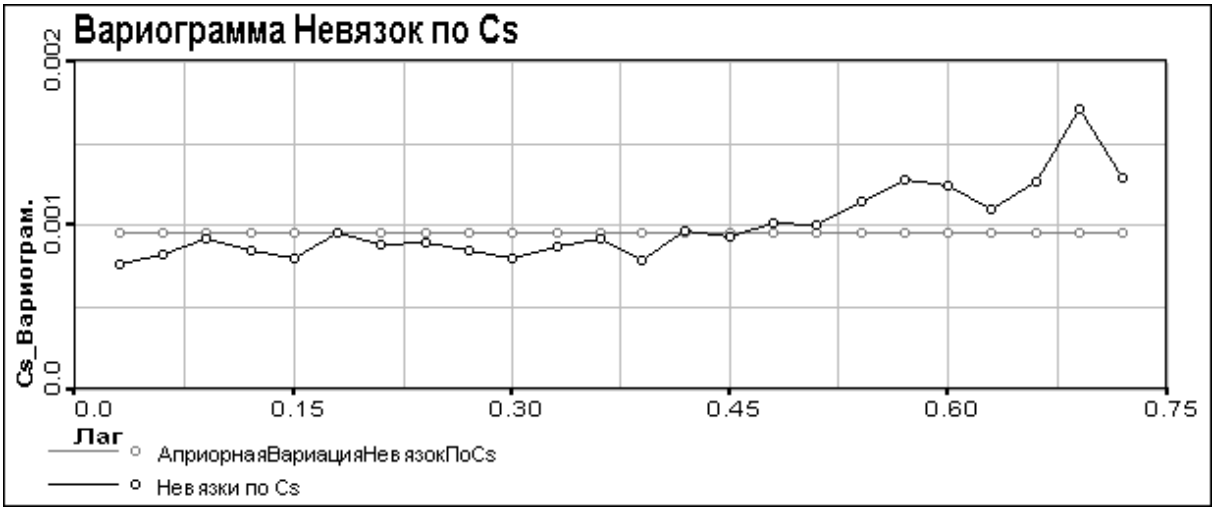

Рис. 4.2.11 Вариограмма невязок по ^{137}Cs

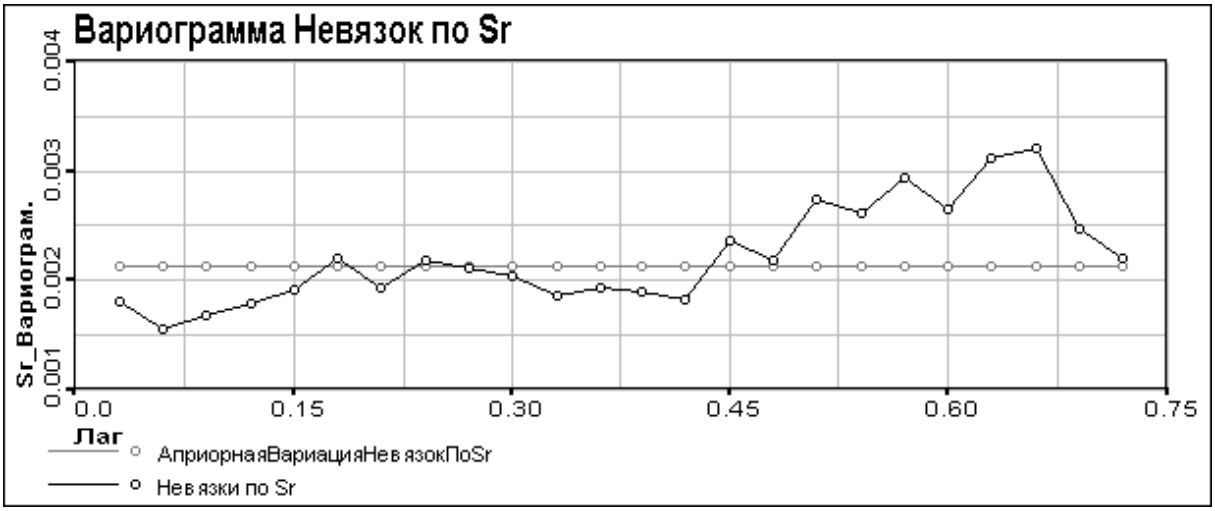

Рис. 4.2.12 Вариограмма невязок по ^{90}Sr

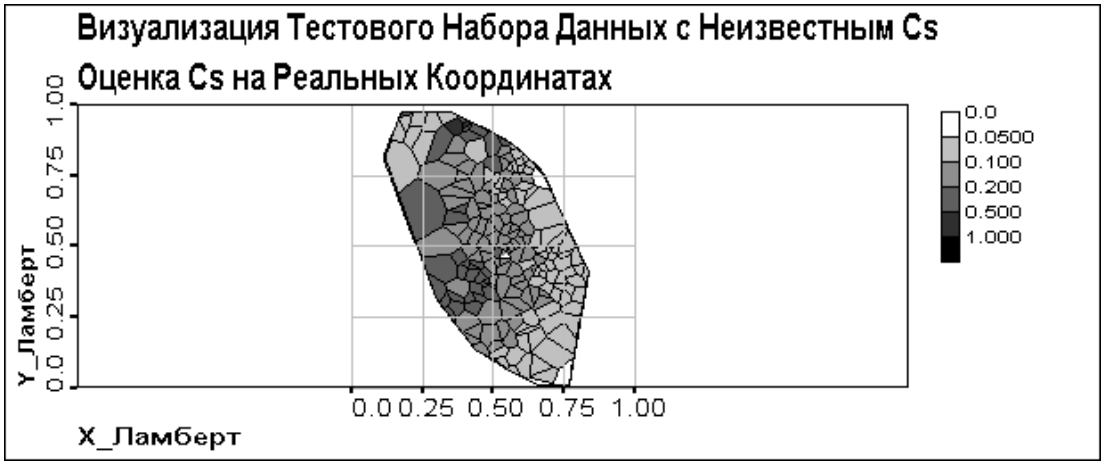

Рис. 4.3.2.1 Визуализация Данных. Оценка ^{137}Cs на реальных координатах

На Рис. 4.3.2.1 можно видеть полученные оценки значения ^{137}Cs, построенные в реальных координатах. Из них видно, что вследствие усреднения сеть не очень хорошо справляетя с оценками значений в точках локальных экстремумов. Значения невязок, представленные на Рис. 4.3.2.2 и 4.3.2.4 подтверждают это.

