Российская Академия Наук

РОССИЙСКАЯ АКАДЕМИЯ НАУК

ИНСТИТУТ ПРОБЛЕМ БЕЗОПАСНОГО РАЗВИТИЯ АТОМНОЙ ЭНЕРГЕТИКИ

RUSSIAN ACADEMY OF SCIENCES

NUCLEAR SAFETY INSTITUTE

Препринт ИБРАЭ № IBRAE-1999-03

Preprint IBRAE-1999-03

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

КАРТИРОВАНИЕ ПРОСТРАНСТВЕННЫХ ДАННЫХ ПРИ ПОМОЩИ МНОГОСЛОЙНОГО ПЕРСЕПТРОНА И ГЕОСТАТИСТИКИ

Москва 1999

Moscow 1999

УДК 502.3

Кравецкий А.С., Демьянов В.В., Каневский М.Ф., Савельева Е.А., Тимонин В.А., Чернов С.Ю. КАРТИРОВАНИЕ ПРОСТРАНСТВЕННЫХ ДАННЫХ ПРИ ПОМОЩИ МНОГОСЛОЙНОГО ПЕРСЕПТРОНА И ГЕОСТАТИСТИКИ. Препринт № IBRAE-99-03. Москва: Институт проблем безопасного развития атомной энергетики РАН. 1999. 41 с. Библиогр.: 12 назв.

Аннотация

В работе исследуется методы картирования пространственных данных при помощи искусственных нейронных сетей и геостатистики. Изучается влияние конфигурации нейронной сети на качество модели. В результате получаются карты загрязнения. Модели применены к реальным данным по радиоактивному загрязнению Брянской области.

Kravetski A.S., Demyanov V.V., Kanevski M.F., Savelieva E.A., Timonin V.A, Chernov S.Y. MAPPING OF SPATIAL DATA WITH MULTILAYER PERCEPTRON AND GEOSTATISTICS. Preprint IBRAE-99-03. Moscow: Nuclear Safety Institute. 1999. 41 p. — Refs.: 12 items.

Abstract

In this work methods of mapping spatial data by Neural Network and Geostatistic are analyzed. The effect of Network configuration on model quality is studied. The results are presented as maps of soil contamination. All models are applied to real radioactive pollution data in Briansk region.

Картирование пространственных данных при помощи многослойного персептрона и геостатистики

А. С. Кравецкий, В. В. Демьянов, М. Ф. Каневский, Е. А. Савельева, 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) 955-11-51, akrav@ibrae.ac.ru

Содержание

1 Введение

В настоящей работе исследуются адаптивные методы анализа и прогнозного картирования пространственно распределенных данных. В этих методах предпринимается попытка извлечь максимум информации из набора данных, учитывая возможные ошибки измерений, неравномерную плотность сети мониторинга, и прочие помехи, встречающиеся при реальных измерениях. Данные по окружающей среде обладают неоднородностью как на крупных, так на мелких масштабах, что затрудняет их анализ. Для их анализа и моделирования в работе применяются совместные методы, являющиеся сочетанием методов геостатистики и адаптивных методов.

При обработке нечеткой или неточной информацией, например в тех случаях, когда “истинные” данные искажены ошибками измерений, детерминистические методы картирования могут давать “плохую” модель. Это обусловлено тем, что в таких методах зависимость данных в пространстве жестко определена детерминистическими функциями, которые выбираются заранее, исходя из предположений о распределении данных. Кроме того детерминистическая модель учитывает все резкие неоднородности данных, что часто бывает нежелательно. К примеру, в одной из точек анализируемого набора за счет ошибки в измерениях может быть сильно завышено значение моделируемой переменной. Детерминистический метод учтет это завышение, и, таким образом, модель будет давать худший результат, чем в случае адаптивных методов.

Одним из методов для решения задач пространственного картирования являются “Искусственные Нейронные Сети” (ИНС) – адаптивные (к данным) математические модели, способные обобщать информацию. Они обладают рядом преимуществ перед детерминистическими моделями: строят модель исходя из всех данных набора, менее чувствительны к выбросам, чем детерминистические, и т. п.

Таким образом, ИНС проявляют себя как достаточно мощный инструмент анализа экспериментально полученной информации.

Однако, часто в невязках остаются мелкомасштабные структуры. То есть, ИНС извлекает не всю информацию из предоставленного ей набора данных. Существуют методы, которые позволяют исследовать тонкую структуру невязок и извлечь оставшуюся в них информацию.

Одним из таких методов является геостатистика. Метод анализа невязок нейронной сети при помощи геостатистики (“Neural Network Residuals Kriging” – NNRK) был предложен [1], развит [2], (приложения метода [3, 4, 5]). Результат анализа невязок при помощи геостатистики будет являться окончательным ответом (если невязки не имеют пространственной структуры, то ответом является, модель построенная ИНС).

В настоящей работе рассматриваются:

- Особенности картирования пространственных данных при помощи ИНС
- Проводится анализ данных при помощи ИНС и кригинга
- Исследуется зависимость качества модели от конфигурации нейронной сети

В качестве набора данных для анализа, в работе использовались реальные данные по загрязнению Брянской области радиоактивными элементами, в результате аварии на Чернобыльской АЭС.

2 Описание работы

2.1 Цели и задачи

Задача – пространственная интерполяция при помощи нейронных сетей и кригинга невязок, исследование устойчивости результатов, анализ неопределенностей и чувствительности.

Цель работы состояла в применении нейронных сетей различной архитектуры к выборкам с различным числом данных и дальнейшем анализе невязок при помощи кригинга. В работе исследовалось влияние архитектуры нейронных сетей на результаты и зависимость качества интерполяции от величины выборки.

Объектом специального исследования являлась пространственная корреляционная структура. Крупномасштабная составляющая корреляционной структуры моделируется нейронной сетью, в то время, как корреляции на локальном масштабе остаются в невязках (Z - Z_net).

В работе использовались ИНС с одним и двумя внутренними слоями. Число нейронов на скрытых слоях варьировалось от 7 до 30. В случае ИНС с двумя скрытыми слоями, на каждом слое бралось примерно одинаковое количество нейронов. Ставилась цель исследовать качество прогнозов ИНС.

После обучения, ИНС применялись к данным, на которых производилось обучение нейронных сетей. Получившиеся при этом невязки анализировались при помощи геостатистики. Т.е. строились вариограммы невязок. Если вариограммы получались стационарными и среднее было равно константе на всей области (обычно тренд исключается не полностью, поэтому принимается гипотеза, что кригинг можно проводить и при частичном исключении тренда), то строились модели вариограмм, и производилась интерполяция невязок при помощи кригинга, с соответствующими моделями вариограмм.

2.2 Основные идеи метода кригинга невязок ИНС

Кригинга невязок ИНС[1-5] используется для картирования пространственных данных.

Поясним суть метода:

Пусть, существует набор пространственных данных (сеть мониторинга). Обычно, данные представляются в виде: X, Y – пространственные координаты, Z – зависящая от них переменная. Нашей задачей является картирование Z на равномерной координатной сетке. Для этого:

1. Конструируется нейронная сеть.
2. Сеть обучается на выбранном наборе данных для обучения.
3. Сеть применяется к данным для обучения, в результате, вычитанием из исходных данных получаются невязки.

4. Невязки анализируются методами геостатистики. Если выясняется, что в них не осталось корреляционной структуры, то ответ ИНС является окончательным ответом.

5. Если выясняется, что в них еще осталась информация (невязки имеют пространственную корреляцию), то проводится анализ при помощи геостатистики.

Анализ при помощи геостатистики проводится следующим образом:

1. Строятся анизотропные вариограммы невязок. Проводится статистика движущегося окна для исследования поведения среднего на всей области.

2. Если среднее на всей области постоянно и вариограммы показывают стационарность процесса, то на основе вариограмм невязок, строится вариограмная модель.

3. Проводится кригинг невязок, с использованием построенной вариограмной модели

После того, как нейронная сеть обучена и решены уравнения обычного кригинга, ИНС применяется к равномерной координатной сетке, и на этой же сетке проводится кригинг. Окончательный результат получается сложением прогноза ИНС и кригинга, в каждой точке равномерной сетки.

3 Многослойный персептрон

Существует множество разновидностей ИНС, более подробную информацию о них можно получить в [6]. Здесь будет кратко рассмотрен частный случай ИНС – многослойный персептрон.

Нейронные сети – это математические системы, способные отыскивать закономерности и строить модели, пользуясь этими закономерностями. ИНС – это адаптивная модель. Она не зависит от модельных параметров, а зависит только от архитектуры ИНС и от самих данных. Кроме того, эта модель нелинейная по параметрам. Нейронные сети позволяют выявить закономерности в наборе данных для обучения и использовать эти закономерности для интерполяции незнакомых данных.

В общем случае нейронные сети подразделяются на использующие алгоритмы обучения с учителем и обучающиеся без учителя. Алгоритм обучения с учителем заключается в том, что используя набор известных значений входов и выходов, сеть подстраивает свои параметры так, чтобы достигнуть наилучшего соответствия между известными значениями выходов и значениями выходов предсказанными сетью.

Многослойный персептрон является нейронной сетью, использующей алгоритм обучения с учителем. В этом случае параметры сети (веса) настраиваются в соответствии с набором данных, состоящим из комбинации входных и выходных данных.

Итак, персептрон состоит из: **нейронов входного слоя**, которые на самом деле являются ячейками, способными принимать определенные числовые величины, то есть независимыми переменными; **связей** (которым приписываются **веса)**, которые отображают маршрут “перемещения” информации, то есть показывают что значения с выхода нейрона, соответствующего началу связи, должно поступить на вход нейрона, соответствующего ее концу; **скрытых нейронов**, обладающих способностью принять информацию со своего входа, преобразовать ее в соответствии с **передаточной функцией** данного нейрона, и отправить результат на свой выход**; выходных нейронов** – ячеек, в которых формируется ответ.

Смысл веса связи заключается в том, что связь не просто передает число от своего начала к концу, а еще и домножает его на вес – численный коэффициент. Веса связей ИНС – это именно те параметры, которые подстраиваются в процессе обучения.

Алгоритмы обучения без учителя применяются для подстройки параметров ИНС в задачах классификации. Эти алгоритмы в работе не рассматриваются.

Рассмотрим поподробнее структуру ИНС:

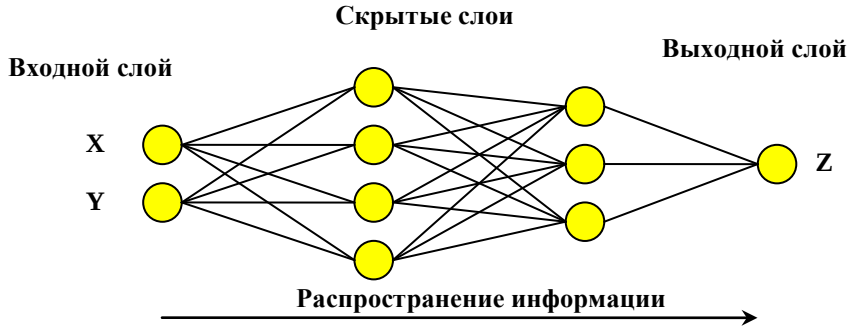


Рис. 1: Общий вид многослойного персептрона

На рис. 1 приведен пример персептрона 2-4-3-1, по структуре такого же, как и используемые в работе. Он состоит из двух нейронов во входном слое, одного нейрона на выходном и семи нейронов в скрытых слоях. Каждый нейрон предыдущего слоя связан с каждым нейроном последующего слоя.

На этом примере мы попытаемся объяснить, что же из себя представляет обработка данных персептроном. Ясно, что вне зависимости от конкретного числа нейронов в каждом слое и числа слоев суть обработки данных нейронной сетью не меняется, поэтому для простоты объяснения мы будем рассматривать ИНС с двумя входами и одним выходом, тем более, что именно такие ИНС применялись во время исследования.

3.1 Обработка информации персептроном

Более подробно функционирование многослойного персептрона описано в [6].

1. На входном слое: z_1 = X и z_2 = Y (X и Y – пространственные координаты). Мы подаем пространственные координаты первой точки из набора данных для обучения на вход ИНС. Ее выход будет функцией от этих двух переменных.

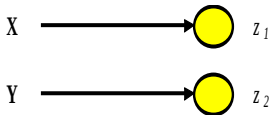


Рис 2: Подача возбуждения на вход

2. Если w_{ji} – это вес связи i-ого нейрона предыдущего слоя с j-ым следующего.
То, на каждый нейрон следующего(скрытого) слоя поступает величина:

$$a_i = \sum_i w_{ij} * z_i \qquad (1)$$

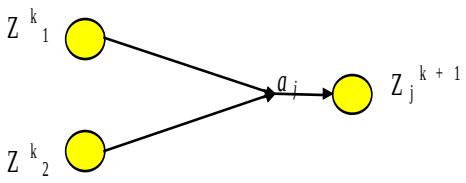


Рис 3: Преобразование информации связями

3. В нейроне величина преобразовывается в выходное значение, как

$z_j = g_j(a_j)$, где $g_j(a_j)$ – активационная функция j-ого нейрона

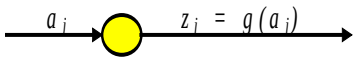


Рис 4: Преобразование данных нейроном

В качестве *g(a)* можно взять любую строго монотонно возрастающую от 0 до 1 непрерывную функцию. Часто в качестве активационной функции берут логистическую (рис. 5), так как для нее легко вычисляется производная:

$$g(a) = \frac{1}{1+\exp(-a)} \tag{2}$$

$$g'(a) = g(a)(1-g(a))$$

Причем обычно для всех нейронов берут одну и ту же активационную функцию. Именно так делалось в работе.

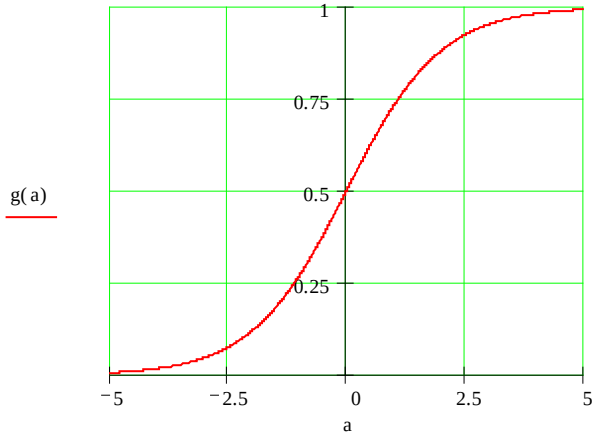


Рис. 5: График функции g(a)

4. Процесс **1-3** повторяется на каждом последующем слое, вплоть до достижения выходного слоя.

5. Величина Z, полученная на выходном слое является реакцией ИНС на раздражение, поданное на вход сети.

3.2 Обучение ИНС

Существует множество методов обучения ИНС. В настоящей работе, обучение ИНС производилось методом обратного распространения ошибки (Back Propagation Error)[6].

Кратко поясним его суть:

Имеется набор данных для обучения. Т. е. набор троек чисел (X,Y,Z). Смысл обучения состоит в том, что надо так подстроить веса связей ИНС, чтобы, подавая на ее вход координаты X и Y, получать на выходе значения Z_net как можно более близкие к реальным. Для этого на вход ИНС подаются поочередно X и Y, для которых величина Z известна. И, в соответствии с ошибкой на каждом нейроне, меняются веса связи.

Обычно, имеющийся набор значений на входе и соответствующих им значений на выходе, делят на два набора – тренировочный и тестовый. Первый из которых, служит для модификации весов ИНС, то есть для ее обучения, а второй для контроля качества этого обучения.

Рассмотрим процесс обучения поподробней:

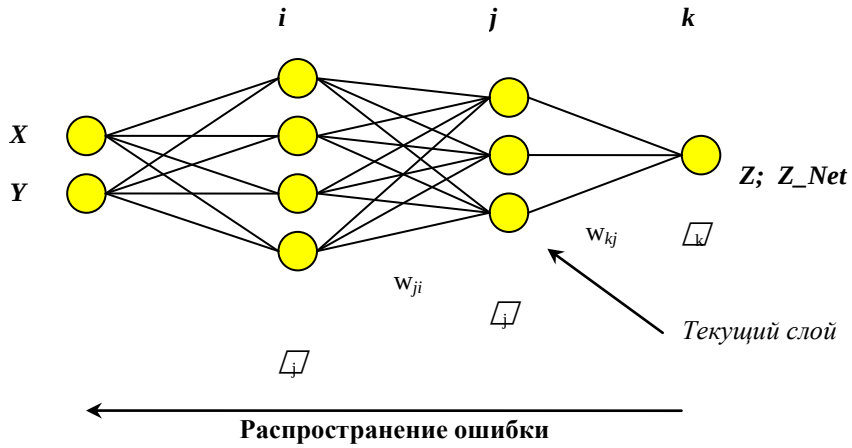


Рис 6: Метод обучения "Обратное распространение ошибки"

Предварительным этапом обучения является инициализация весов ИНС. Обычно, веса задаются случайным образом на отрезке, включающем ноль.