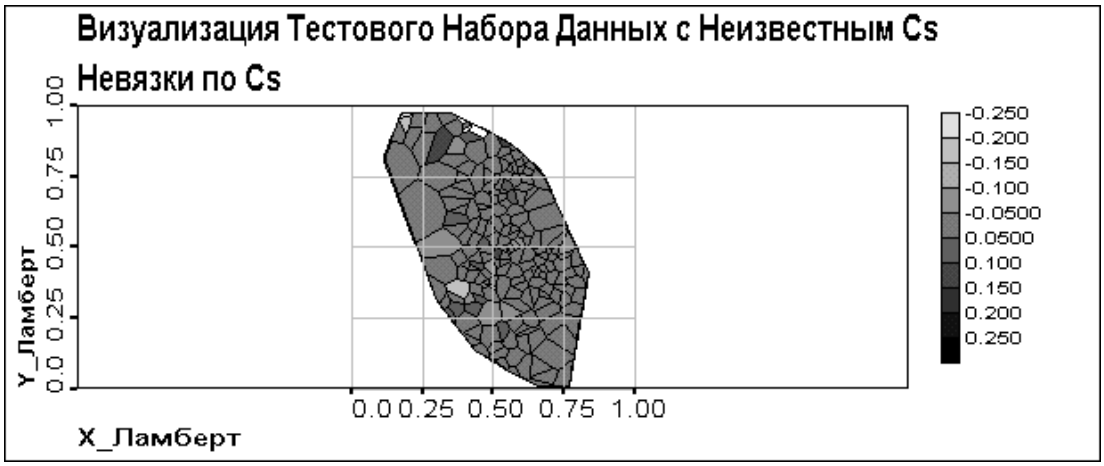

Рис. 4.3.2.2. Невязки по ^{137}Cs

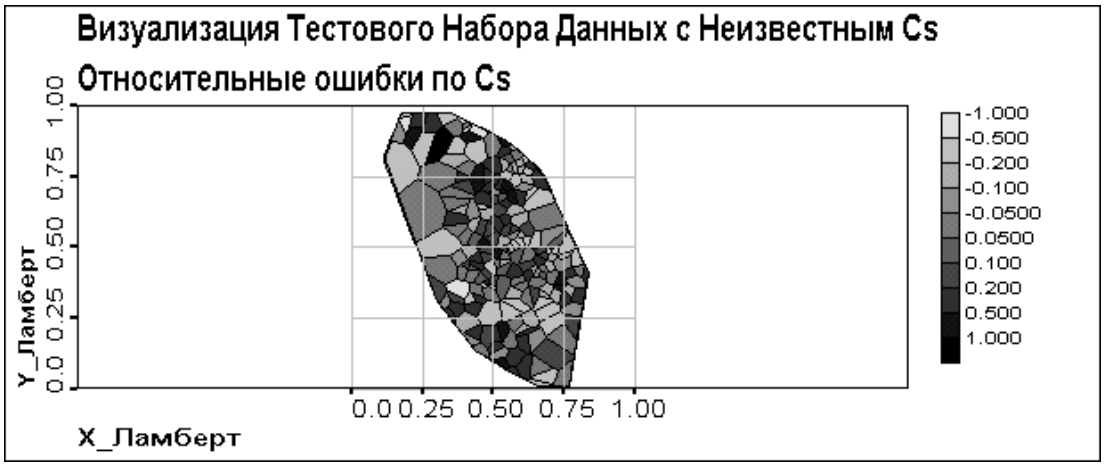

Рис. 4.3.2.3. Относительные Ошибки по ^{137}Cs

На Рис. 4.3.2.4 представлены статистические гистограммы результатов оценки и исходных данных. Как мы видим, гистограмма оценки ^{137}Cs практически не отклоняется от гистограммы исходных данных.

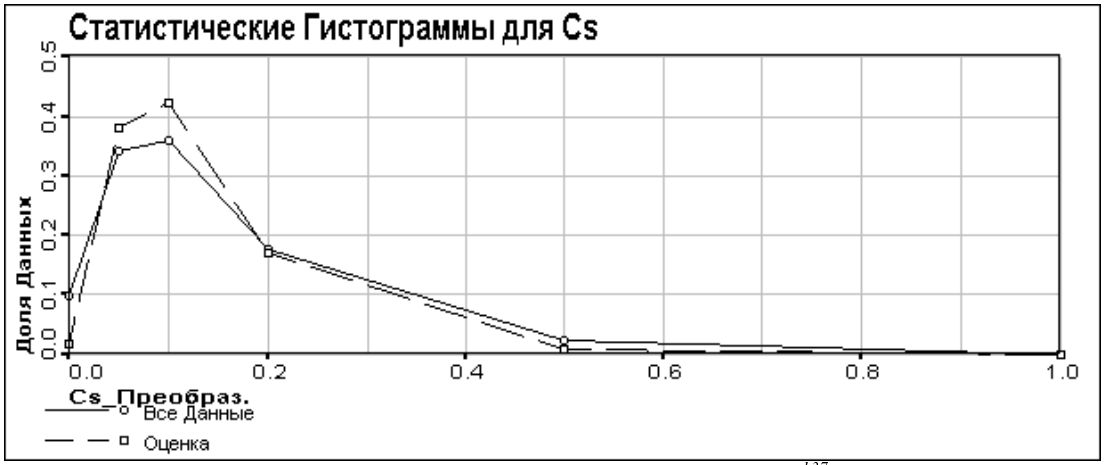

Рис. *4.3.2.4 Статистические гистограммы для ^{137}Cs*

На Рис. 4.3.2.5 можно видеть, что значения вариограммы полученных оценок меньше, чем значения вариограмм исходных данных и локальных средних, но структура вариограммы оценок воспроизводит структуру вариограмм данных.

На Рис. 4.3.2.6. мы можем видеть вариограмму полученных невязок и заключить из нее, что невязки в данном случае имеют пространственную корреляцию.

В итоге мы можем сказать, что обученная сеть достаточно хорошо оценивает исходные данные с отсутствием одной из компонент.