1. На вход ИНС подаем X и Y из тренировочного набора, после ее применения на выходе получаем Z_net – величину, спрогнозированную ИНС, в соответствии с текущими значениями весов.

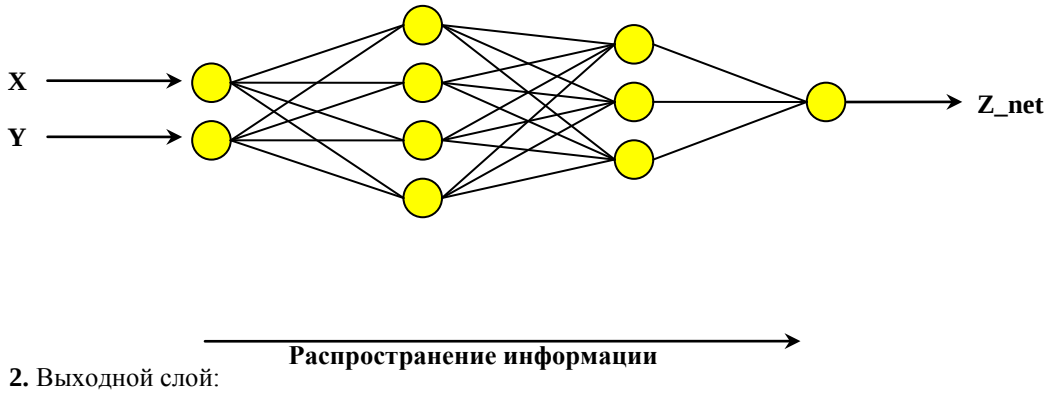


Рис 7: Получение "промежуточного" прогноза

2. Выходной слой:

Z – известное значение (цель)
Z_net – отклик ИНС

$\delta_i = Z\_net - Z$

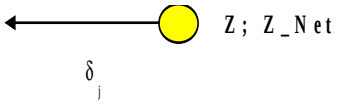


Рис 8: Вычисление ошибки

На выходном слое находим ошибку – разность между истинным значением Z и откликом ИНС. Если бы выходов было несколько, то ошибку надо было бы искать для каждого выхода.

3. На каждом скрытом слое:

$$\delta_j = z_j(1 - z_j)\sum w_{ij} * \delta_k \qquad (3)$$

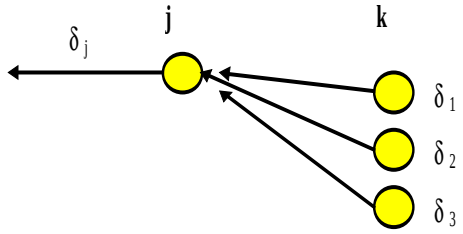


Рис 9: Распространение ошибки

где $z_j\ (1 - z_j)$ – это $g'(a_j) = g(a_j)\ (1 - g(a_j))$;
z_j – величины, на выходе текущего нейрона;
δ_k – ошибки на предыдущем слое.

Вычисляем ошибки для внутренних слоев.

4. Веса модифицируются по правилу:

$$\Delta w_{ji} = -\eta \delta_j x_i ,$$

где x_i – величина, поступившая на i - ый нейрон;
η – скорость обучения (в работе менялась от 0.1 до 0.5, в зависимости от количества данных и архитектуры ИНС).

5. Рассчитывается ошибка на тестовом наборе, если она меньше, чем на предыдущей итерации, то веса ИНС временно считаются наилучшими.

6. Если условие окончания обучения не выполнено (в работе – 200000 циклов без уменьшения ошибки на тестовом наборе), то подается следующая точка (X,Y,Z). Если выполнено, то берется последний набор временно наилучших весов, его считают наилучшим и обучение останавливают.

4 Алгоритм проделанной работы

Особый интерес представляет зависимость качества прогнозов ИНС от конфигурации нейронной сети. Очевидно, что при количестве нейронов в скрытых слоях много меньшем количества данных для обучения, прогноз ИНС будет резко отличаться от данных, на которых строилась модель. При слишком большом количестве скрытых нейронов может наблюдаться переобучение, т. е. ИНС будет давать хорошие результаты на данных для тренировки и плохие на незнакомых ей данных.

Для анализа зависимости качества прогнозов ИНС от конфигурации нейронной сети были сконструированы 6 нейронных сетей (2-7-1, 2-15-1, 2-30-1, 2-7-7-1, 2-15-12-1, 2-30-25-1).

1. В качестве набора данных для обучения и валидации были взяты измерения загрязненности почвы ^{137}Cs в брянской области [7]. Всего использовалось 665 данных. Из них случайным образом было получено 3 набора данных для обучения, по 100, 300 и 500 точек. Остаток: 565, 365 и 165 точек соответственно, считался данными для валидации (рис. 10).

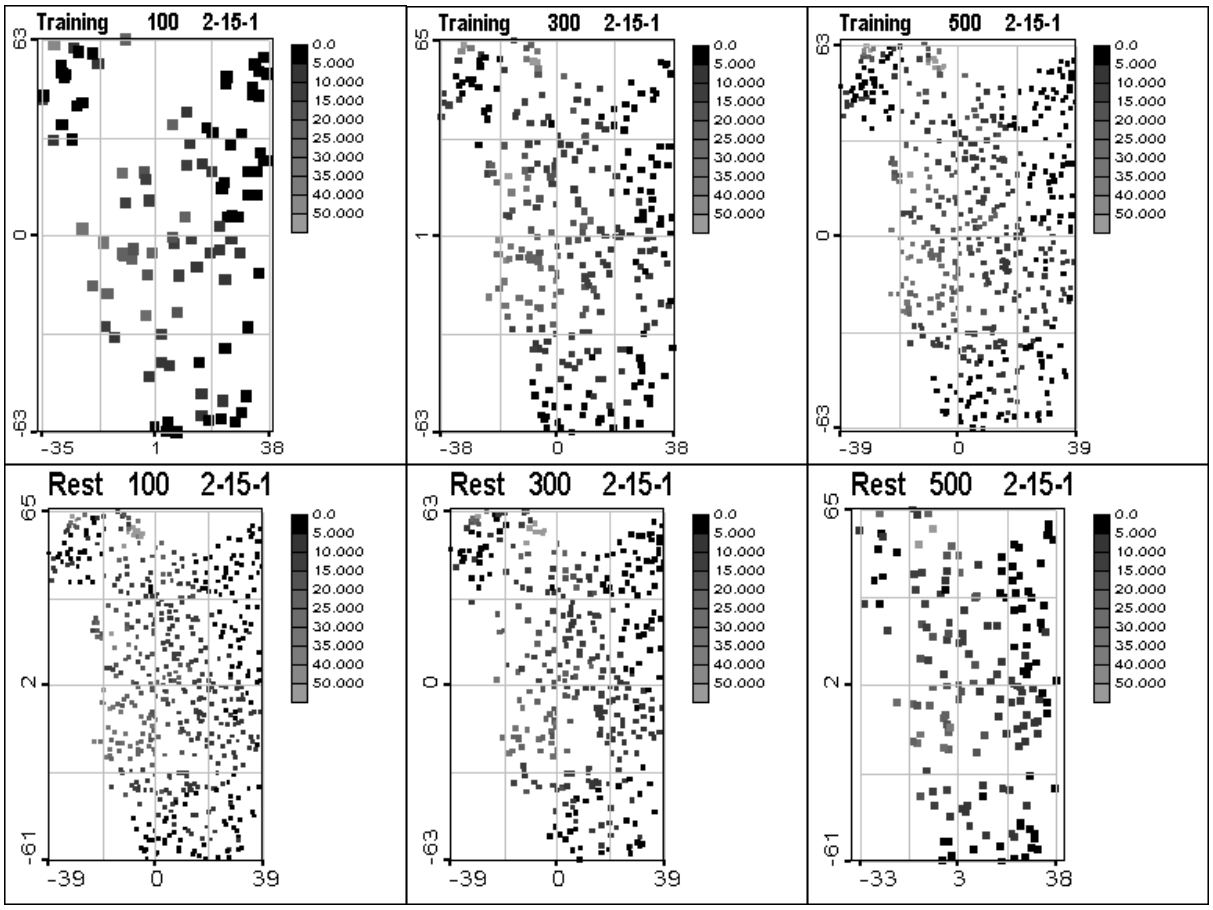


Рис 10: Карты данных для обучения и для валидации

2. Из каждого набора данных для обучения выделялись случайным образом 20% точек, далее они использовались, как тестовые, т. е. при достижении на них минимума ошибки, производилась запись весов ИНС. Остальные 80% являлись данными для тренировки, на которых обучалась ИНС. Обучение производилось методом обратного распространения ошибки.

Для построения, обучения и применения ИНС использовалась программа NeuroShell 2

3. Начальные веса выбирались случайным образом на отрезке [-0.3; 0.3], данные нормировались на интервал (-1; 1). Тренировка прекращалась, когда в течение 200 тыс. циклов ошибка на тестовых данных не становилась меньше (ошибка на тренировочных данных ~ 0.01).

4. После тренировки, ИНС использовалась для интерполяции на равномерной прямоугольной координатной сетке (-38.9> 80x0.990 : -62.5> 125x1.02), для визуального контроля качества построенной модели.

5. Далее ИНС применялась к данных для обучения, вычислялись: среднее невязок, RMSE – корня из средней квадратичной ошибки, и т. д. После этого вычислялись коэффициенты корреляции, строились гистограммы распределения невязок и т. п.

6. То же проводилось для валидационных данных.

7. Строились изотропные вариограммы, проводилась статистика движущегося окна.

8. Проверялась стационарность распределения значений невязок при помощи анализа вариограмм (вариограммы имели sill).

9. Среднее не было постоянным на всей области, однако, принималась гипотеза, что среднее меняется настолько слабо, что можно считать условия применимости кригинга выполненными.

10. Строилась анизотропная модель вариограмм.

11. На основании построенной модели решались уравнения кригинга с невязками, в качестве параметров.

12. В точках для тренировки, для валидации и на равномерной прямоугольной координатной сетке производилась интерполяция данных при помощи кригинга и ИНС.

13. Программное обеспечение:

    Весь статистический анализ, кроме вариограмм, проводился в программе STATISTICA. Вариограммы, карты распределения ошибок и карты на равномерной сетке строились в GEO STAT OFFICE [8, 9], 3PLOT [10]. Кригинг производился при помощи программы GSLIB[12].

14. Результаты моделирования представляются в виде:

- Карт оценок ИНС на равномерной сетке
- Гистограмм распределения данных
- Вариограмм невязок
- Трехмерных графиков распределения ошибок, эффективных радиусов и т. п.
- Таблиц со статистикой моделей ИНС
- Примеров распределения ошибки, представленных в виде полигонов Вороного
- Карт оценок невязок при помощи кригинга

Условные обозначения:

2-7-1	Конфигурация ИНС (2 входных нейрона, 7 внутренних, 1 выходной)
100	Число данных, на которых производилось обучение
Rest	Валидационные данные
Training	Тренировочные данные
Z_net	Величина Z, предсказанная ИНС
Dif = Residual=Z - Z_net	Невязки

$RMSE = \frac{\sum_{i=1}^{N} \sqrt{(Z_i - Z\_net_i)^2}}{N}$ – Средняя квадратичная ошибка по ансамблю.

5 Обсуждение результатов использования ИНС

Все выводы относительно ИНС следует рассматривать, как результаты, полученные при работе в среде NeuroShell 2.

Согласно алгоритму работы, была произведена интерполяция данных при помощи ИНС различных конфигураций.

При обучении ИНС применялся метод, при котором данные для обучения разбиваются на два набора. На одном сеть тренируется (тренировочный набор), на другом при каждой итерации вычисляется ошибка (тестовый набор). При обучении ИНС по такому алгоритму стремятся к минимизации ошибки на тестовом наборе. Таким образом, ИНС при тренировке не «переучивается». Дело в том, что при минимизации ошибки на тренировочном наборе ИНС, начиная с определенного момента, выучивает их «слишком хорошо» и теряет способность обобщать, т.е. уменьшается ошибка на тренировочных данных, зато растет ошибка на данных, незнакомых нейронной сети. При использовании же обучения с минимизацией ошибки на тестовом наборе этот эффект удается снизить.

Выяснилось, что количество данных для обучения, гораздо сильнее влияет на точность полученной модели (критерием точности служила RMSE), чем конфигурация ИНС (прил. 12). Однако, при любом количестве данных для обучения существует **оптимальная** конфигурация ИНС (прил. 17). **Это говорит о том, что нельзя до бесконечности увеличивая число связей повышать точность построенной модели.** По-видимому, если пользоваться алгоритмом с минимизацией ошибки на тестовом наборе, для каждого конкретного набора данных существует предел точности, которую можно достичь посредством обработки данных с помощью нейронной сети.

Из графиков 15, 16, 17 следует, что количество связей оказывает ощутимое влияние на RMSE, только в том случае, когда ИНС применяется к данным для обучения. В случае же валидационных данных явная связь наблюдается только при очень небольшом количестве связей. **Таким образом, количество связей не оказывает существенного влияния на невязки в точках, которые не применяются для обучения.**

Видно, что общую структуру данных отражают все модели. Различия заключаются в мелких деталях: формах линий уровня, положениях максимумов, и т. п. Судя по всему, это объясняется флуктуациями в пределах погрешности моделей. **Отсюда следует, что качественное представление о пространственном распределении данных можно получить даже при использовании ИНС с небольшим числом нейронов.**

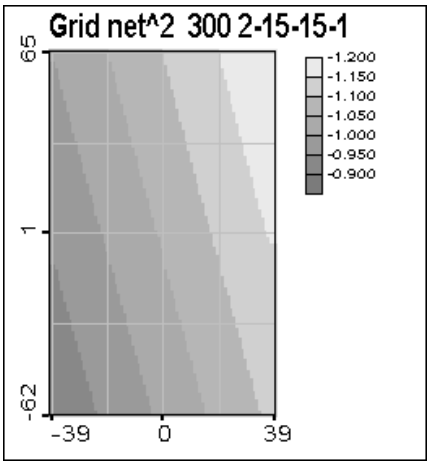


Рис 11: Результат обработки невязок при помощи ИНС

Была предпринята попытка использовать новую ИНС для интерполяции невязок, полученных при интерполяции значений зависимой переменной (Z). Однако, результатом такой интерполяции оказался слабый линейный тренд (рис 11).

Еще одой численной характеристикой качества модели может служить линейная корреляция между исходными значениями величин и соответствующими им предсказанными значениями. Здесь также оказывается, что существуют ИНС с наилучшими показателями (коэффициентом корреляции). Причем не всегда эти показатели улучшаются при увеличении числа нейронов (прил. 4).

Более подробный анализ показывает, что ИНС занижает значения в максимумах и завышает значения в минимумах (прил. 5, 10). Особенно хорошо это видно на валидационных данных. Таким образом, основной вклад в RMSE вносят несколько точек с относительно большими значениями Z. Эти точки находятся в максимумах, причем обычно сильно отличаются от близлежащих по оцениваемой величине. И действительно, при отбрасывании этих точек корреляция заметно увеличивается (прил. 9, 11). Значит, ИНС лучше предсказывает значения близкие к среднему, чем сильно от него отклоняющиеся. Это дает качественную оценку применимости модели, построенной ИНС.

Следует отметить, что ИНС могла извлечь не всю информацию из анализируемых данных. Это значит, что невязки ИНС могут иметь пространственную структуру. Существуют методы, с помощью которых это можно обнаружить и извлечь оставшуюся информацию из невязок.

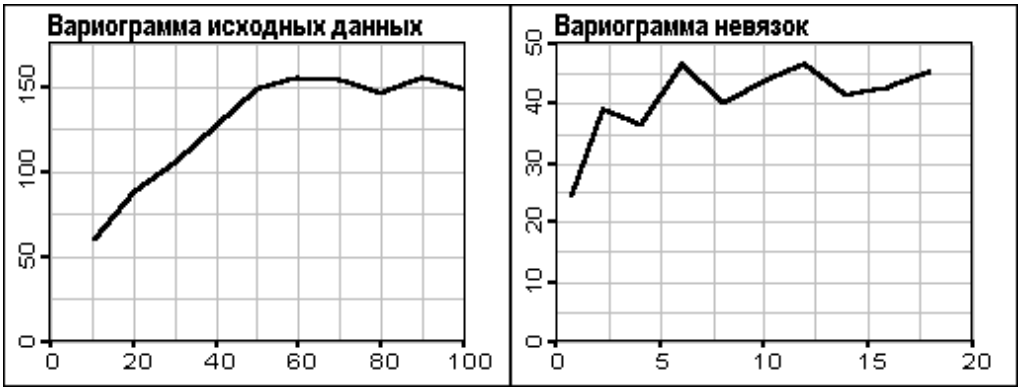


Рис. 12: Пример вариограмм данных для обучения

Невязки ИНС можно проанализировать с помощью методов геостатистики. Построив вариограммы, можно обнаружить, что у невязок есть эффективный радиус, т. е. на расстояниях порядка 5 км. невязки коррелированны между собой. Однако, на расстояниях больше 5 км. такой корреляции не наблюдается. Если принять гипотезу о том, что среднее меняется слабо, то к ним можно применять кригинг, результатом которого станет “точная” модель.

Из рис. 12 видно, что у исходных данных эффективный радиус был порядка 60 км., в то время как у невязок он порядка 7 км. это означает, что ИНС выделила крупномасштабный тренд, оставив только мелкомасштабную структуру.

Из графиков 20 и 21 видно, что существует оптимальная конфигурация ИНС, выделяющая наиболее мелкомасштабную структуру. Это явление, как и существование ИНС с наименьшей RMSE на тренировочных данных, говорит о том, что имеет смысл использовать ИНС с количеством связей, принадлежащем определенному промежутку, зависящему от количества данных. Слишком маленькое или слишком большое количество связей ведет к ухудшению качества модели. В качестве метода для отыскания оптимума, можно предложить построение определенного числа моделей с различным количеством связей, и нахождения максимума с помощью методов математики (например, построив график), статистики или искусственного интеллекта (например, генетические алгоритмы).

6 Кригинг

6.1 Система уравнений кригинга

Существует метод, который может быть использован практически для всех пространственных данных, обладающих корреляцией. Этот метод называется кригинг [11].

Основная идея метода заключается в том, что значение прогнозируемой переменной в точке ищется как взвешенная сумма значений этой переменной в известных точках. Решение системы уравнений кригинга (6) относительно весов со значениями прогнозируемой переменной в известных точках в качестве параметров позволяет прогнозировать значения этой переменной в точках, где ее значение неизвестно.

Выпишем уравнения обычного кригинга. Именно он использовался в работе.

Обычный Кригинг [11] – "лучший" несмещенный линейный оцениватель. Лучший в смысле минимизации вариации ошибок (5). Линейный, так как оценки являются взвешенными линейными комбинациями доступных данных (4). Несмещенный, так как ошибка имеет нулевое среднее значение. Кригинг не является детерминистическим методом.

Оценка в точке x_0:

$$V^*(x_0) = \sum_{j=1}^{n} w_j * V(x_j) \tag{4}$$

Ошибка оценки кригингом:

$$\sigma_R^2 = \sum_{i=1}^{n} w_i \gamma_{i0} + \mu \tag{5}$$

Из **Системы уравнений кригинга** (n+1 уравнение) находятся "веса" w_j**:**

$$\sum_{j=1}^{n} w_j \gamma_{ij} - \mu = \gamma_{i0}$$
$$\sum_{i=1}^{n} w_i = 1 \qquad (6)$$

где g **– Вариограмма** – вариация разницы значений переменной в двух точках как функция расстояния между ними и направления:

$$\gamma(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{i=1}^{N(\mathbf{h})} (V(x_i) - V(x_i + \mathbf{h})) \qquad (7)$$

Условия применимости Обычного кригинга:

Ограничением в применении Обычного кригинга является условие стационарности. Распределение называется стационарном в строгом смысле, если оно инвариантно относительно сдвига. Однако, на практике такое сильное предположение неприменимо. Вместо этого используют более слабые предположения, такие как внутренняя (intrinsic) гипотеза и стационарность второго порядка.

Функция называется внутренней, если математическое ожидание существует и не зависит от положения в пространстве, а вариограмма существует и ограничена. Функция обладает внутренней стационарностью второго порядка, если математическое ожидание существует и не зависит от положения в пространстве, а ковариация для каждой пары значений случайной переменной зависит только от расстояния между ними.

6.2 Кригинг невязок

Как уже упоминалось выше, если невязки обладают пространственной корреляцией, то их можно анализировать при помощи методов геостатистики. Для этого следует проверить выходит ли вариограмма на стационарный уровень. Если выходит, то можно построить модель вариограммы и на ее основе решить уравнения кригинга. После этого, окончательный ответ будет складываться из результата интерполяции при помощи ИНС и интерполяции невязок при помощи кригинга (Рис 13, 14).

Точность пространственной модели при этом существенно повышается.

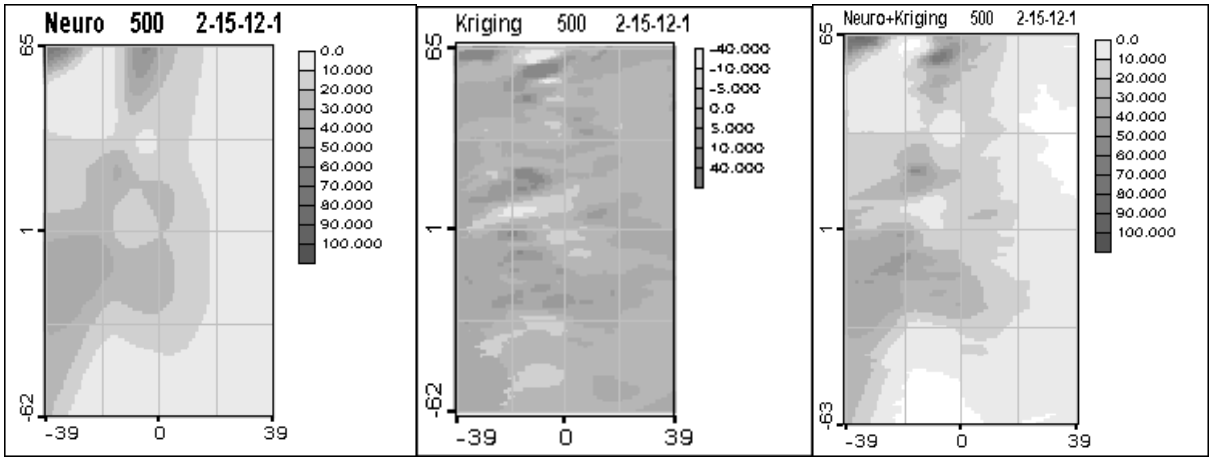


Рис 13: Получение окончательного ответа из анализа ИНС и кригинга невязок для конфигурации 2-15-12-1

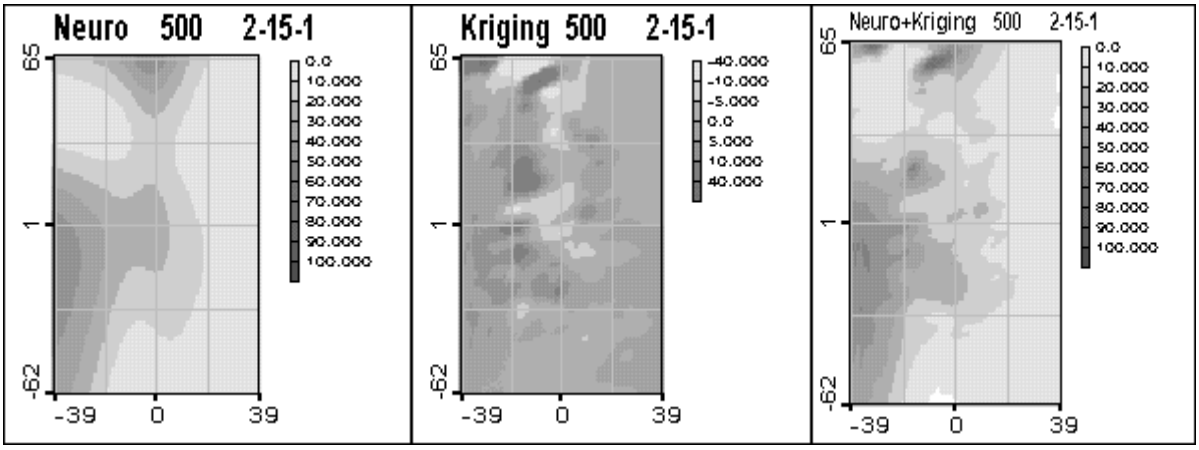


Рис 14: Получение окончательного ответа из анализа ИНС и кригинга невязок для конфигурации 2-15-1

Таблица 1. Характеристики модели, построенной ИНС и модели ИНС+кригинг

ИНС	2-15-12-1		2-15-1	
	Данные для обучения	Валидационные данные[1]	Данные для обучения	Валидационные данные[1]
Количество данных	500	165	500	165
RMSE после ИНС	6. 0	5. 0	7.5	5. 4
RMSE после кригинга	4. 7	2. 8	4. 4	2. 7
Корреляция оценки кригинга и истинного значения невязки	0. 62	0. 79	0. 8	0. 76

[1]Из валидационных данных при расчете RMSE была изъята одна точка (изолированный максимум), вносящая существенный вклад в ошибку.

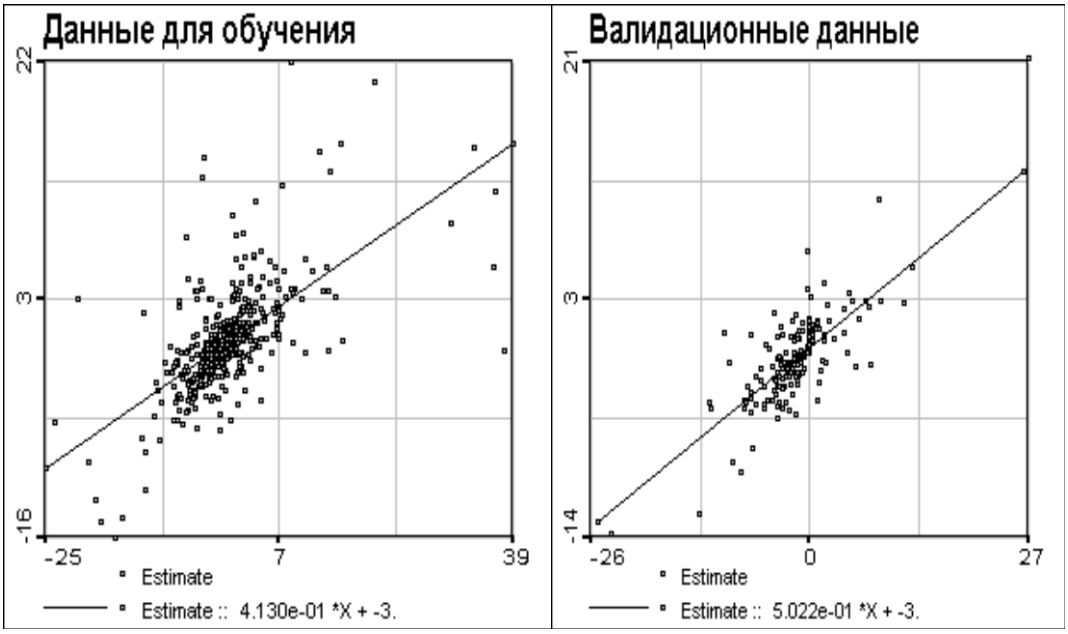


Рис. 15: Графики зависимости предсказанной величины невязки от истинной для конфигурации 2-30-25-1

Согласно алгоритму работы, были построены вариограммы невязок. Анализ вариограмм показал, что распределение обладает стационарностью. Было принята гипотеза, что среднее можно считать равным константе на всей области (рис.16). После этого были построены модели вариограмм (рис.17, 18) и проведен кригинг.

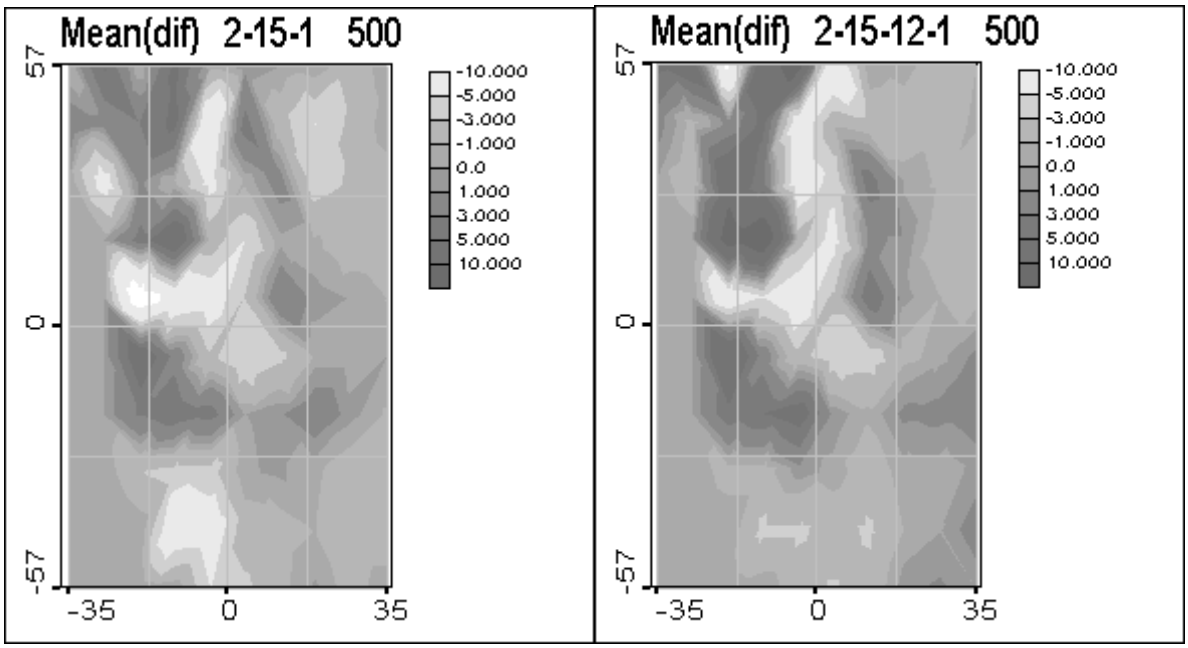


Рис.16: Распределение среднего невязок для конфигурации 2-15-1 и 2-15-12-1, 500 данных