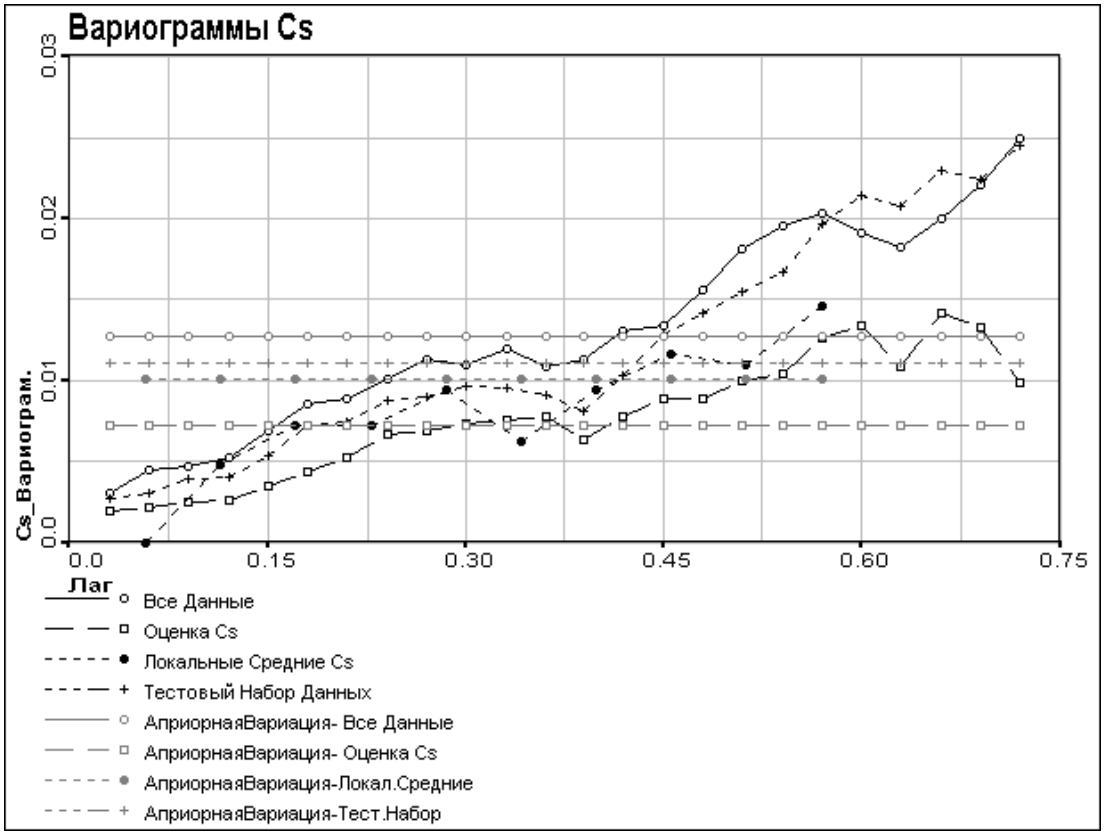

Рис. 4.3.2.5 Вариограммы ^{137}Cs: исходные данные, оценка и локальные средние

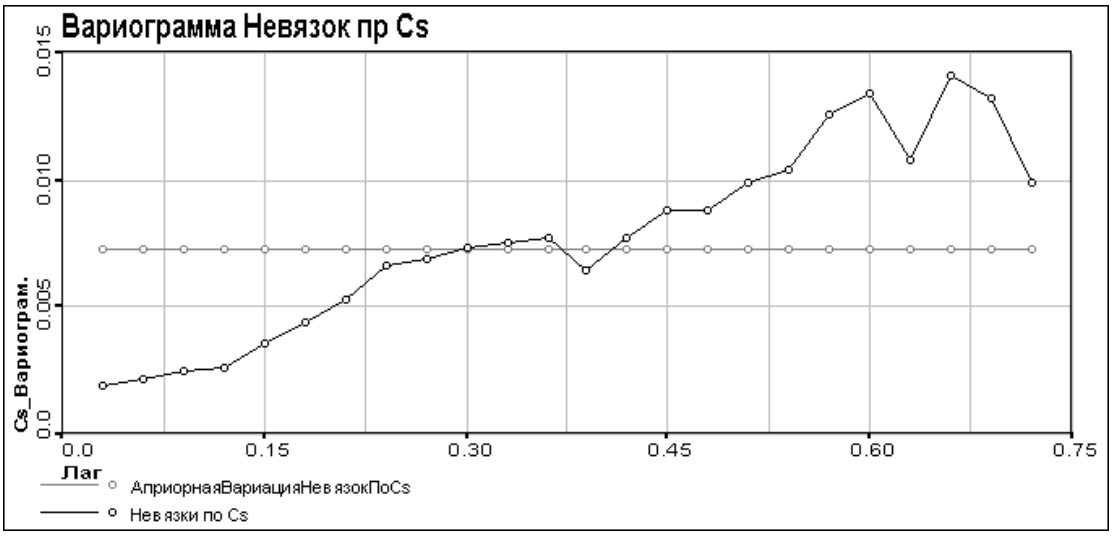

Рис 4.3.2.6 Вариограмма невязок по ^{137}Cs

4.3.3 Дополнение пропущенных значений ^{90}Sr в Тестовом Наборе Данных

Для сравнения с результатами визуализации по обученной карте тестового набора данных с неизвестным ^{137}Cs проводилась визуализация тестового набора с неизвестным ^{90}Sr (Алгоритм работы см. в частях 4.1 и 4.3.1) Результаты приведены на Рис. 4.3.3.1-4.3.3.3 Затем полученные оценки анализировались при помощи гистограмм и вариограмм (см. Рис. 4.3.3.4-4.3.3.6).

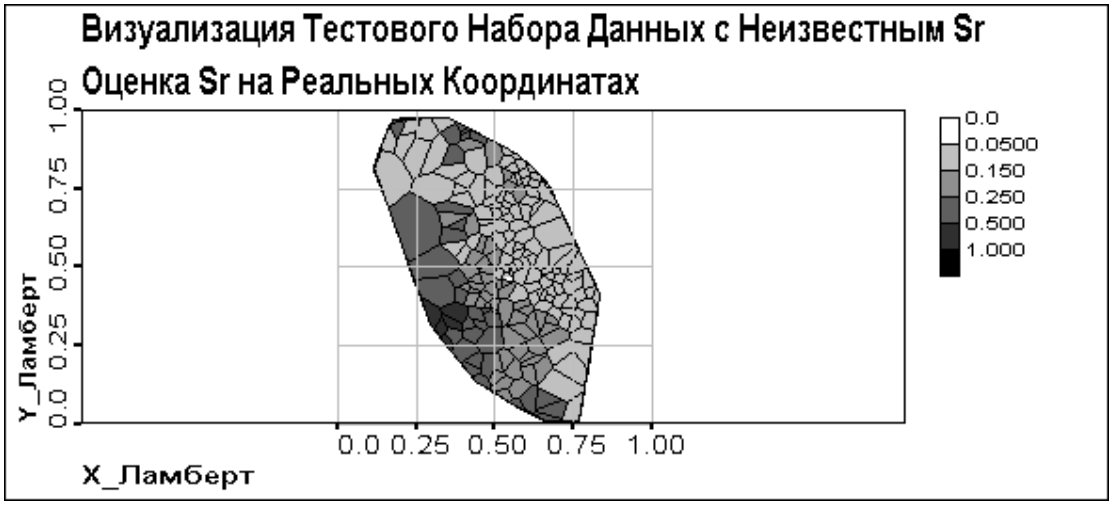

Рис. 4.3.3.1 Визуализация Данных. Оценка ^{90}Sr на реальных координатах

На Рис. 4.3.3.1 можно видеть полученные оценки значения ^{90}Sr, построенные в реальных координатах. Из них видно, что, как и в случае с визуализацией данных без ^{137}Cs, сеть, вследствие усреднения, не очень хорошо справляетя с оценками значений в точках локальных экстремумов. Значения невязок, представленные на Рис. 4.3.3.2 и 4.3.3.3 подтверждают это.

Рис. 4.3.3.2 Невязки по ^{90}Sr

Как можно видеть из Рис. 4.3.3.3, значения относительных ошибок в данном случае достигают еще более значительных значений, чем в предыдущем.

На Рис. 4.3.3.4 представлены статистические гистограммы результатов оценки и исходных данных. Как мы видим, гистограмма оценки ^{90}Sr практически не отклоняется от гистограммы исходных данных. Как и в остальных случаях, алгоритм слегка увеличивает долю данных со средними значениями и, соответственно, уменьшает с экстремальными.

На Рис. 4.3.2.5 мы можем видеть вариограмму полученных невязок и заключить из нее, что невязки в данном случае имеют явную пространственную корреляцию.

На Рис. 4.3.2.6 можно видеть, что вариограмма полученных оценок хорошо соответствует значениям вариограмм исходных данных и локальных средних.

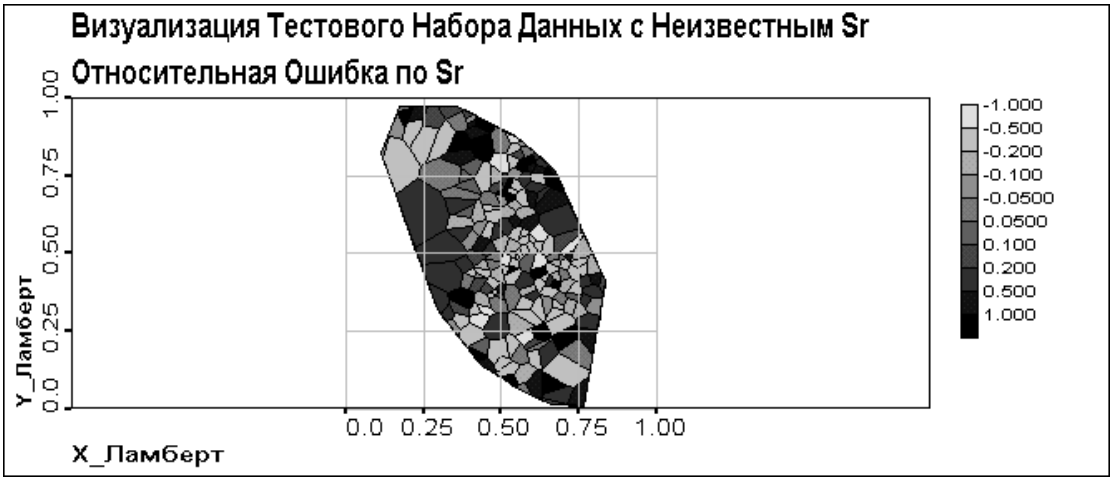

Рис. 4.3.3.3 Относительные Ошибки по ^{90}Sr

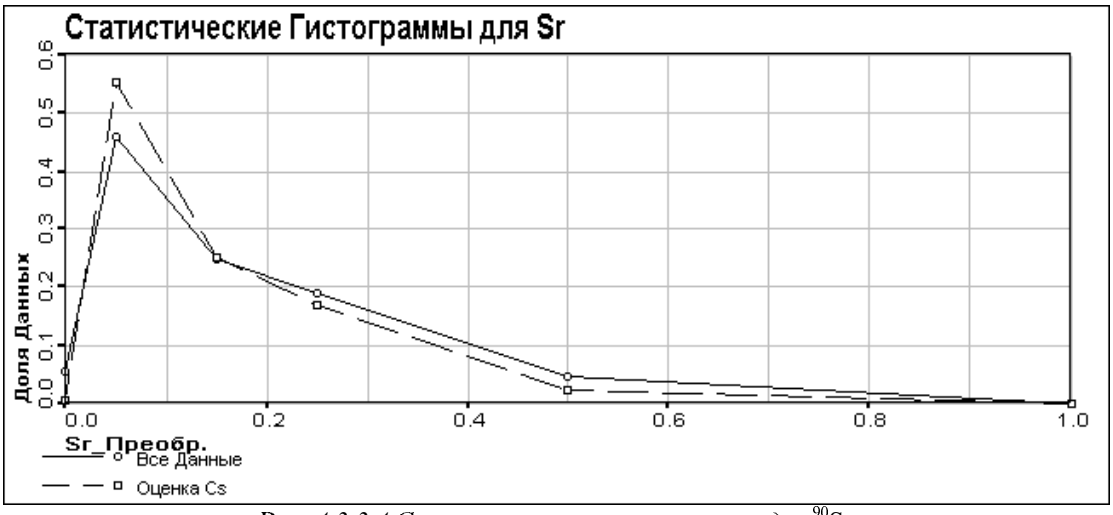

Рис. *4.3.3.4 Статистические гистограммы для* ^{90}Sr

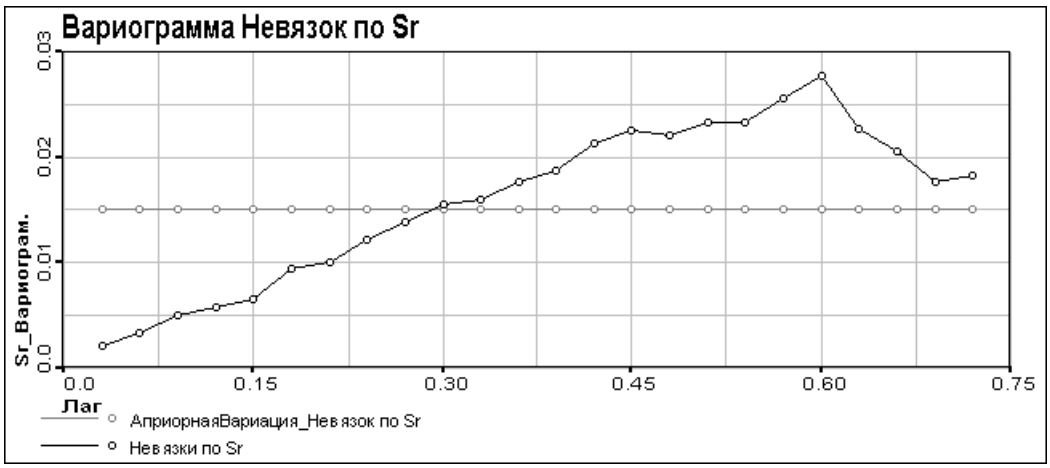

Рис. 4.3.3.5 Вариограмма невязок по ^{90}Sr

В заключение этой части можно сказать, что большие, чем в предыдущем случае, расхождения результатов вызваны большим расхождением в исходных данных между обучающим и тестовым наборами по ^{137}Cs, чем по ^{90}Sr.

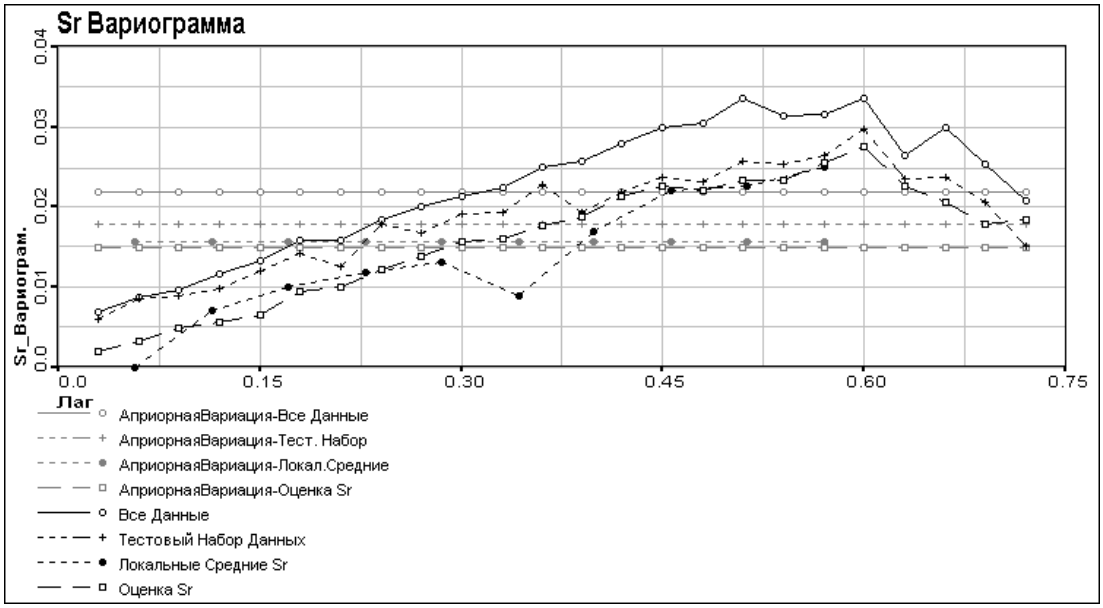

Рис. 4.3.3.6 Вариограммы ^{90}Sr: исходные данные, оценка и локальные средние

4.3.4 Дополнение Неизвестных ^{137}Cs и ^{90}Sr в Тестовом Наборе Данных

Заключительным этапом работы была попытка совместного восстановления пропущенных ^{137}Cs и ^{90}Sr по обученной ранее карте. Это проводилось на тестовом наборе данных. Результаты приведены на Рис. 4.3.4.1-4.3.4.6 Полученные оценки анализировались на воспроизведение гистограмм и вариограмм исходных данных (см. Рис. 4.3.4.7-4.3.4.12).

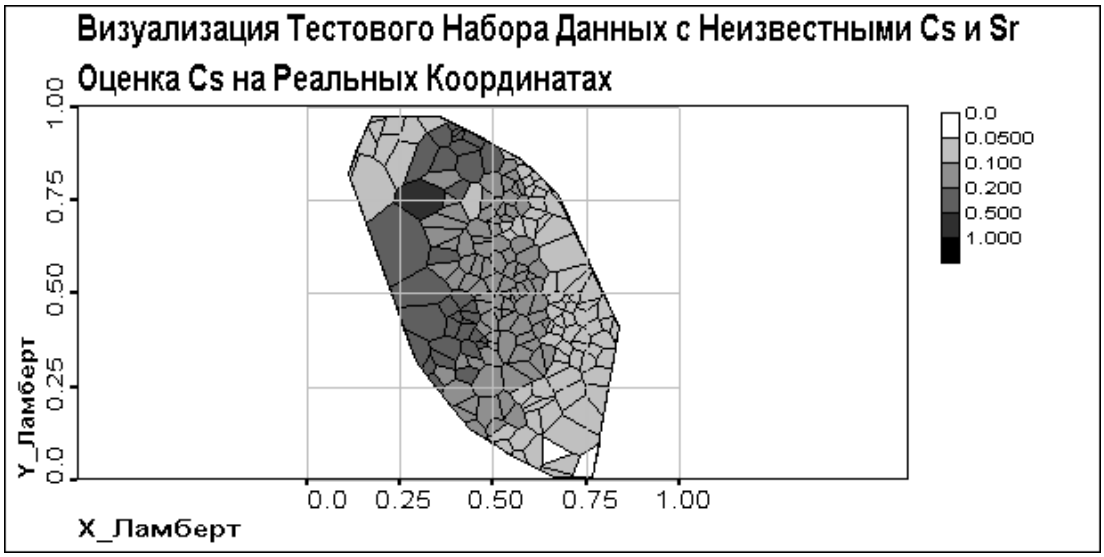

Рис. 4.3.4.1 Визуализация Данных. Оценка ^{137}Cs на реальных координатах

Как видно из Рис. 4.3.4.1 и 4.3.4.2, оценки пропущенных данных достаточно хорошо воспроизводят тренд исходных данных, но, как и в предыдущих случаях, сглаживает экстремумы. В данном случае, вследствие того, что здесь производится оценка только по координатам, это сглаживание более заметно, чем в предыдущих случаях.