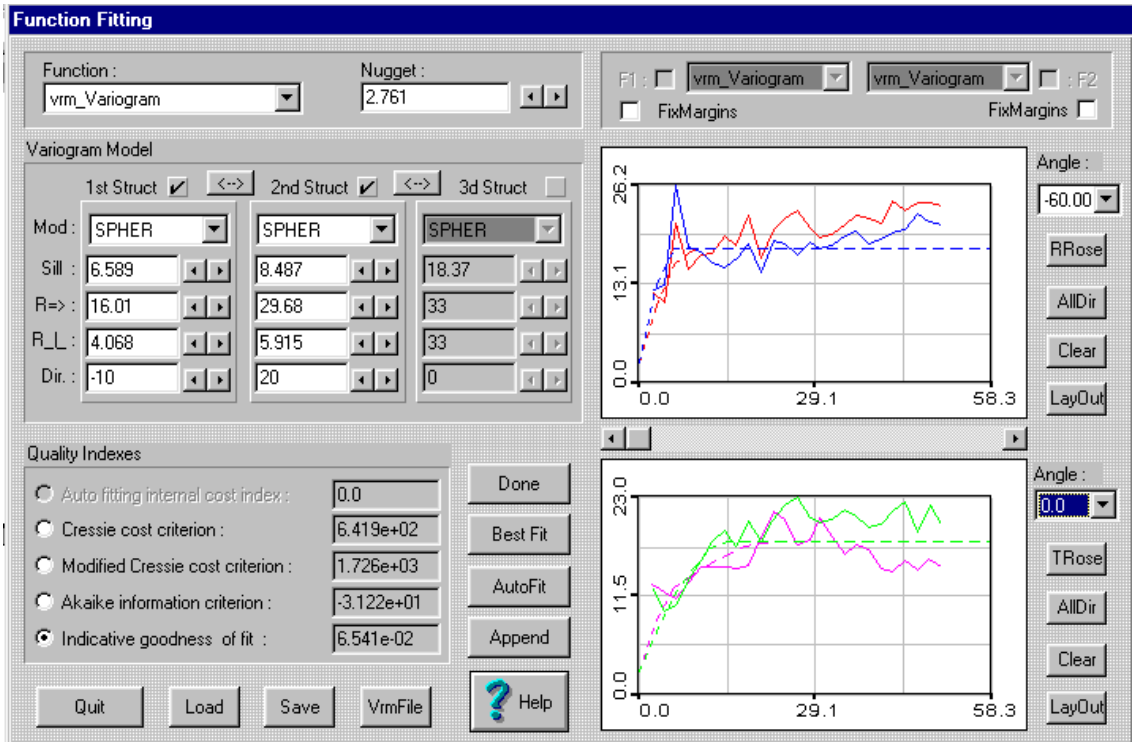


Рис.17: Модель вариограммы для конфигурации 2-15-12-1

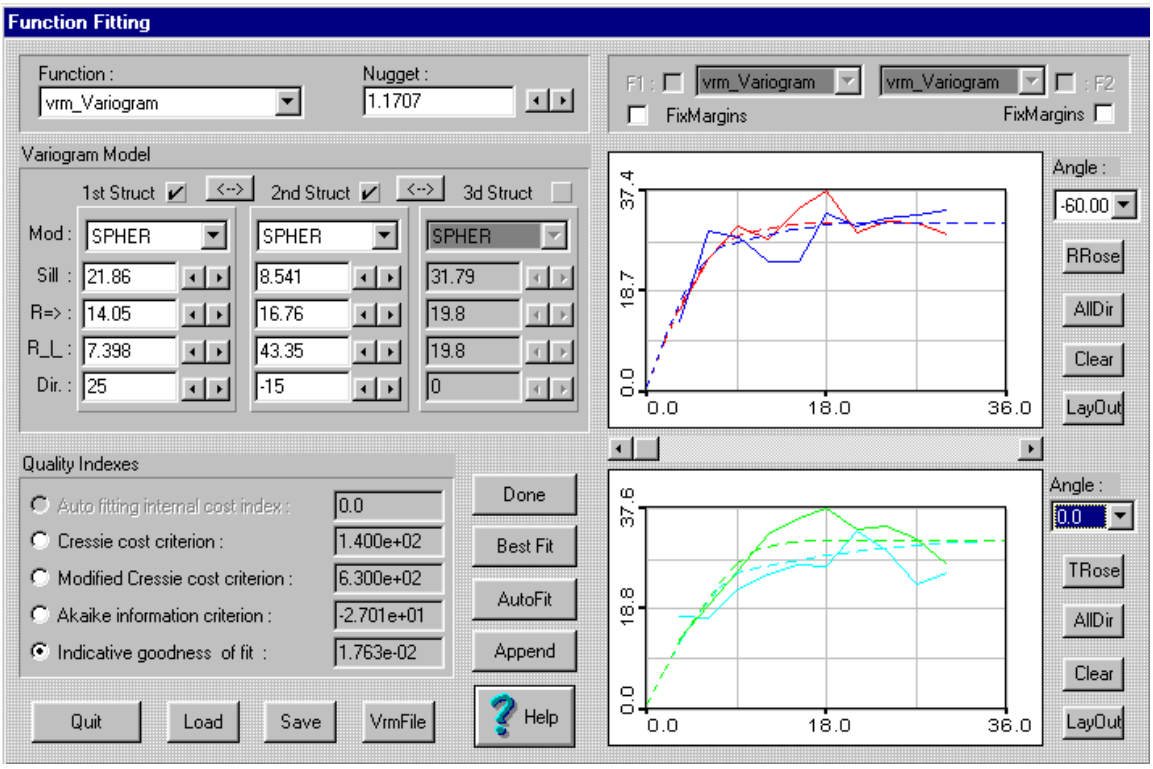


Рис.18: Модель вариограммы для конфигурации 2-15-1

Точность интерполяционной модели после кригинга заметно возросла (таб. 1).

Важным результатом является совпадение порядков величин у валидационных данных и данных для обучения. Это соответствует основной цели прогнозирования – предсказанию переменных в неизвестных точках. В данном случае для валидационных данных ошибка даже меньше, возможно, это объясняется тем, что их было меньше.

Если построить поверхность вариации оценки кригинга невязок, то наглядно видно, что, чем меньше плотность данных в области, тем больше в ней вариация (Рис 19).

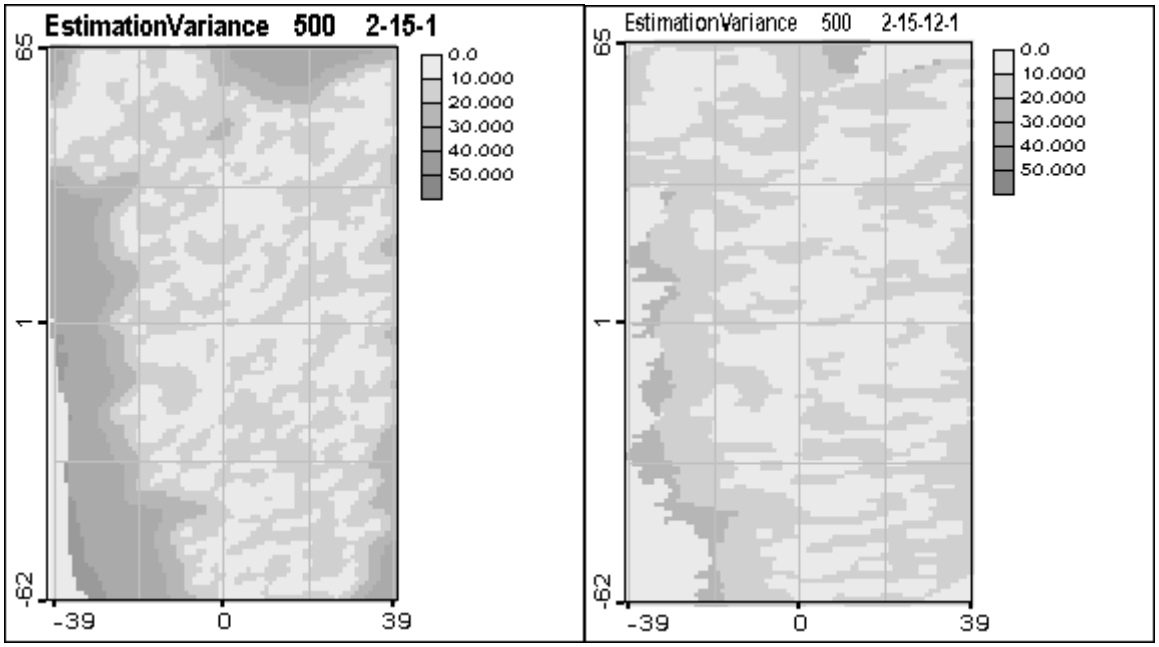
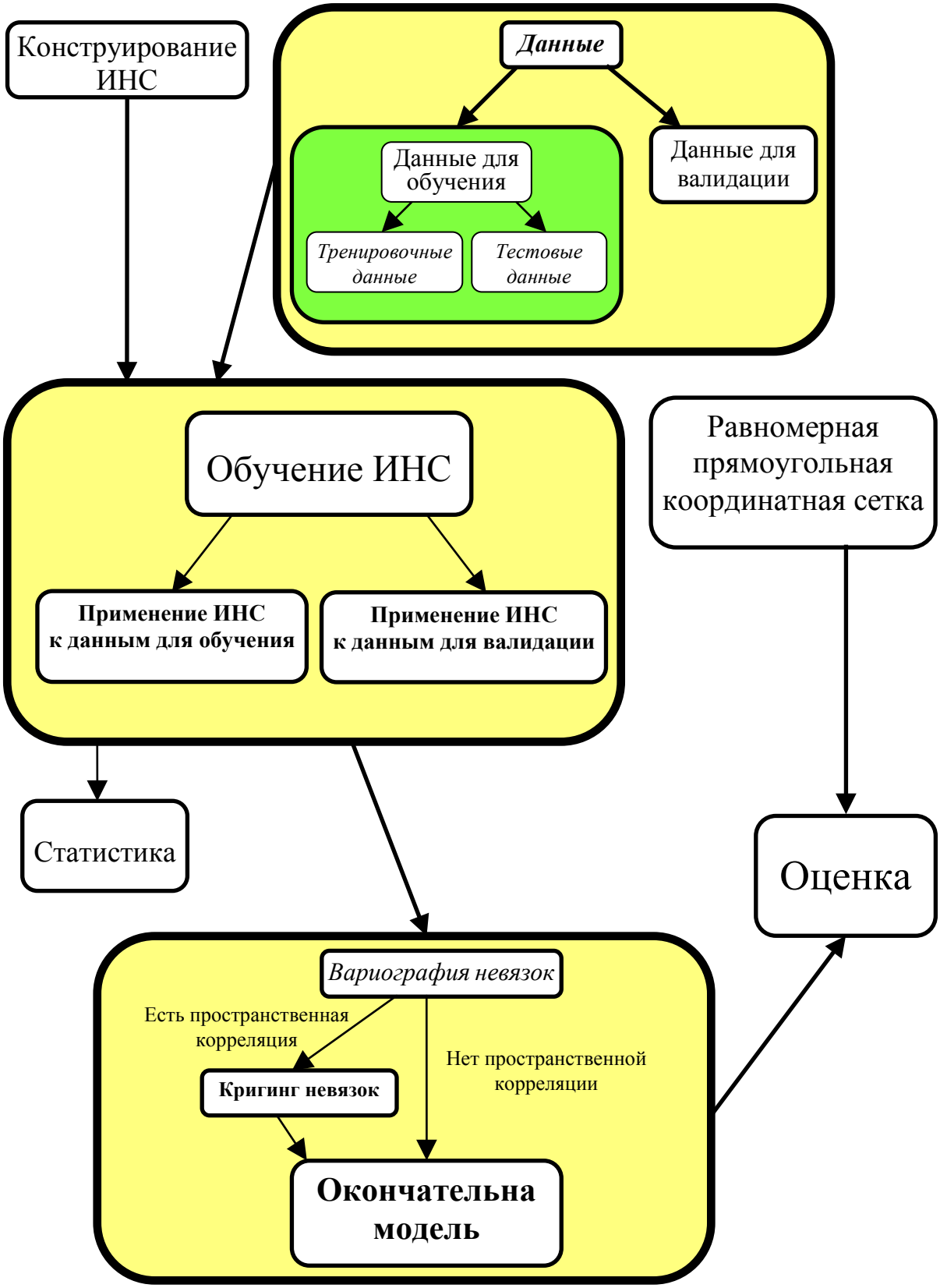


Рис 19: Вариация оценки кригинга невязок

7 Блок-схема метода картирования при помощи ИНС и геостатистики



8 Заключение

В работе рассматривался комбинированный метод моделирования пространственных данных при помощи ИНС и кригинга. Причем ИНС применялись для моделирование крупномасштабных структур. Кроме того анализировалась зависимость модели, построенной ИНС, от количества данных для обучения и конфигурации нейронной сети. Была рассмотрена возможность обработки невязок при помощи еще одного персептрона.

Выводы:

1. Для каждого набора данных существует ИНС, с минимальной RMSE модели. Нельзя, увеличивая количество нейронов до бесконечности, уменьшать RMSE.

2. Зависимость RMSE от количества нейронов в скрытых слоях ИНС наблюдается только на тренировочных данных, и не обнаруживается на валидационных.

3. Количество данных для обучения влияет на качество модели гораздо сильнее, чем конфигурация ИНС.

4. Модель построенная ИНС имеет крупномасштабную структуру. Невязки имеют мелкомасштабную структуру.

5. Кригинг невязок заметно улучшает модель, построенную ИНС.

6. При анализе невязок с помощью ИНС не удается уточнить модель.

7. Для исследуемых данных можно выделить класс ИНС, зависящий от соотношения количества нейронов и количества данных, при которых метод совместного анализа является устойчивым.

8. Ни при одной конфигурации ИНС не удалось достигнуть оверфиттинга.

Все выводы относительно ИНС сделаны на основе результатов, полученных при работе в среде NeuroShell2.

Благодарности

Работа выполнена при частичной поддержке гранта ИНТАС 96 – 1957 и гранта для молодых ученых “искусственные нейронные сети и генетические алгоритмы для анализа и моделирования пространственной информации по окружающей среде”.

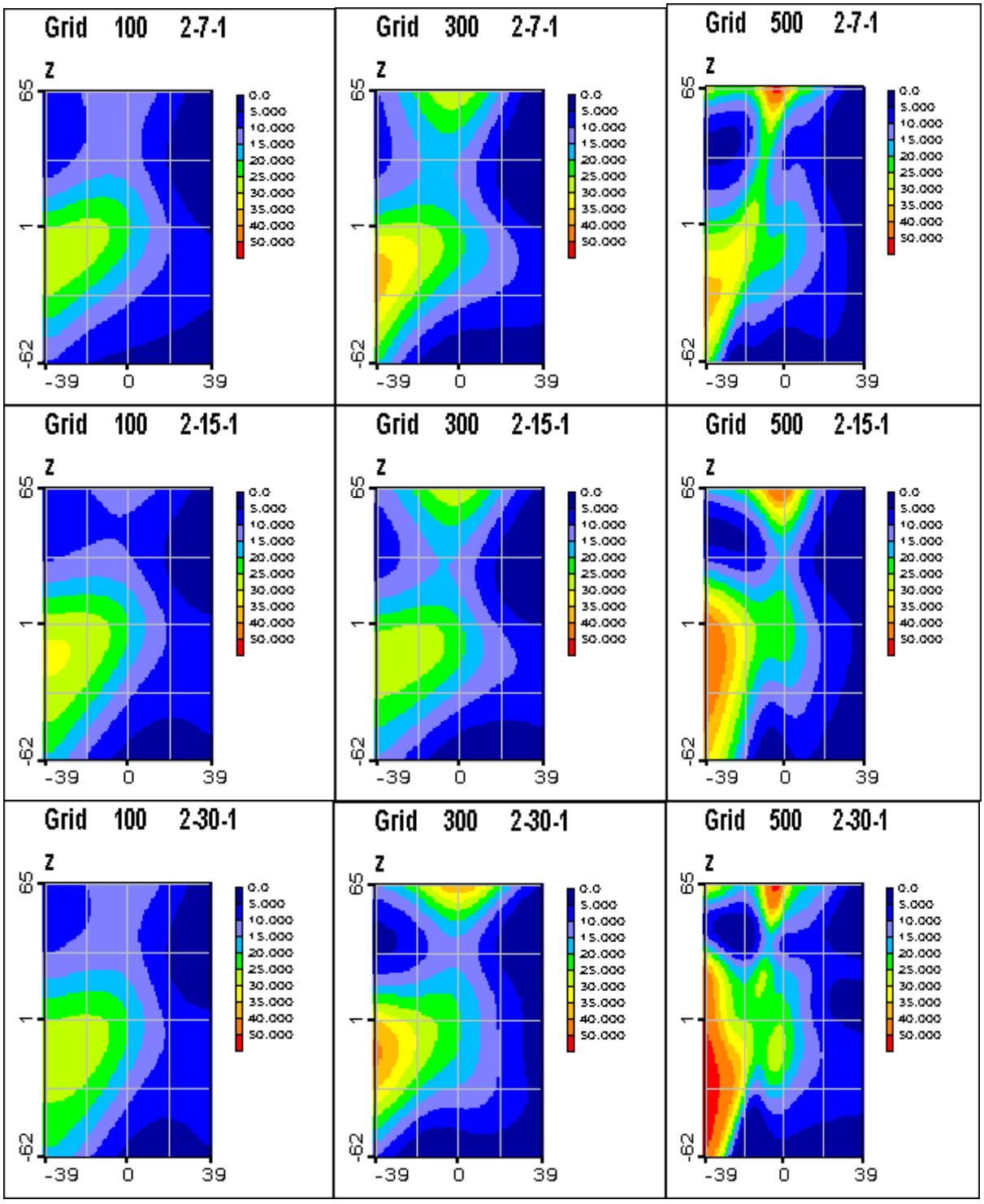
9 Список литературы

1. M. Kanevsky, R. Arutyunyan, L. Bolshov, V. Demyanov, M. Maignan. Artificial neural networks and spatial estimations of Chernobyl fallout. Geoinformatics, 1996, vol. 7, nos. 1-2., p. 5-11.
2. M. Kanevski M., Demyanov V., and Maignan M. Mapping of Soil Contamination by Using Artificial Neural Networks and Multivariate Geostatistics. Artificial Neural Networks ICANN '97. 7th International Conference, Lausanne, Switzerland, October 1997. Proceedings. W. Gerstner, A. Germond, M. Hasler, J.-D. Nicould (eds.). Lecture Notes in Computer Science, Springer, 1997, pp. 1125-1130.
3. M. Kanevski, M. F. Maignan, V. Demyanov and M. F. Maignan. Environmental decision-oriented mapping with algorithms imitating nature. IAMG'97 Proceedings of The Third Annual Conference of the International Association of Mathematical Geology. Ed. V. Pawlowsky Glan. Barcelona, Spain, CIMNE, 1997, ISBN 84-87867-97-9, vol. 2, p.520.
4. M. Kanevski, V. Demyanov and M. Maignan. Spatial estimations and simulations of environmental data by using geostatistics and artificial neural networks. IAMG'97 Proceedings of The Third Annual Conference of the International Association of Mathematical Geology. Ed. V. Pawlowsky Glan. Barcelona, Spain, CIMNE, 1997, ISBN 84-87867-97-9, vol. 2, pp.533-538.