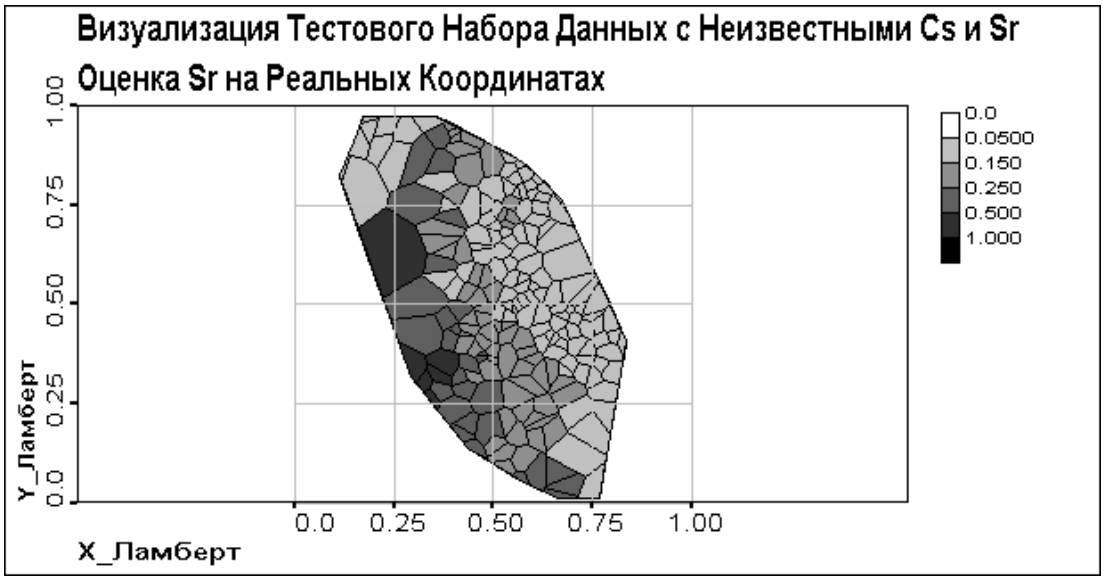

Рис. 4.3.4.2 Визуализация Данных. Оценка ^{90}Sr на реальных координатах

Рис. 4.3.4.3 Невязки по ^{137}Cs

Рис. 4.3.4.4 Невязки по ^{90}Sr

Из Рис. 4.3.4.3-4.3.4.6 видно, что в случае совместного восстановления ^{137}Cs и ^{90}Sr невязки и относительные ошибки достигают больших значений по сравнению с предыдущими случаями, когда восстанавливались только ^{137}Cs или только ^{90}Sr. Хорошо видно, что невязки и относительные ошибки оценок ^{90}Sr больше, чем оценок ^{137}Cs. Это может быть обусловлено большим расхождением между обучающим и тестовым наборами данных по ^{90}Sr.

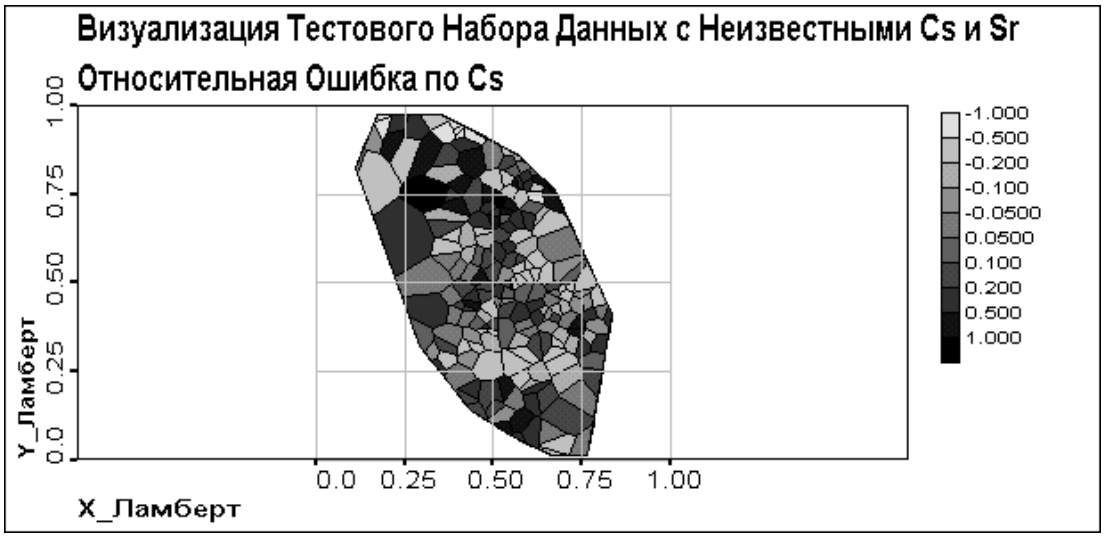

Рис. 4.3.4.5 Относительные Ошибки по ^{137}Cs

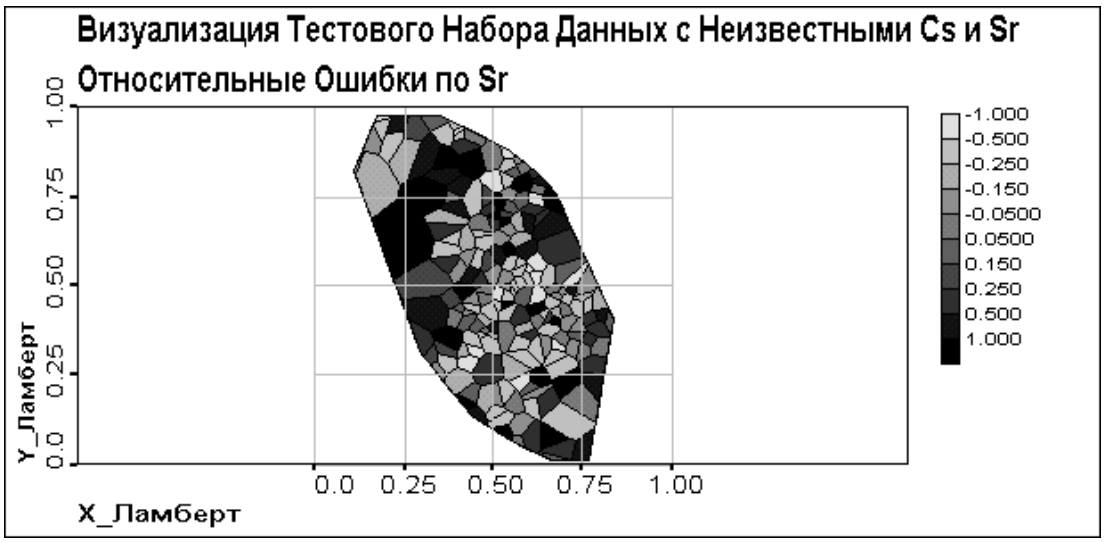

Рис. 4.3.4.6 Относительные Ошибки по ^{90}Sr

Статистические гистограммы оценок ^{137}Cs и ^{90}Sr, представленные на Рис. 4.3.4.7 и 4.3.4.8, демонстрируют то же, что и в предыдущих случаях – снижение доли экстремальных и увеличение доли средних значений.

Из представленных на Рис. 4.3.4.9 вариограмм для ^{137}Cs видно, что вариограмма оценок ^{137}Cs имеет явно меньшие значения, чем вариограммы исходных данных и локальных средних, но эффективный радиус корреляции и форма сохраняются.

На Рис. 4.3.4.10 можно видеть вариограммы для ^{90}Sr. Из них видно, что вариограммы оценок ^{90}Sr достаточно хорошо соответствуют вариограммам исходных данных.