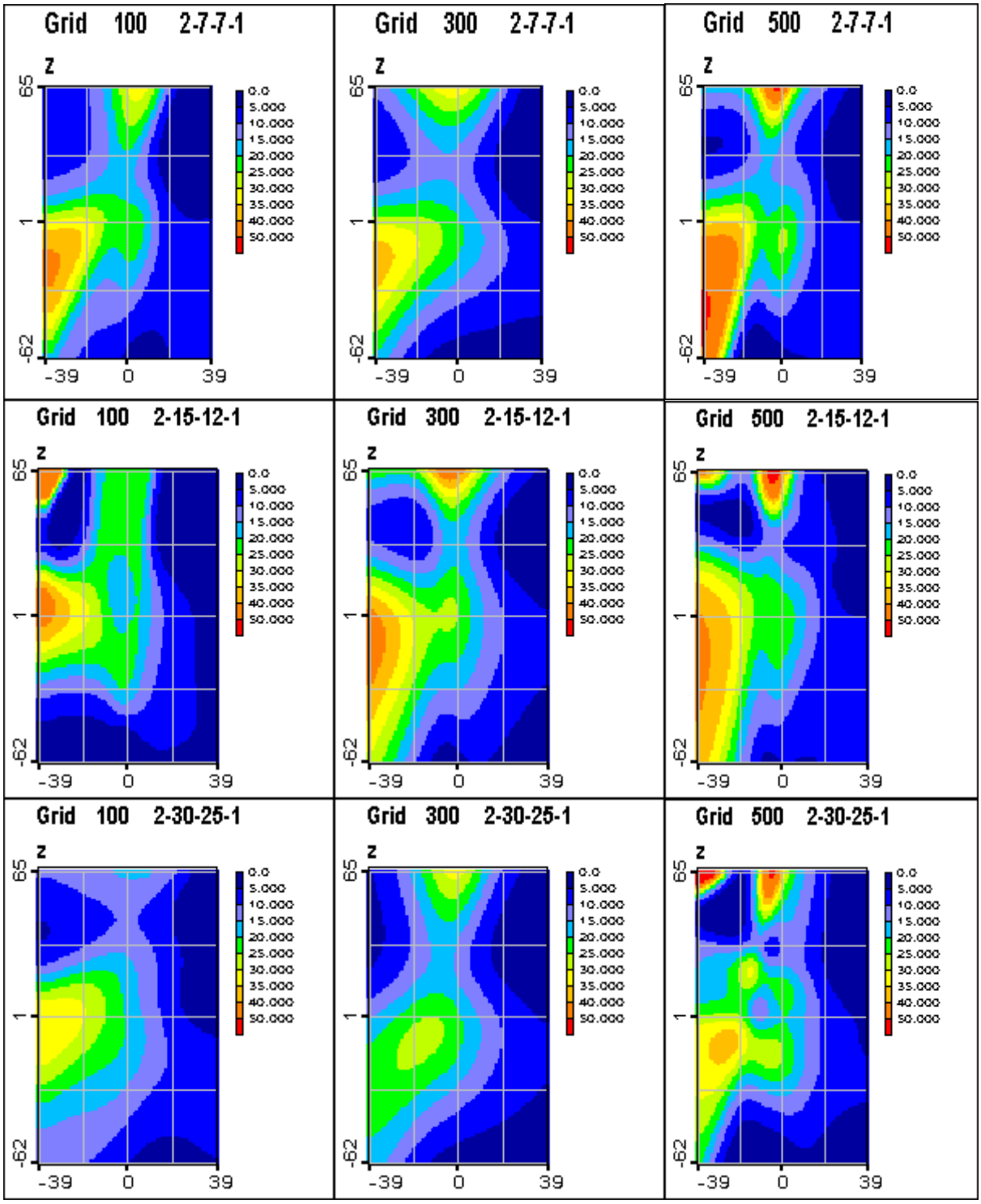
5. M. Kanevsky, R. Arutyunyan, L. Bolshov, S. Chernov, V. Demyanov, I. Linge, N. Koptelova, E. Savelieva, T. Haas, M. Maignan. Chernobyl Fallouts: Review of Advanced Spatial Data Analysis, First European Conference on Geostatistics for Environment Applications geoENV 96, Lisbon, Portugal, 20-22 November 1996, also in geoENV I — Geostatistics for Environmental Applications, ed. A. Soares, J. Gomez-Hernandes, R. Froidvaux, Kluwer Academic Publishers, 1997, pp. 389-400.
6. Christopher M. Bishop. Neural Network for Pattern Recognition. Department of Computer Science and Applied Mathematics Aston University, Birmingham, UK. Clarendon Press, Oxford.
7. M. Kanevsky M., Arutyunyan R., Bolshov L., Demyanov V., Linge I., Savelieva E.,. Shershakov V., Haas T., Maignan M. Geostatistical Portrayal of the Chernobyl Fallout. Geostatistics Wollongong '96, ed. E.Y. Baafi, N.A. Schofield, Kluwer Academic Publishers, 1996, volume 2, pp.1043-1054.
8. M. Kanevski, V. Demyanov, S. Chernov, E. Savelieva, V. Timonin. *Environmental Spatial Data Analysis with GEOSTAT OFFICE Software*. IAMG '98 the 1998 Annual Conference of the International Association for Mathematical Geology Ischia Island, Italy, 5-9 October 1998, ed. A. Buccianti, G. Nardi, R. Potenza, De Frede Editore Napoli, 1998, pp.161-166.
9. S. Chernov, V. Demyanov, M. Kanevski, E. Savelieva. VarRose- a Way of Variogram Analysis. Preprint IBRAE-98-03. Moscow 1998, 27 p. In English.
10. M. Kanevski, S. Chernov, V. Demyanov. 3Plot Software: Advanced Spatial Data Plot. Preprint IBRAE 97-02, Moscow, 1997, 32 p. In English.
11. Isaaks E.H., Shrivastava R.M. An Introduction to Applied Geostatistics. Oxford University press, Oxford, 1989.
12. Clayton V. Deutch, Andre G. Journel. Geostatistical Software Library and User's Guide. New York, Oxford, Oxford University Press, 1998.

10 Приложения

1 Карты оценок ИНС различной архитектуры с одним скрытым слоем на равномерной сетке, CS137, западная часть Брянской обл.



2 Карты оценок ИНС различной архитектуры с двумя скрытыми слоями на равномерной сетке, CS137, западная часть Брянской обл.



3 Статистика невязок ИНС на тренировочных данных

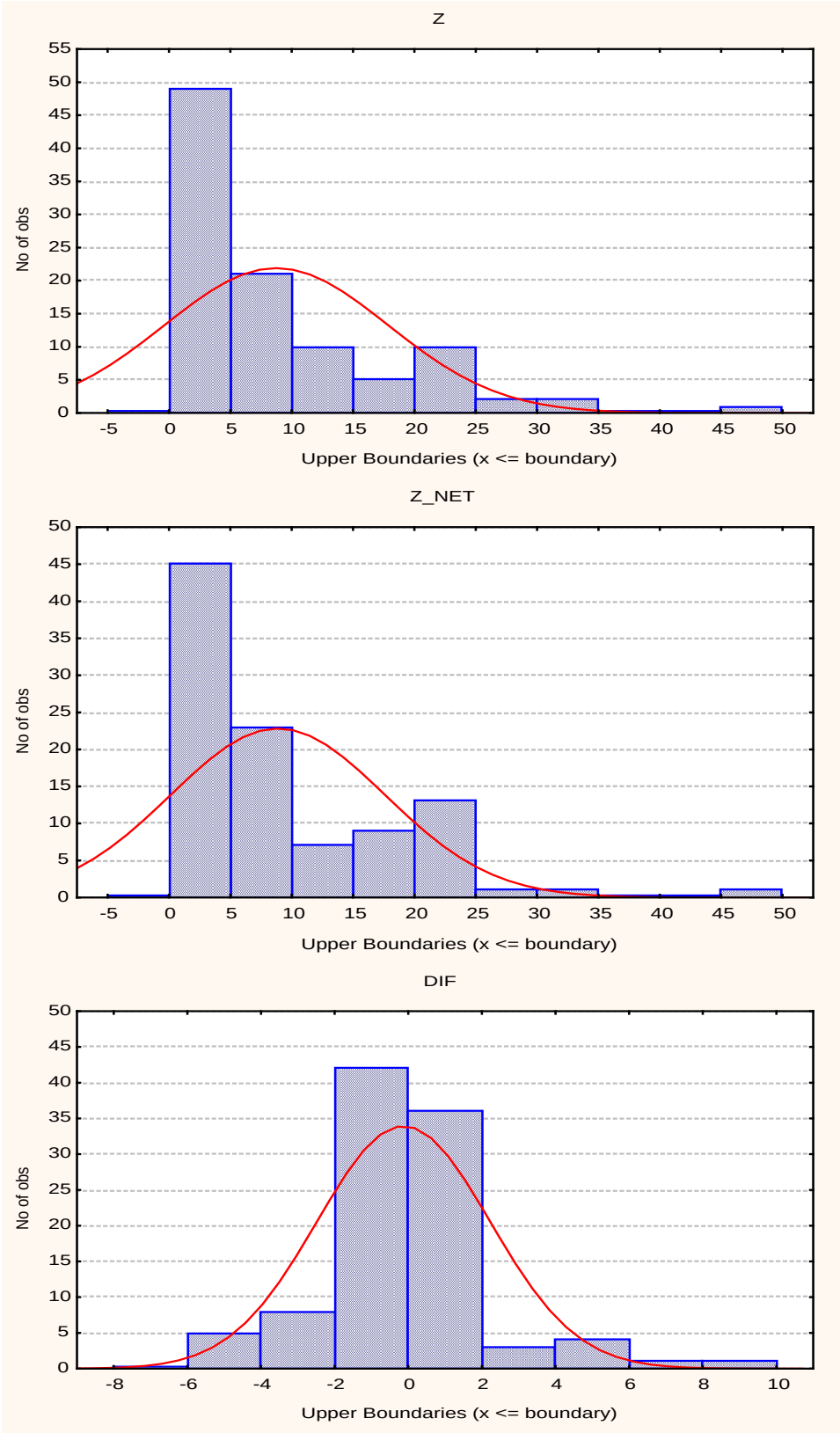
ИНС	Количество	Ср.	Медиана	Минимум	Максимум	Ниж.	Верх.	Размах	Размах	RMSE
	Данных	Знач.				Квартиль	Квартиль		Квартили	
2-7-1	100	-0.47	-1.02	-8	40	-3.05	0.42	47.9	3.47	5.7
2-15-1	100	-0.49	-1.16	-10	39	-2.98	0.46	49.4	3.44	5.8
2-30-1	100	-0.78	-1.35	-12	39	-3.16	0.03	50.9	3.19	5.9
2-7-7-1	100	-0.46	-0.85	-8	40	-2.98	0.51	48.4	3.49	5.3
2-15-12-1	100	**-0.13**	-0.17	-6	9	-1.21	0.66	14.8	1.88	**2.4**
2-30-25-1	100	-0.66	-1.24	-9	37	-3.06	0.19	46.7	3.25	5.3
2-7-1	300	-0.61	-1.24	-16	56	-4.08	1.02	72.3	5.10	7.9
2-15-1	300	-0.40	-1.10	-16	55	-3.77	1.32	70.7	5.09	7.9
2-30-1	300	**-0.13**	-1.10	-20	54	-3.21	1.11	73.6	4.32	7.5
2-7-7-1	300	-0.53	-1.41	-18	55	-3.79	0.86	72.4	4.65	7.7
2-15-12-1	300	-1.13	-1.97	-20	50	-4.27	0.24	70.2	4.51	**7.4**
2-30-25-1	300	0.27	-0.93	-13	62	-3.18	1.42	75.2	4.61	***8.0***
2-7-1	500	0.44	-0.09	-24	48	-2.33	2.14	71.3	4.47	6.9
2-15-1	500	**0.09**	-0.90	-22	53	-3.01	1.59	74.6	4.60	7.6
2-30-1	500	***-0.68***	-1.34	-23	46	-3.41	1.12	68.6	4.53	6.5
2-7-7-1	500	-0.53	-1.37	-23	54	-3.38	0.97	77.6	4.36	7.2
2-15-12-1	500	-0.58	-1.07	-24	38	-3.14	1.16	62.1	4.29	6.0

4 Корреляция измерений (Z) с значениями оценок (Z_NET) и невязок (DIF) на тренировочных данных

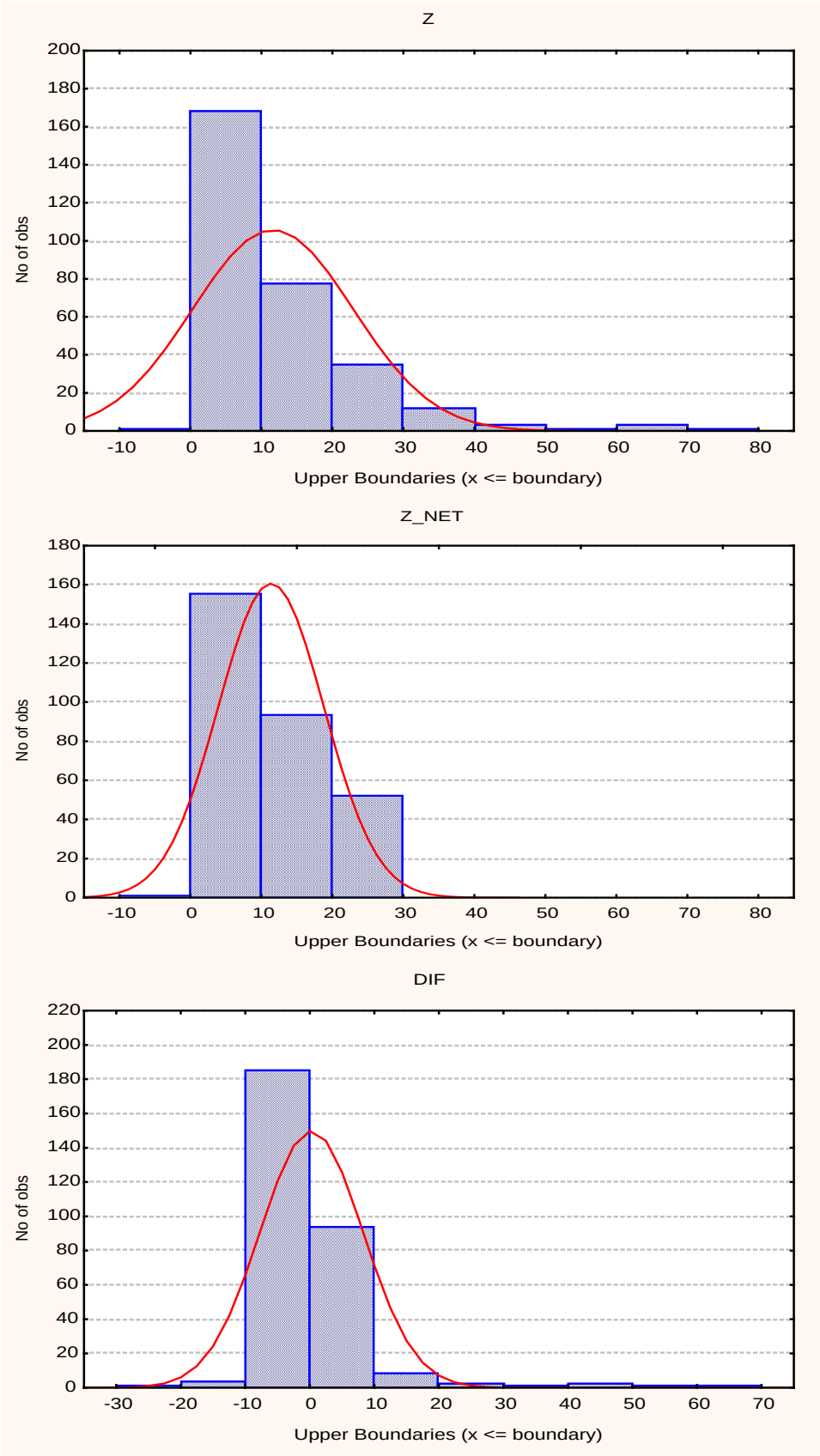
ИНС	Кол-во трен.дан.	Z_NET	DIF
2-7-1	100	0.78	0.64
2-15-1	100	0.77	0.62
2-30-1	100	0.77	0.63
2-7-7-1	100	0.82	0.57
2-15-12-1	100	0.97	0.28
2-30-25-1	100	0.81	0.63
2-7-1	300	0.72	0.73
2-15-1	300	0.72	0.75
2-30-1	300	0.75	0.73
2-7-7-1	300	0.73	0.74
2-15-12-1	300	0.76	0.69
2-30-25-1	300	0.71	0.75
2-7-1	500	0.81	0.71
2-15-1	500	0.75	0.73
2-30-1	500	0.82	0.64
2-7-7-1	500	0.78	0.71
2-15-12-1	500	0.85	0.62
2-30-25-1	500	0.88	0.58