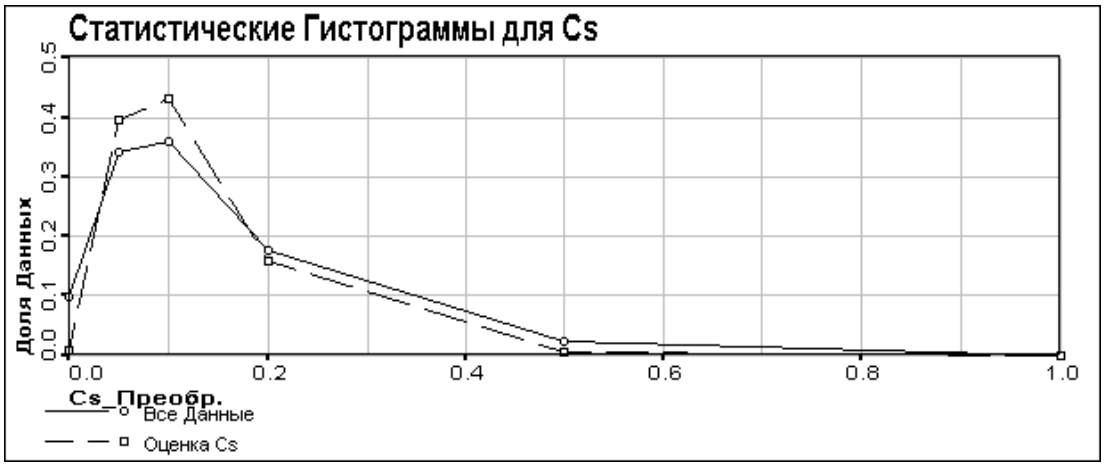

Рис. 4.3.4.7 Статистические гистограммы для ^{137}Cs

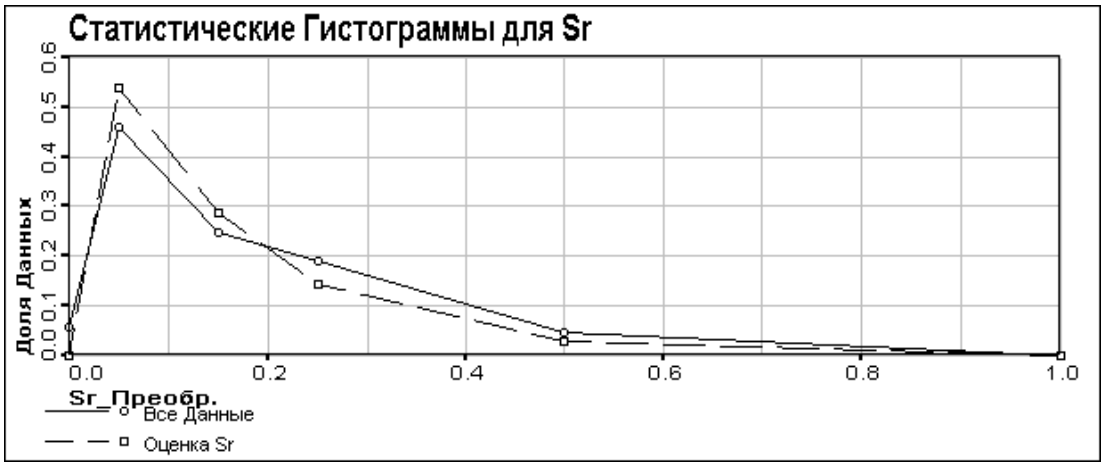

Рис. 4.3.4.8 Статистические гистограммы для ^{90}Sr

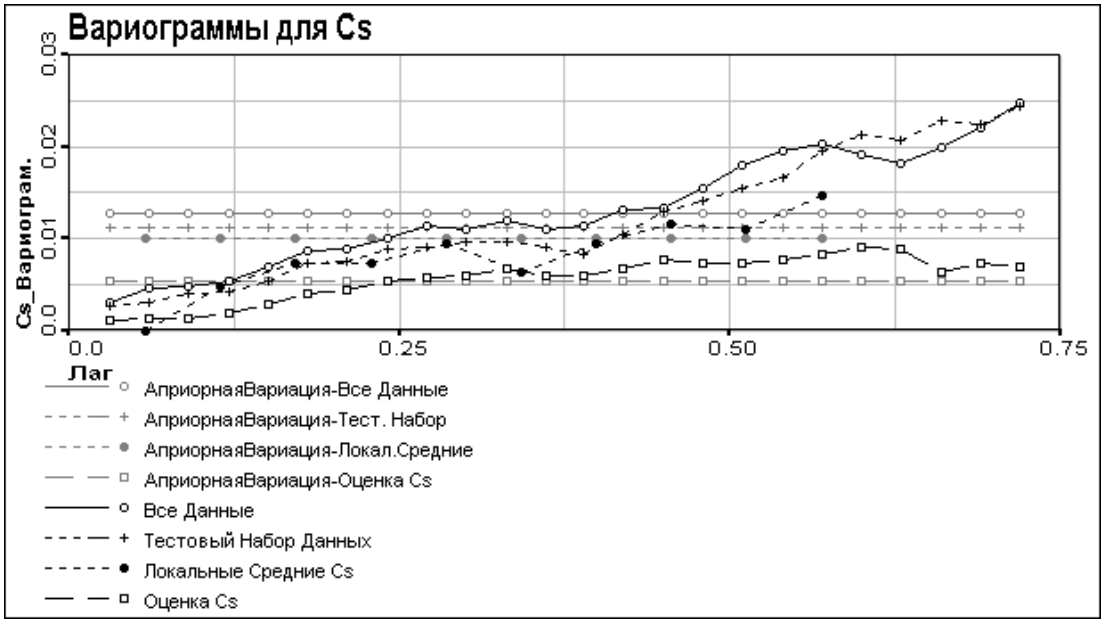

Рис. 4.3.4.9 Вариограммы ^{137}Cs: исходные данные, оценка и локальные средние

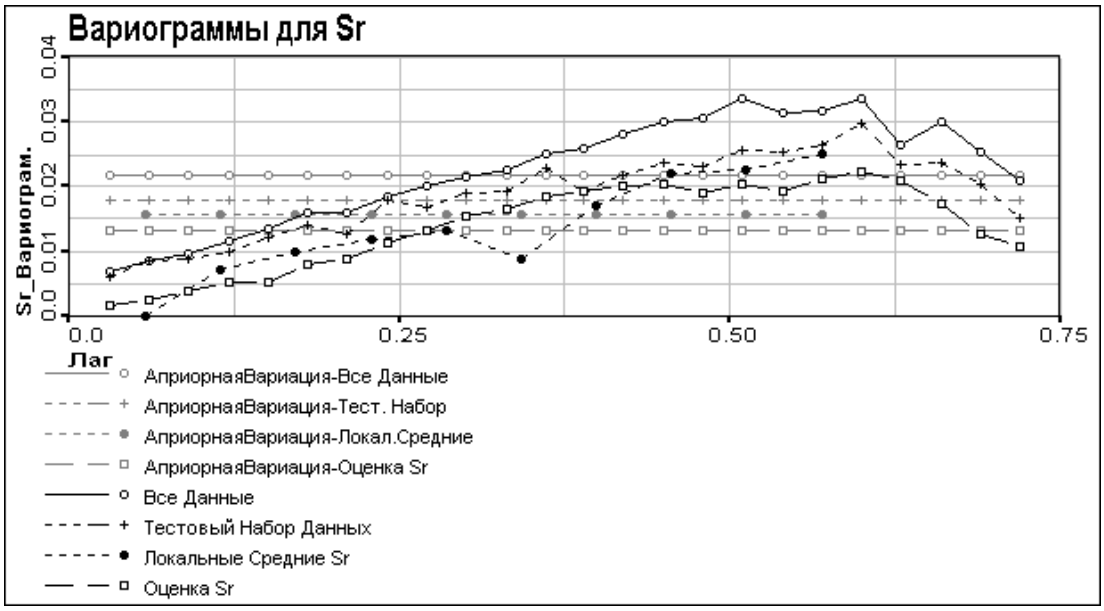

Рис. 4.3.4.10 Вариограммы ^{90}Sr: исходные данные, оценка и локальные средние

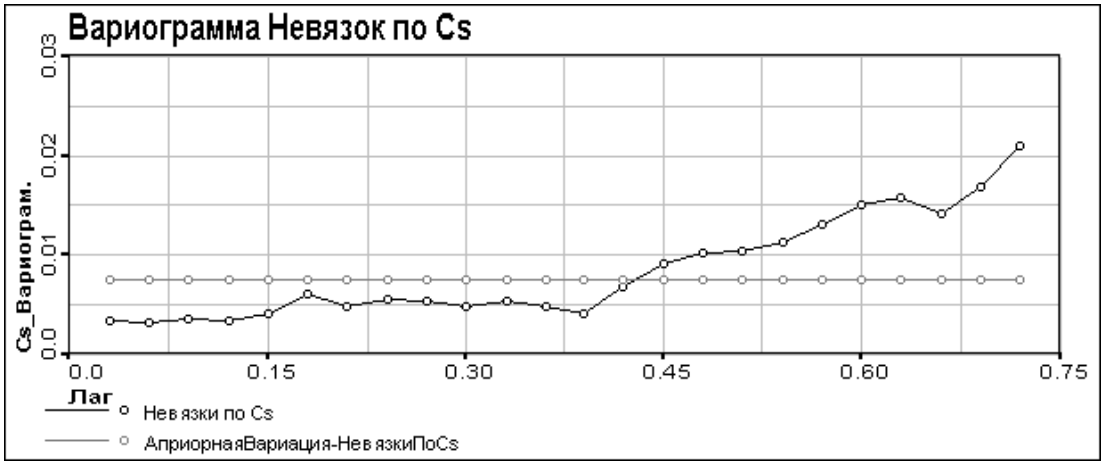

Рис. 4.3.4.11 Вариограмма невязок по ^{137}Cs

Оценки ^{137}Cs и ^{90}Sr в данном случае хорошо повторяют структуру данных, но сильно сглаживают локальные экстремумы. Интересно, что вариограммы оценок ^{90}Sr лучше соответствуют вариограммам исходных данных, чем вариограммы оценок ^{137}Cs, несмотря на меньшее расхождение данных и лучшее соответствие оценок по ^{137}Cs, чем по ^{90}Sr.

Из представленных на Рис. 4.3.4.11 и 4.3.4.12 вариограмм невязок по ^{137}Cs и ^{90}Sr видно, что невязки имеют пространственную корреляцию. Она обусловлена сглаживанием оценок.

4.3.5 Сравнение Результатов Различных Вариантов Анализа

После проведения различных вариантов тестирования обученной карты на тестовом наборе с восстановлением данных проведено сравнение полученных результатов между собой, а также с результатом оценки ^{137}Cs и ^{90}Sr при классификации тестового набора с полными данными.

На Рис. 4.3.5.1 приведены вариограммы различных вариантов оценки ^{137}Cs в сравнении с вариограммами исходных данных. Можно видеть, что с увеличением числа восстанавливаемых переменных значения вариограммы ^{137}Cs становятся все меньше и меньше. При оценке по всем переменным и без ^{90}Sr вариограммы оценок ^{137}Cs практически полностью совпадают с вариограммой исходных данных. При восстановлении данным по ^{137}Cs и при совместном восстановлении ^{137}Cs и ^{90}Sr