5 Гистограммы исходных данных (Z), оценок ИНС (Z_NET) и невязок (DIF) в точках для тренировки

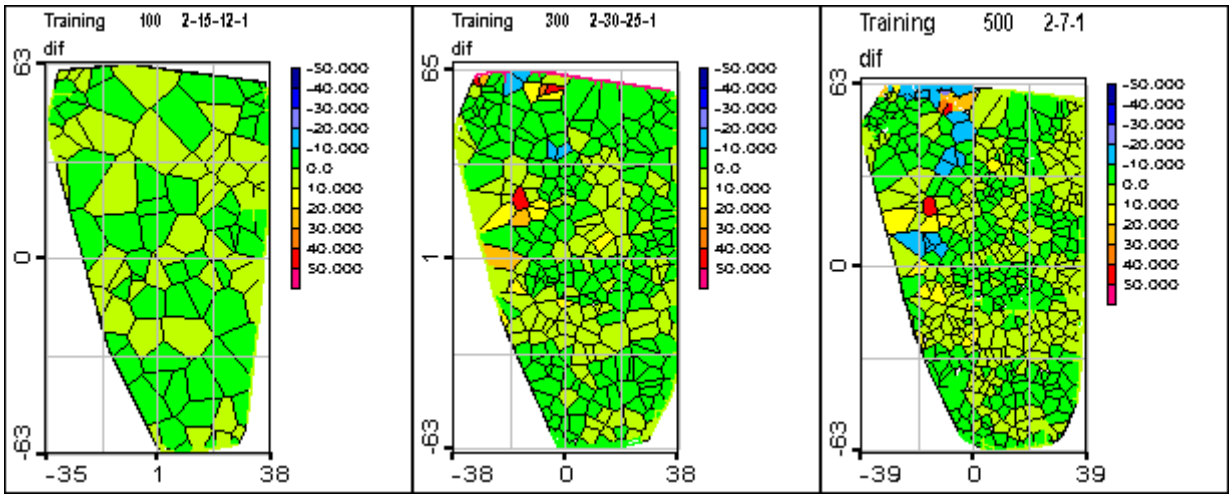
5.1 Набор 100 тренировочных данных, ИНС 2-15-12-1 (самая маленькая RMSE)



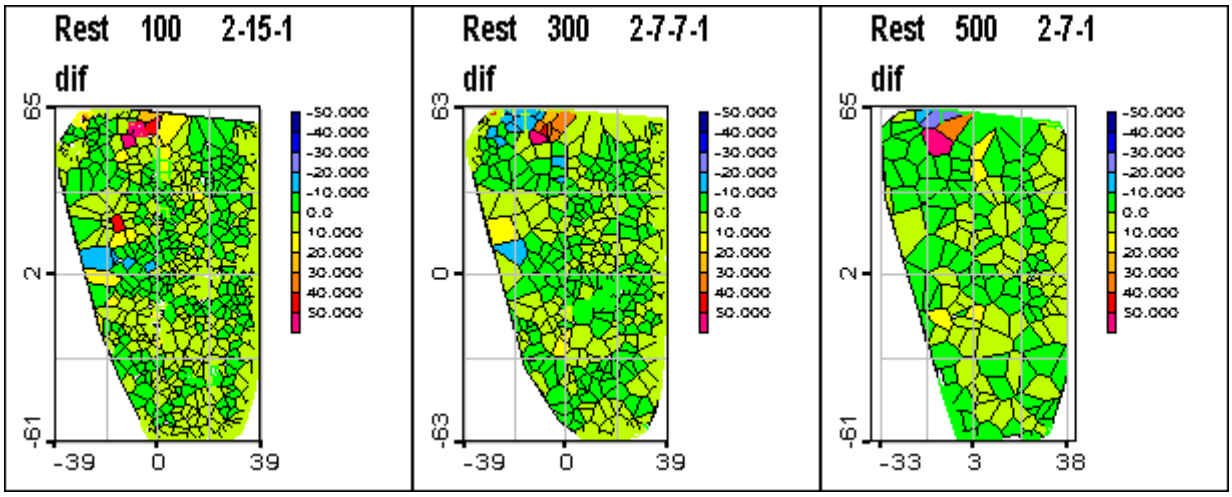
5.2 Набор 300 тренировочных данных, ИНС 2-30-25-1 (большая RMSE)



6 Примеры распределения невязок ИНС на тренировочных данных в виде полигонов Вороного

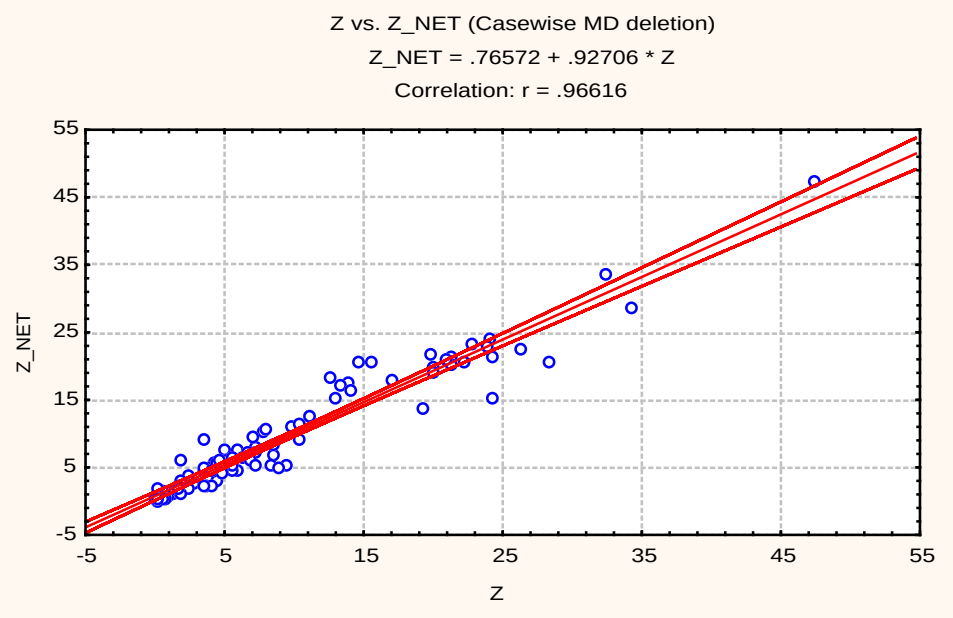


Примеры распределения невязок для валидационных данных в виде полигонов Вороного

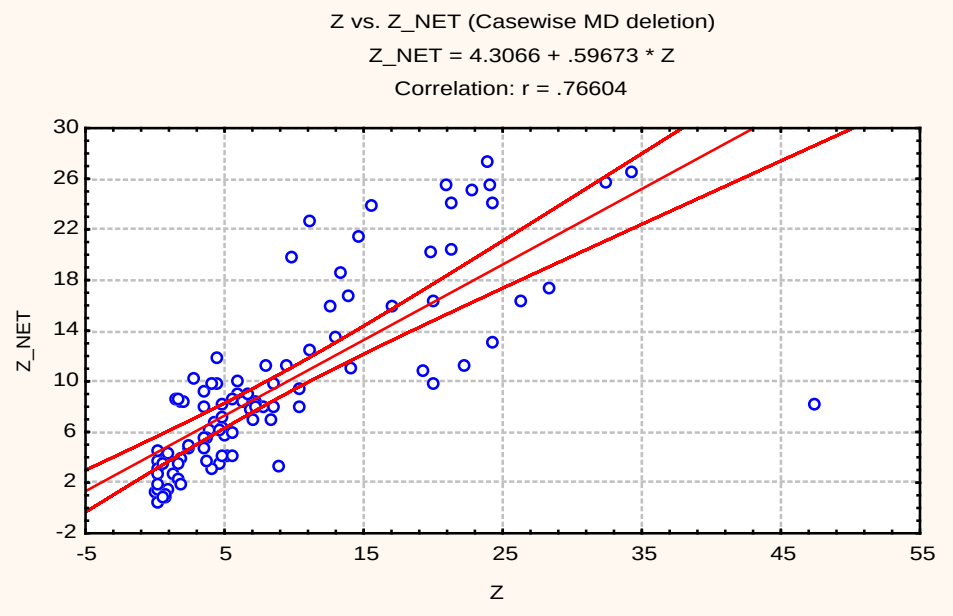


7 Зависимость оценки ИНС (Z_NET) от измерений (Z) в тренировочных точках

7.1 Набор 100 тренировочных данных, ИНС 2-15-12-1



7.2 Набор 100 тренировочных данных, ИНС 2-30-1



8 Статистика невязок ИНС в валидационных точках

ИНС	Кол. трен.дан.	Кол. вал.дан.	Среднее	Медиана	Минимум	Максимум	RMSE
2-7-1	100	465	***1.15***	-0.76	-12.8	83.1	9.4
2-15-1	100	465	***1.32***	-0.62	-13.1	84.5	***9.5***
2-30-1	100	465	0.85	-1.01	-12.2	83.7	***9.5***
2-7-7-1	100	465	0.41	-0.53	-16.0	80.6	9.2
2-15-12-1	100	465	-0.32	-0.17	-36.4	74.4	**8.8**
2-30-25-1	100	465	0.78	-1.04	-17.8	84.0	9.4
2-7-1	300	265	-0.64	-1.07	-16.1	72.7	8.3
2-15-1	300	265	-0.45	-1.19	-16.1	73.9	8.3
2-30-1	300	265	**-0.20**	-0.95	-20.4	76.0	8.1
2-7-7-1	300	265	-0.52	-1.35	-19.0	72.9	8.1
2-15-12-1	300	265	-1.12	-1.46	-22.3	76.8	**8.0**
2-30-25-1	300	265	0.37	-0.65	-15.1	74.1	8.1
2-7-1	500	65	**0.03**	-0.41	-20.8	72.6	**7.7**
2-15-1	500	65	-0.50	-1.07	-21.7	76.8	8.2
2-30-1	500	65	***-1.12***	-1.57	-18.4	75.7	7.9
2-7-7-1	500	65	-1.05	-1.65	-20.6	72.9	7.9
2-15-12-1	500	65	-1.06	-1.55	-23.9	77.8	7.9
2-30-25-1	500	65	-0.67	-0.90	-19.0	71.1	**7.1**

9 Корреляция исходных значений (Z) с оценкой ИНС (Z_Net) и невязками (Dif)

все данные (слева)

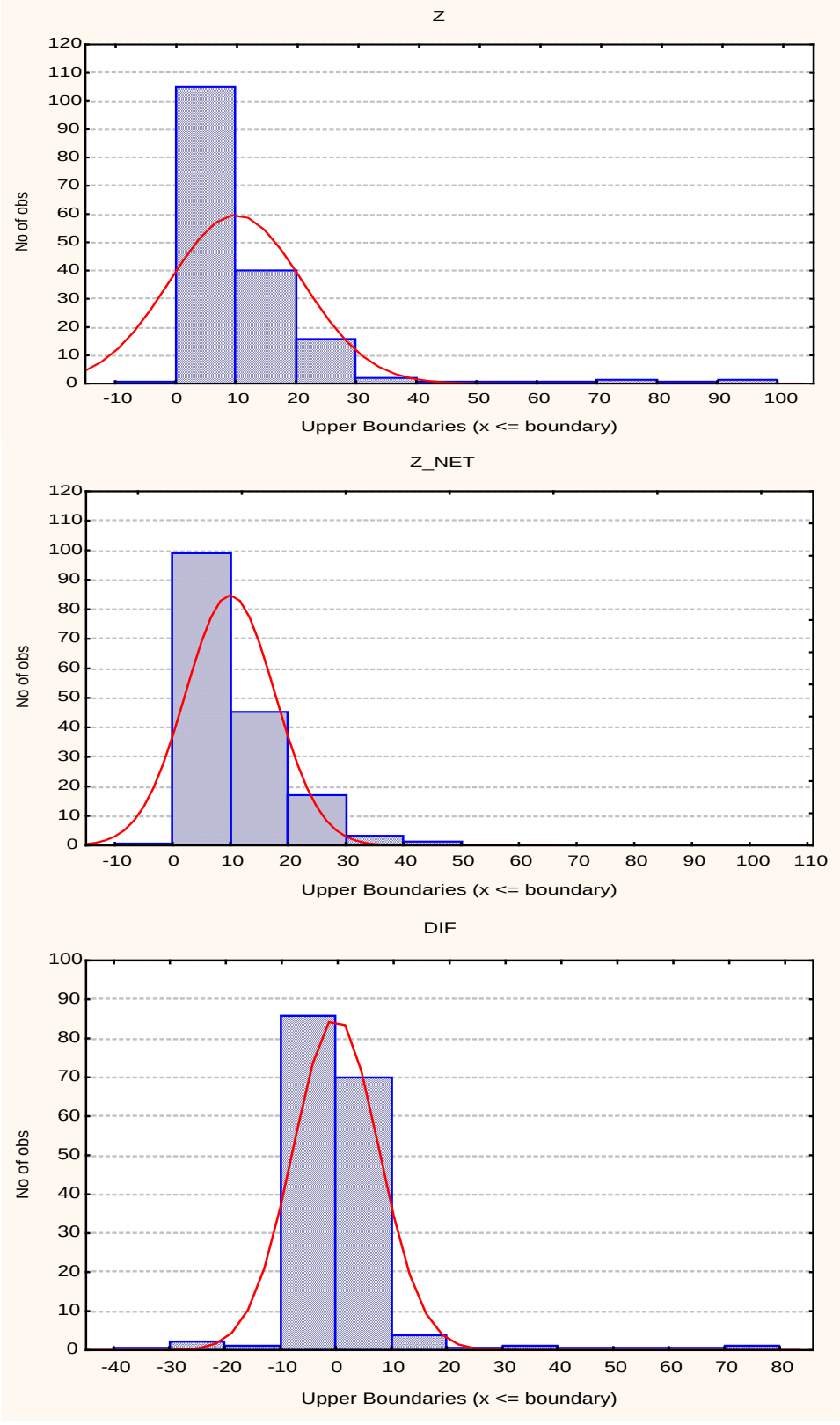
ИНС	Кол-во вал.дан.	Z-Net	Dif
2-7-1	100	0.59	0.85
2-15-1	100	0.57	0.84
2-30-1	100	0.58	0.84
2-7-7-1	100	0.62	0.78
2-15-12-1	100	0.66	0.66
2-30-25-1	100	0.59	0.84
2-7-1	300	0.69	0.76
2-15-1	300	0.69	0.78
2-30-1	300	0.71	0.77
2-7-7-1	300	0.71	0.77
2-15-12-1	300	0.71	0.73
2-30-25-1	300	0.70	0.78
2-7-1	500	0.72	0.71
2-15-1	500	0.67	0.70
2-30-1	500	0.70	0.67
2-7-7-1	500	0.70	0.70
2-15-12-1	500	0.70	0.69
2-30-25-1	500	0.77	0.64

без пяти точек с самым большим отклонением (справа)