значения вариограммы падают, причем даже ниже, чем у вариограммы локальных средних. Это происходит вследствие усреднения данных, сглаживания высоких значений.. Несмотря на падение значений, эффективный радиус корреляции примерно одинаков во всех случаях, что означает сохранение корреляционной структуры.

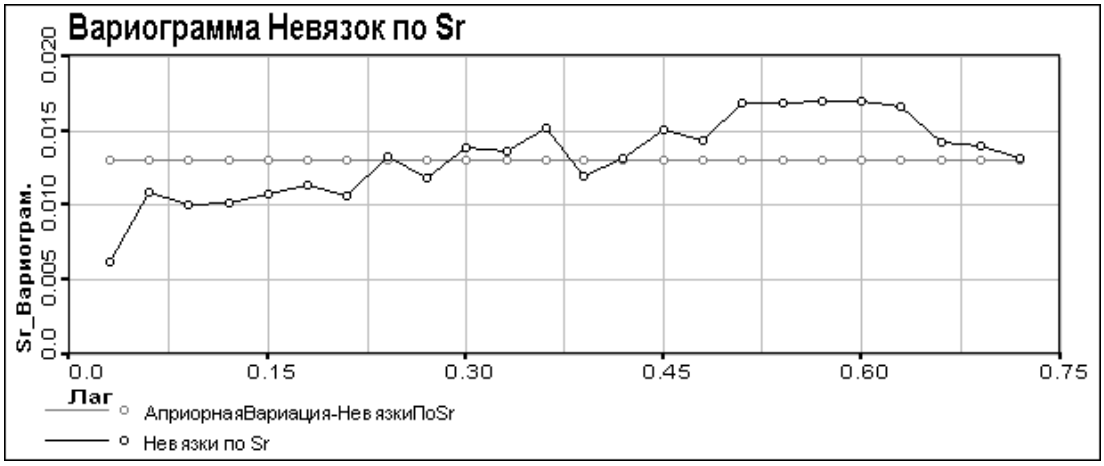

Рис. 4.3.4.12 Вариограмма невязок по ^{90}Sr

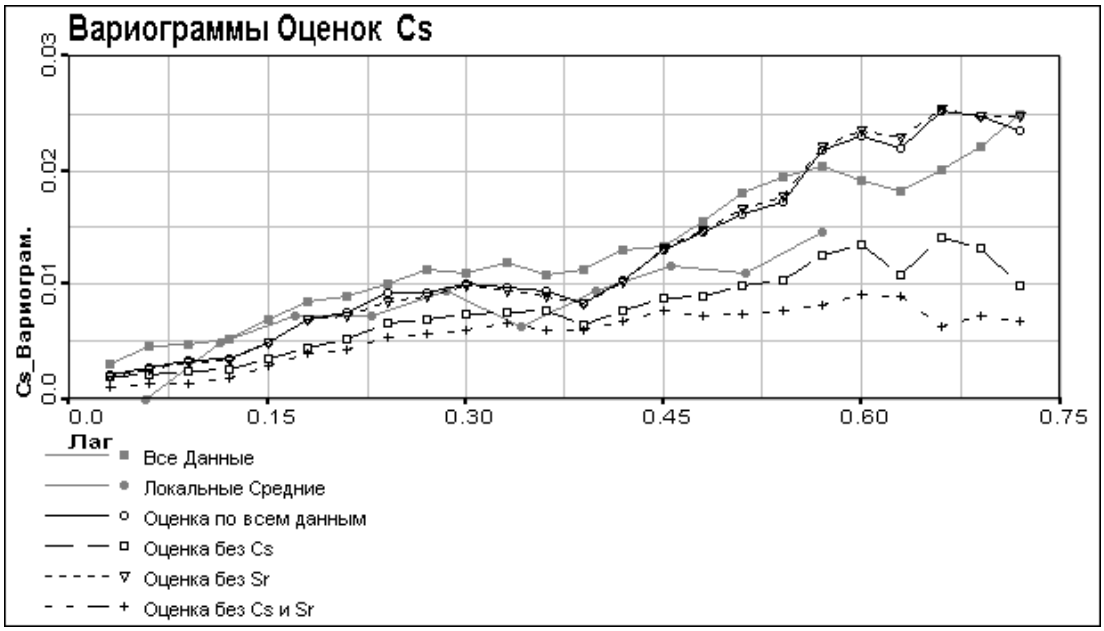

Рис. 4.3.5.1 Вариограммы Оценок ^{137}Cs

Из вариограмм для оценок ^{90}Sr, приведенных на Рис. 4.3.5.2. видно, что все они совпадают между собой и с вариограммой для локальных средних, оставаясь меньше вариограммы всех данных. Это означает, что несмотря на расхождения между обучающим и тестовым наборами, которые по ^{90}Sr заметнее, чем по ^{137}Cs (см. Часть 3.3), в целом структура выпадений ^{90}Sr более простая и сильнее коррелирована с координатами.

На Рис. 4.3.5.3 и 4.3.5.4 приведены сравнения вариограмм невязок по ^{137}Cs и ^{90}Sr для различных вариантов визуализации. Из этих вариограмм можно видеть, что вариограммы невязок при визуализации с неполными данными ощутимо больше, чем с полными. Это можно объяснить резким возрастанием самих значений невязок при визуализации неполных данных. Эффективный радиус корреляции примерно одинаков для всех вариантов визуализации неполных данных как для невязок по ^{137}Cs, так и по ^{90}Sr.