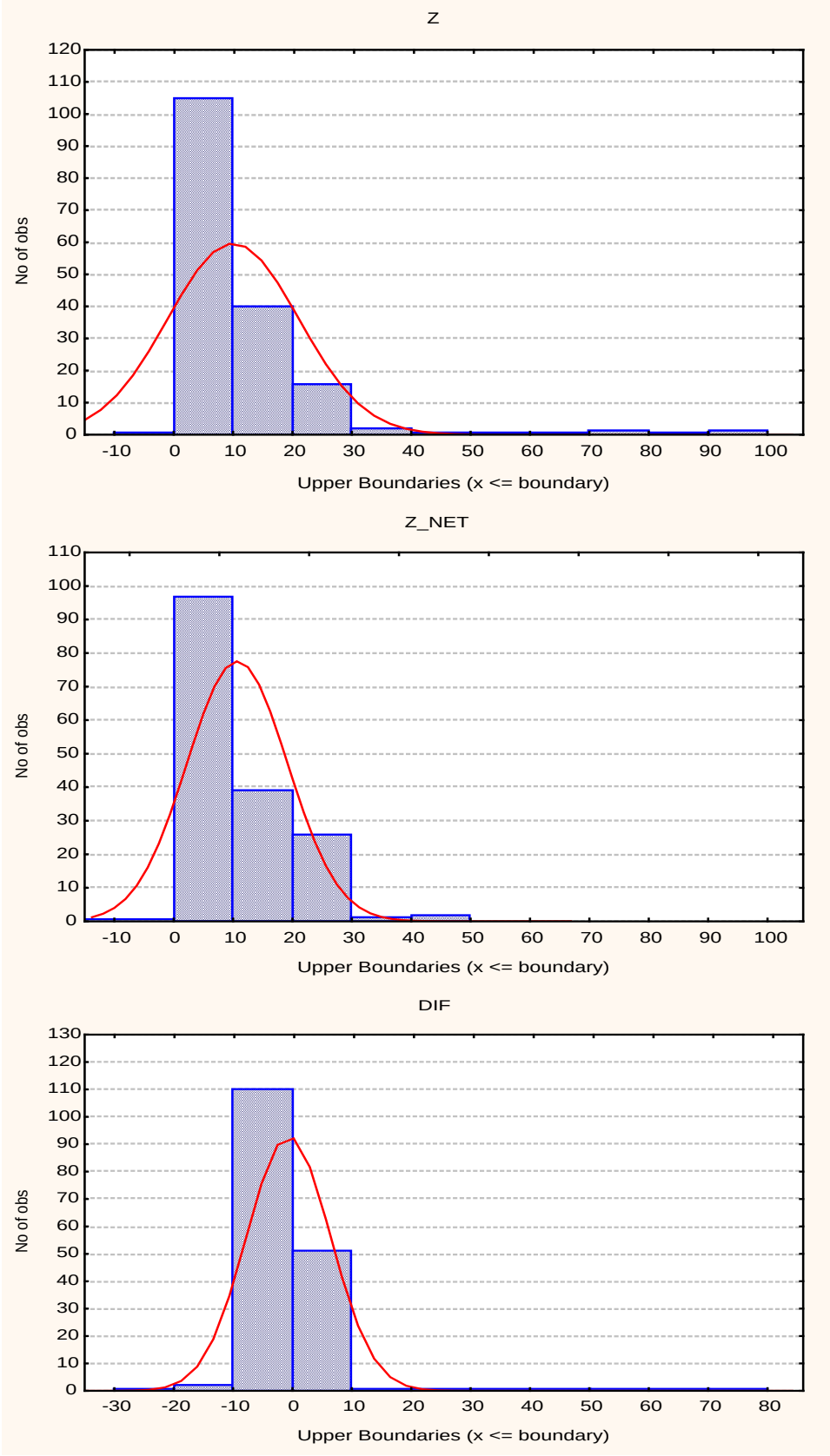
ИНС	Кол-во вал.дан.	Z-Net	Dif
2-7-1	100	0.69	0.78
2-15-1	100	0.68	0.77
2-30-1	100	0.68	0.78
2-7-7-1	100	0.71	0.69
2-15-12-1	100	0.72	0.52
2-30-25-1	100	0.69	0.77
2-7-1	300	0.77	0.57
2-15-1	300	0.77	0.60
2-30-1	300	0.79	0.58
2-7-7-1	300	0.78	0.59
2-15-12-1	300	0.79	0.50
2-30-25-1	300	0.79	0.61
2-7-1	500	0.84	0.18
2-15-1	500	0.84	0.10
2-30-1	500	0.85	0.09
2-7-7-1	500	0.83	0.12
2-15-12-1	500	0.85	0.11
2-30-25-1	500	0.90	0.02

10 Гистограммы исходных данных (Z), оценок ИНС (Z_NET), невязок (DIF) в точках для валидации

10.1 Набор 500 трен. данных, ИНС 2-7-1 (маленькая RMSE, самое маленькое среднее)

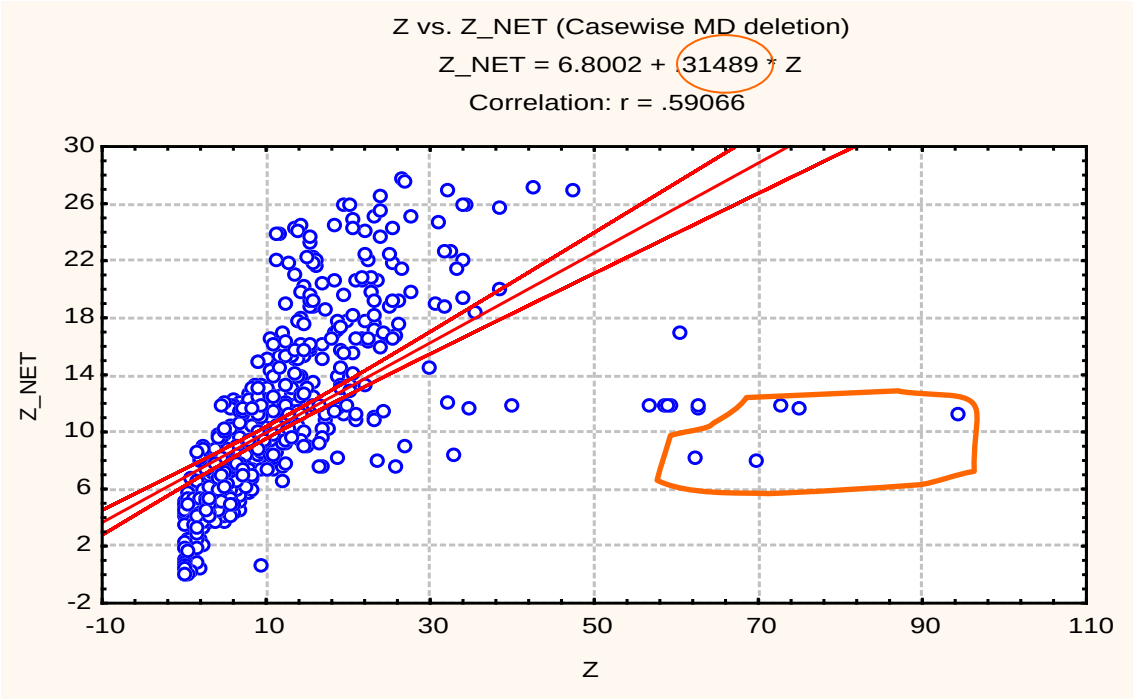


10.2 Набор 500 тренировочных данных на 2-30-25-1 (самая маленькая RMSE)

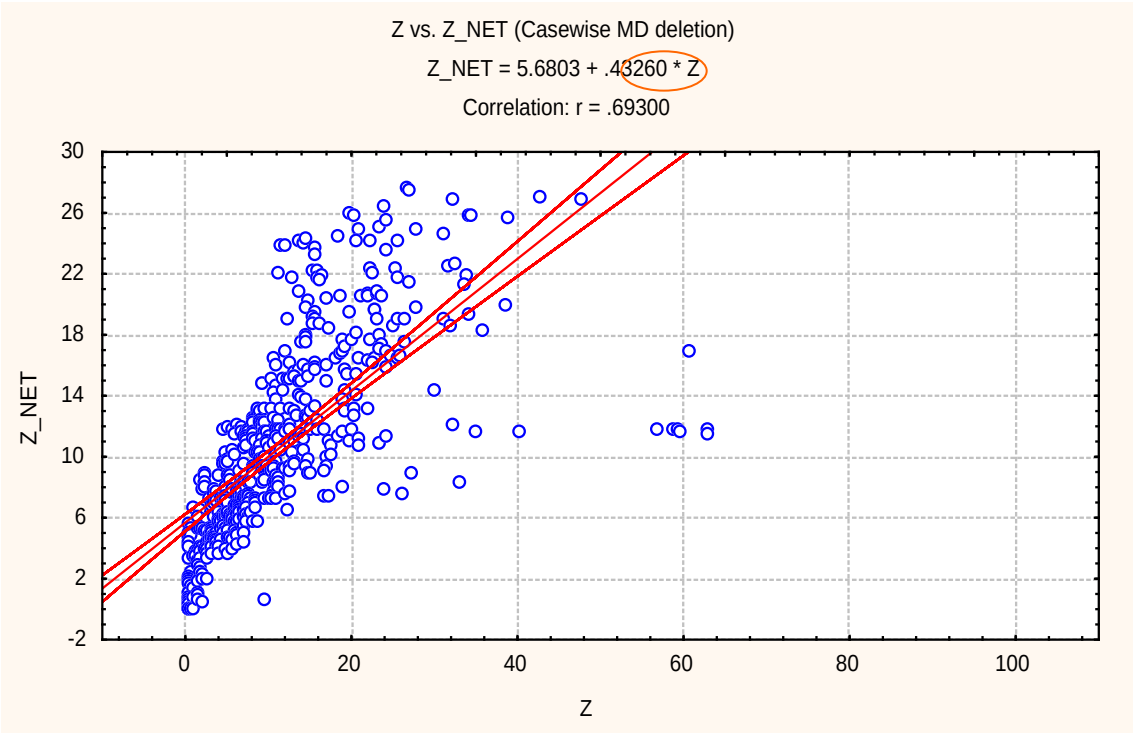


11 Зависимость оценки ИНС (Z_NET) от измерений (Z) в валидационных точках

11.1 Набор 100 тренировочных данных, ИНС 2-7-1

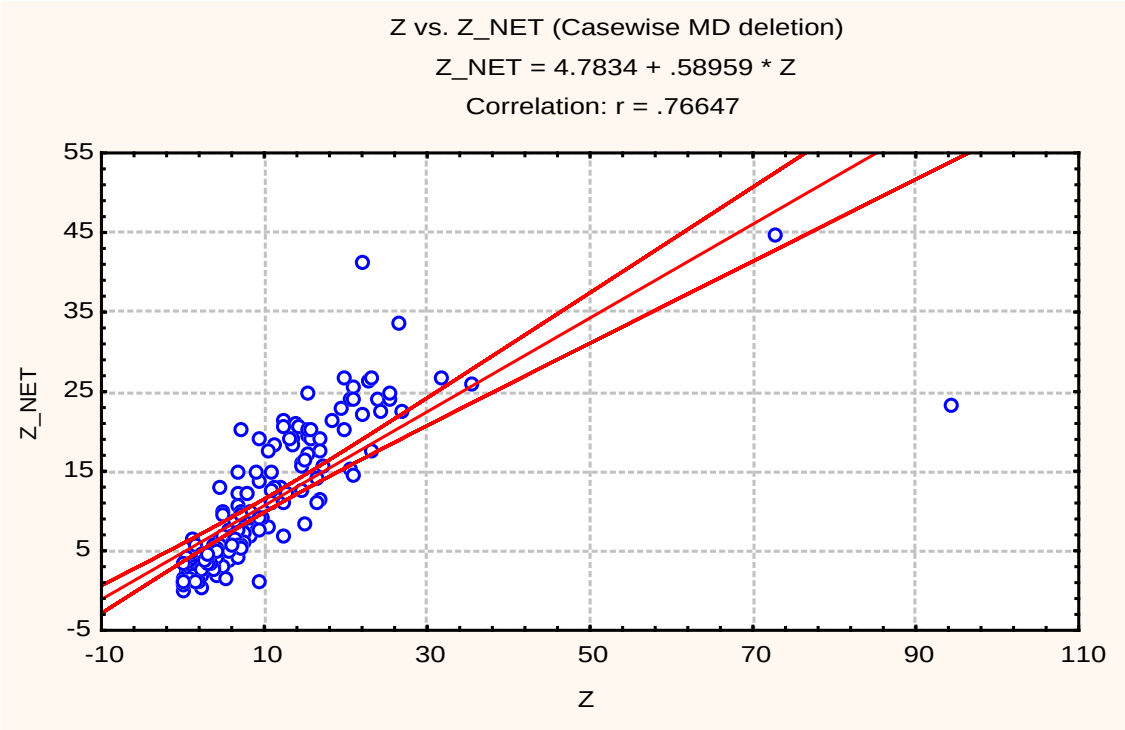


11.2 Набор 100 тренировочных данных, ИНС 2-7-1 без пяти точек с наибольшим отклонением

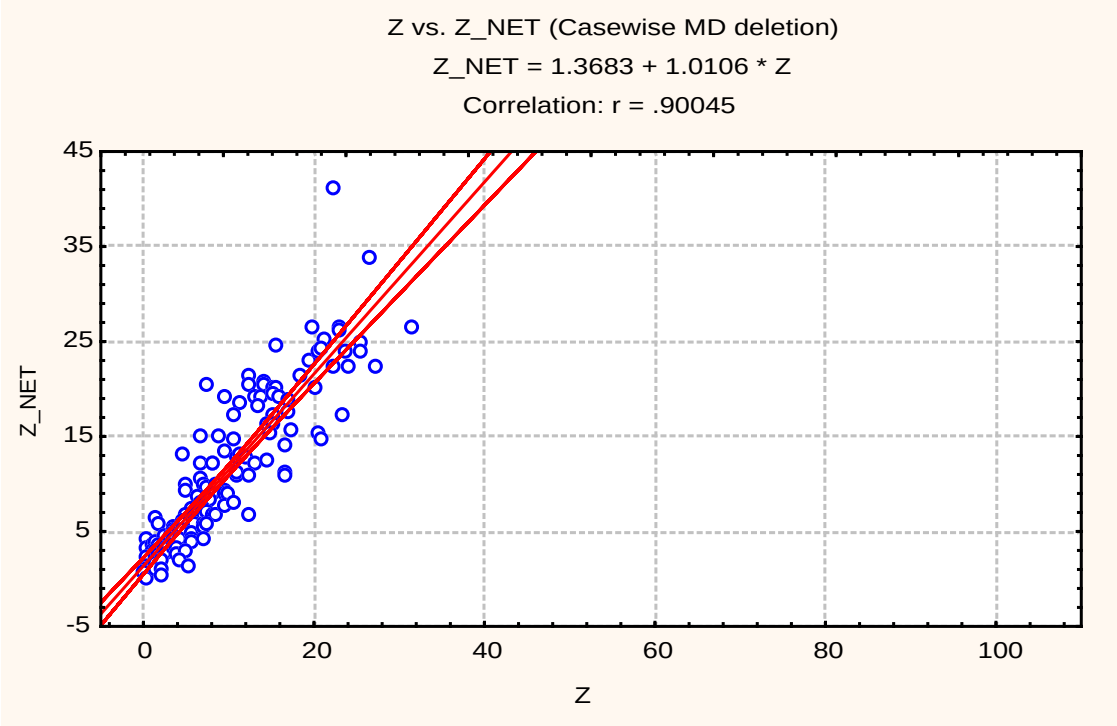


Обведенные точки на (12.1) были удалены. Результат удаления приведен на (12.2). После этого коэффициент корреляции существенно увеличился. Это говорит о том, что ИНС дает сильную ошибку в небольшом количестве точек. Коэффициент при линейном члене обведен в заголовке графиков.

12.3 Набор 500 тренировочных данных, ИНС 2-30-25-1



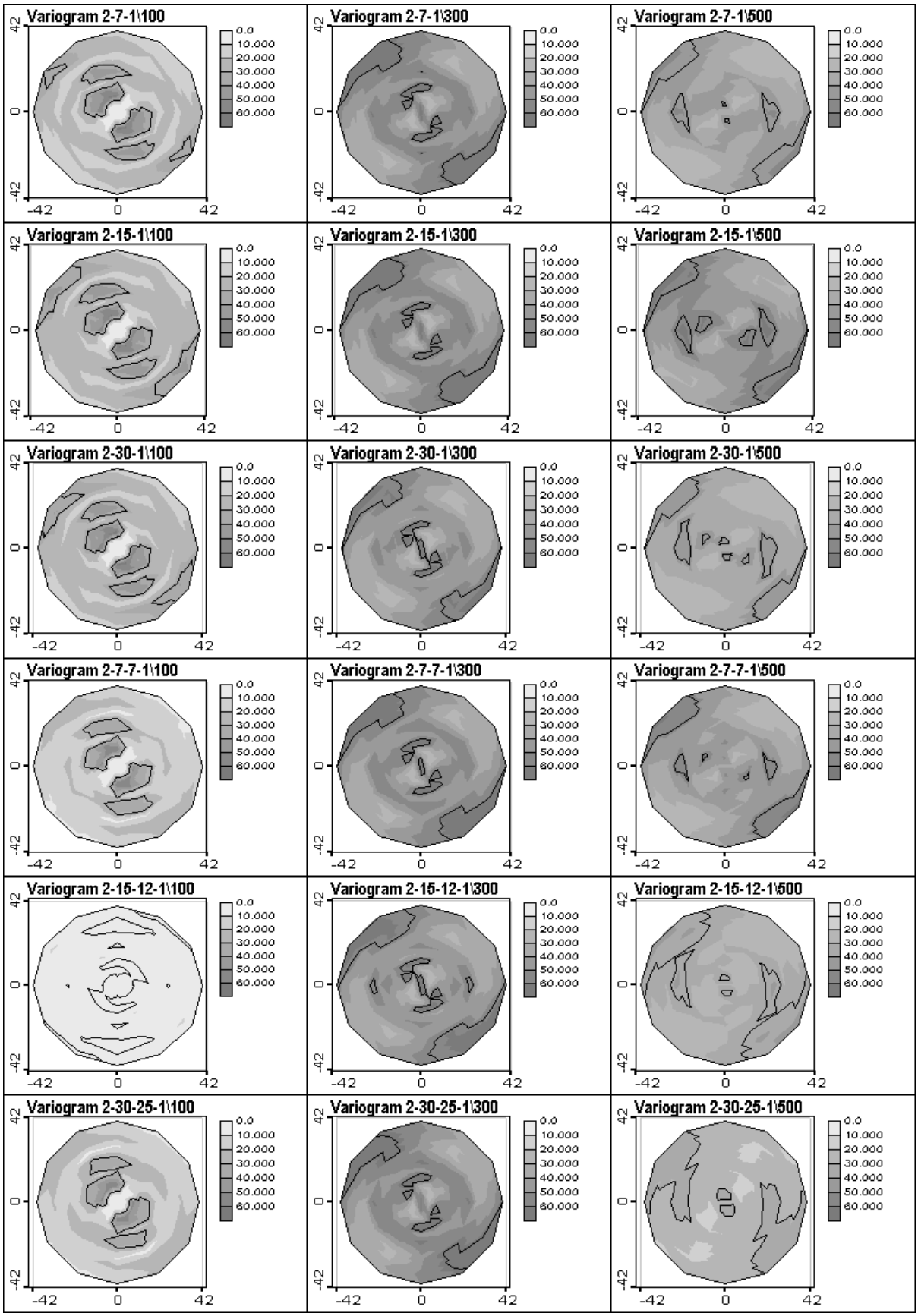
12.4 Набор 500 тренировочных данных, ИНС 2-30-25-1 без пяти точек с наибольшим отклонением