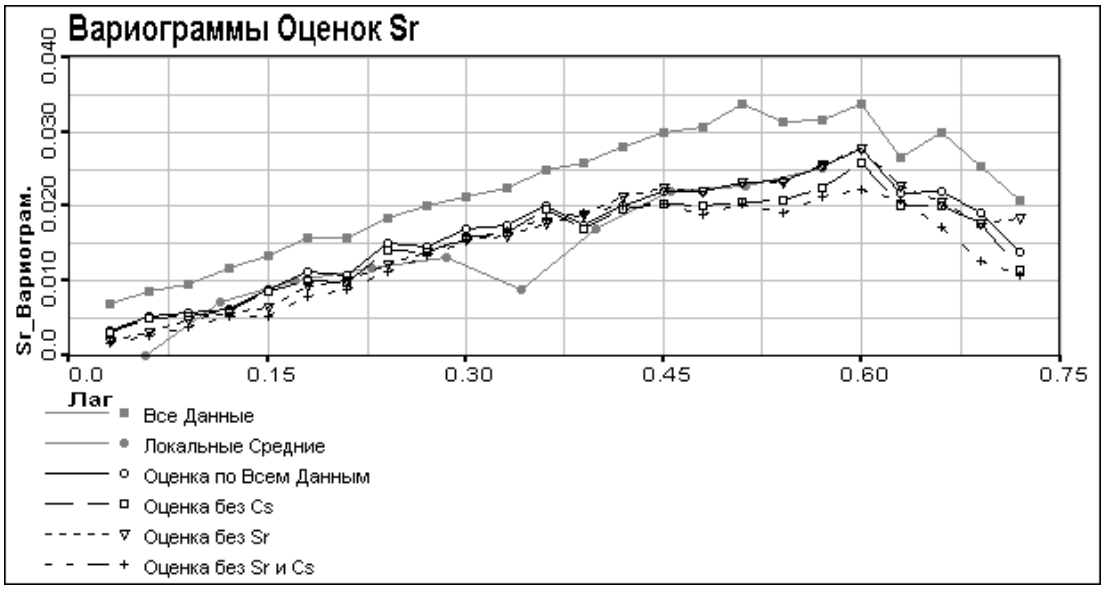

Рис. 4.3.5.2 Вариограммы Оценок ^{90}Sr

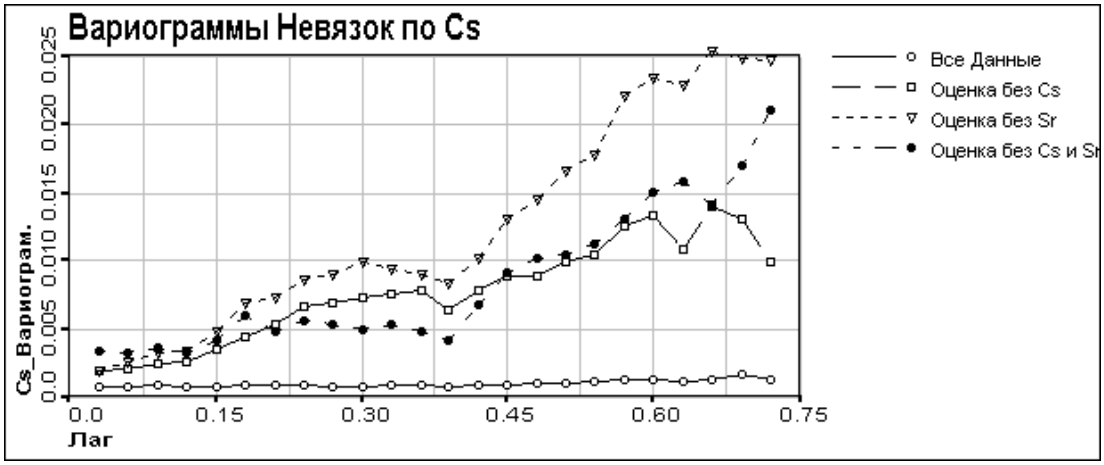

Рис. 4.3.5.3 Вариограммы Невязок по ^{137}Cs

5 Заключение

В заключение можно сделать следующие выводы:

- самоорганизующиеся карты Кохонена бесполезно применять для данных при отсутствии в них взаимной коррелированности;
- самоорганизующаяся карта Кохонена достаточно хорошо воспроизводит данные, на которых учится;
- обученная самоорганизующаяся карта успешно классифицирует данные без пропущенных переменных, соответсвующие (по свойствам) данным на которых проводилось ее обучение;
- для восполнения неизвестных значений в наборе с пропущенными данными самоорганизующаяся карта дает локально усредненное значение;
- невязки при восстановлении пропущенных значений имеют пространственную корреляцию, а значит могут быть уточнены, например с помощью кригинга;
- достоинство метода в том, что он позволяет работать с большим числом переменных;
- метод целесообразнее использовать для предварительной классификации неполных данных, это позволяет использовать влияние большого числа переменных, коррелированных с оцениваемой

переменных. Уточненное значение следует получать используя другой метод, учитывающий пространственную корреляцию невязок.

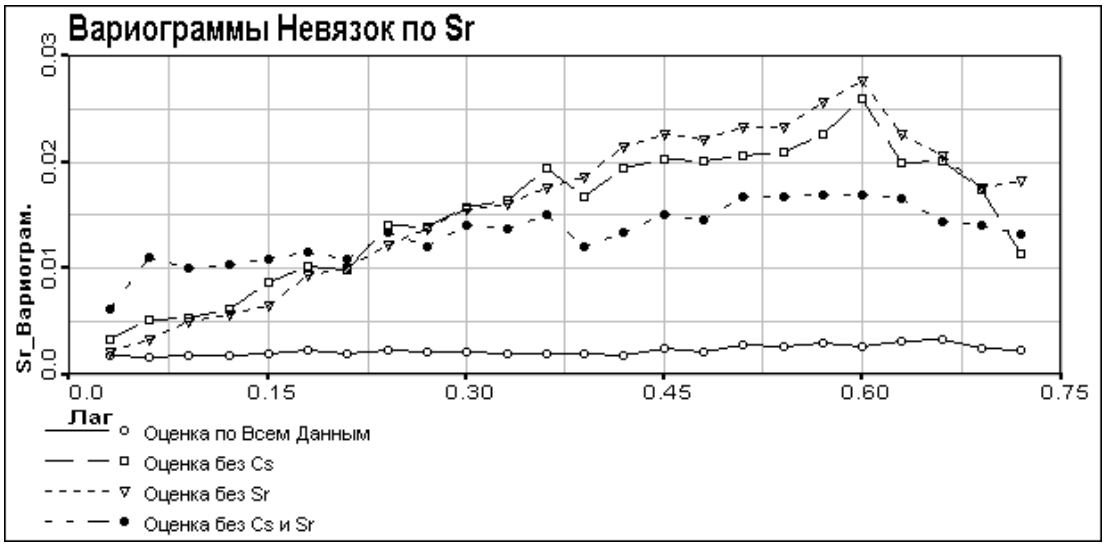

Рис. 4.3.5.4 Вариограммы Невязок по ^{90}Sr

Благодарности

Работа выполнена при частичной поддержке грантов ИНТАС 96-1957, ИНТАС 97-31726 и гранта для молодых ученых «Искусственные нейронные сети и генетические алгоритмы для анализа и моделировнаия пространственной информации по окружающей среде».

Литература

1. М.Ф. Каневский, В.В. Демьянов, С. Ю. Чернов. Совместный пространственный анализ ^{137}Cs и ^{90}Sr чернобыльских выпадений. Препринт IBRAE-96-04. Москва, 1996.
2. T. Kohonen. Self-Organizing Maps. Springer-Verlag, Berlin, Heidelberg, 1995.
3. А.А. Ежов, С.А. Шумский. Нейрокомпьютинг и его приложения в экономике и бизнесе. М.:МИФИ, 1999.
4. T. Kohonen. Self-Organization and Associative Memory. Springer-Verlag, Berlin, Heidelberg, 1984. 3rd ed. 1989.
5. http://neuron-ai.tuke.sk/NCS/VOL1/p4_html/node35.html
6. T. Kohonen, J. Hynninen, J. Kangas, J. Laaksonen. SOM_PAK: The Self Organizing Map Program Package, Version 3.1. Helsinki University of Technology, Espoo, Finland, 1995.
7. S. Kaski. Data Exploration Using Self-Organizing Maps. ACTA POLYTECHNICA SCANDINAVICA. Mathematics, computing and management in engineering series No.82. Helsinki university of Technology,Espoo, Finland, 1997.