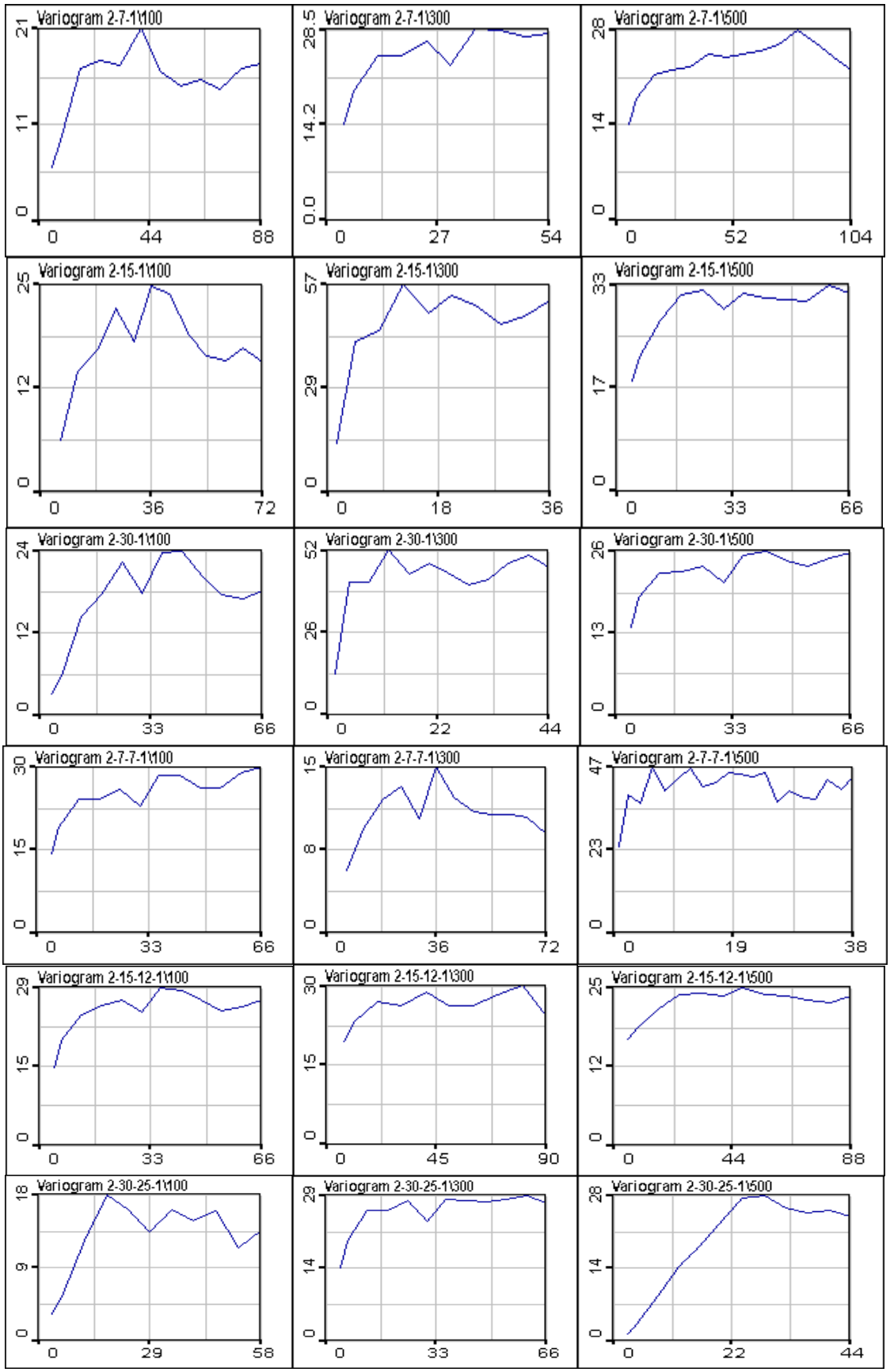
12 Сводная таблица статистики невязок ИНС

ИНС	Кол-во тр.дан.	Кол-во вал.дан.	Данные/Связи	Среднее вал.	RMSE вал.	Среднее трен.	RMSE трен.
2-7-1	100	465	7.14	***1.15***	9.4	-0.47	5.7
2-15-1	100	465	3.33	***1.32***	***9.5***	-0.49	5.8
2-30-1	100	465	1.67	0.85	***9.5***	-0.78	5.9
2-7-7-1	100	465	1.02	0.41	9.2	-0.46	5.3
2-15-12-1	100	465	0.28	**-0.32**	**8.8**	**-0.13**	**2.4**
2-30-25-1	100	465	0.07	0.78	9.4	-0.66	5.3
2-7-1	300	265	21.43	-0.64	8.3	-0.61	7.9
2-15-1	300	265	10.00	-0.45	8.3	-0.40	7.9
2-30-1	300	265	5.00	**-0.20**	8.1	**-0.13**	7.5
2-7-7-1	300	265	3.06	-0.52	8.1	-0.53	7.7
2-15-12-1	300	265	0.83	-1.12	**8.0**	-1.13	**7.4**
2-30-25-1	300	265	0.20	0.37	8.1	0.27	***8.0***
2-7-1	500	65	35.71	**0.03**	**7.7**	0.44	6.9
2-15-1	500	65	16.67	-0.50	8.2	**0.09**	7.6
2-30-1	500	65	8.33	***-1.12***	7.9	***-0.68***	6.5
2-7-7-1	500	65	5.10	-1.05	7.9	-0.53	7.2
2-15-12-1	500	65	1.39	-1.06	7.9	-0.58	6.0

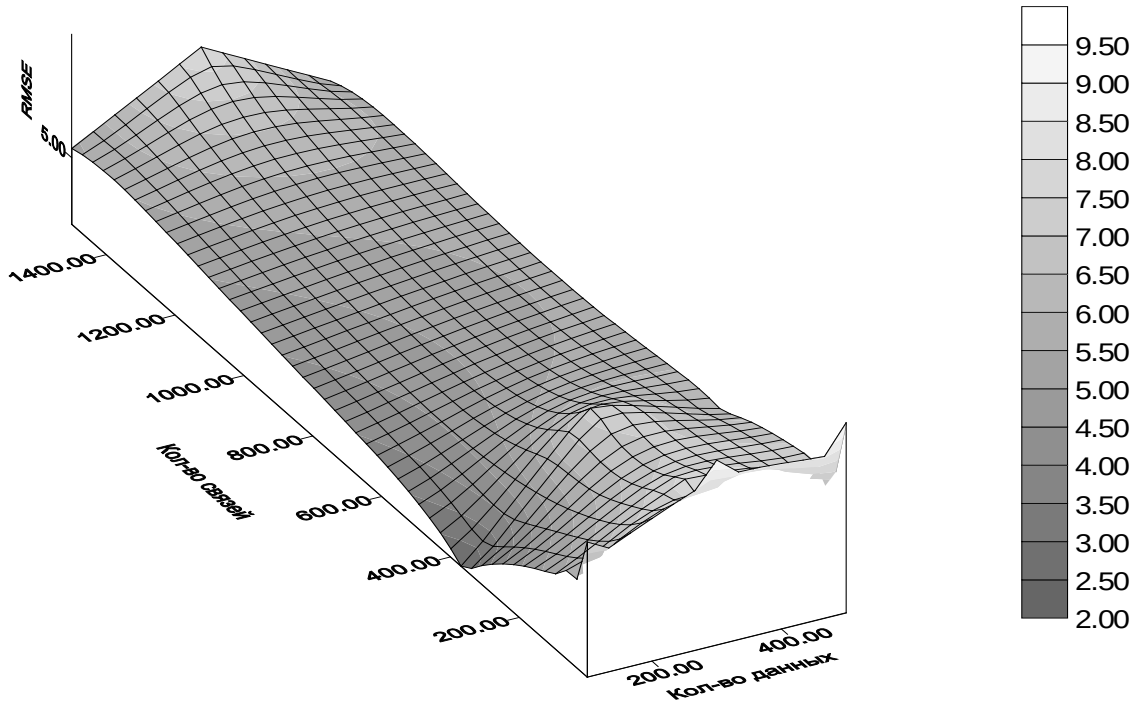
13 Вариограмные «розы» невязок ИНС



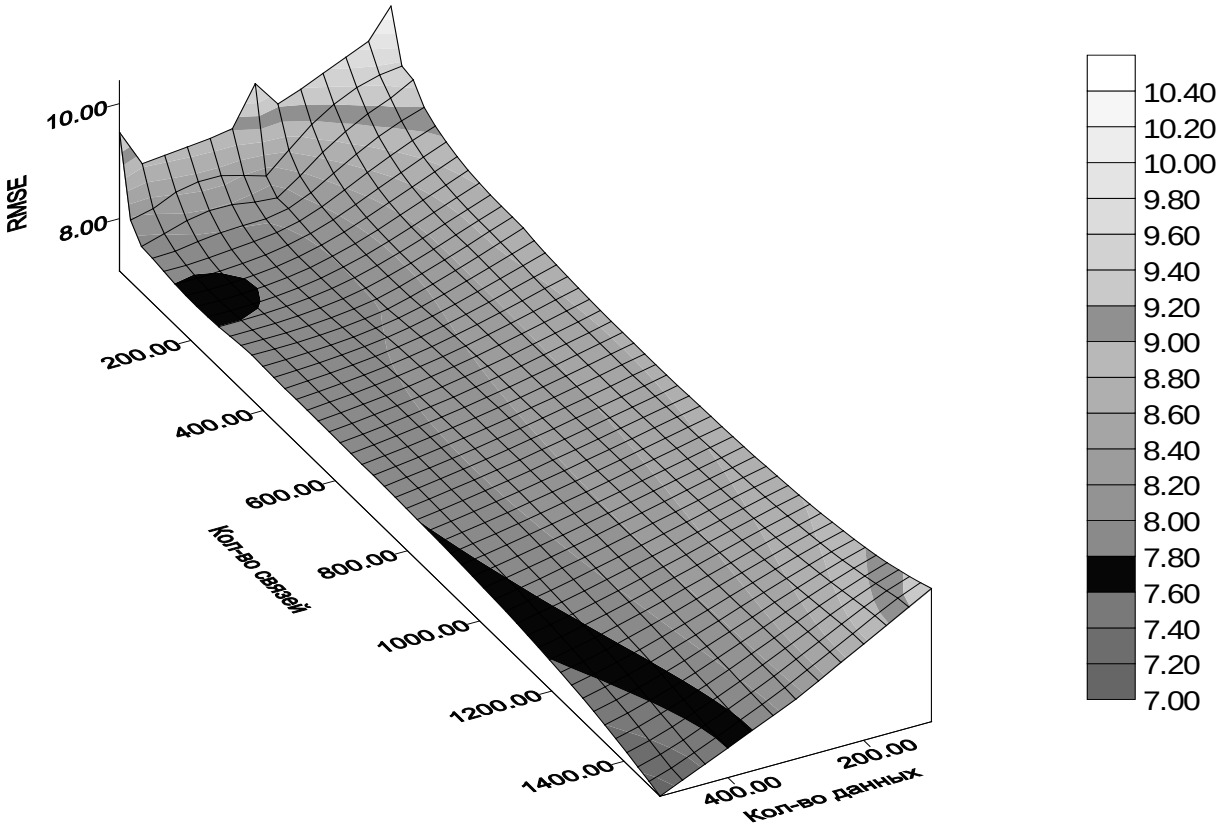
14 Вариограммы по всем направлениям для невязок ИНС



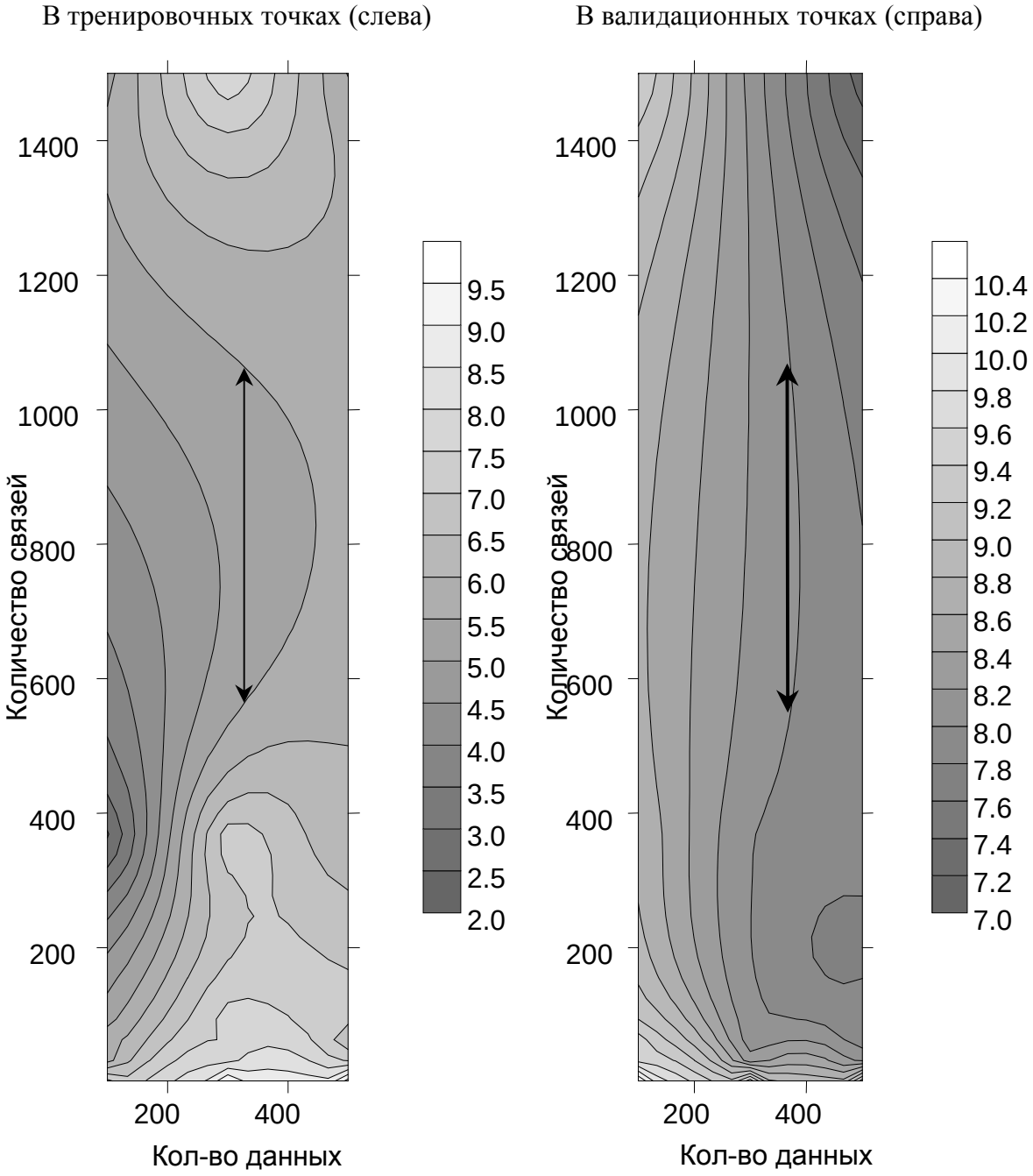
15 Распределение RMSE на тренировочных данных в зависимости от числа связей и числа данных для тренировки



16 Распределение RMSE на валидационных данных в зависимости от числа связей и числа данных для тренировки

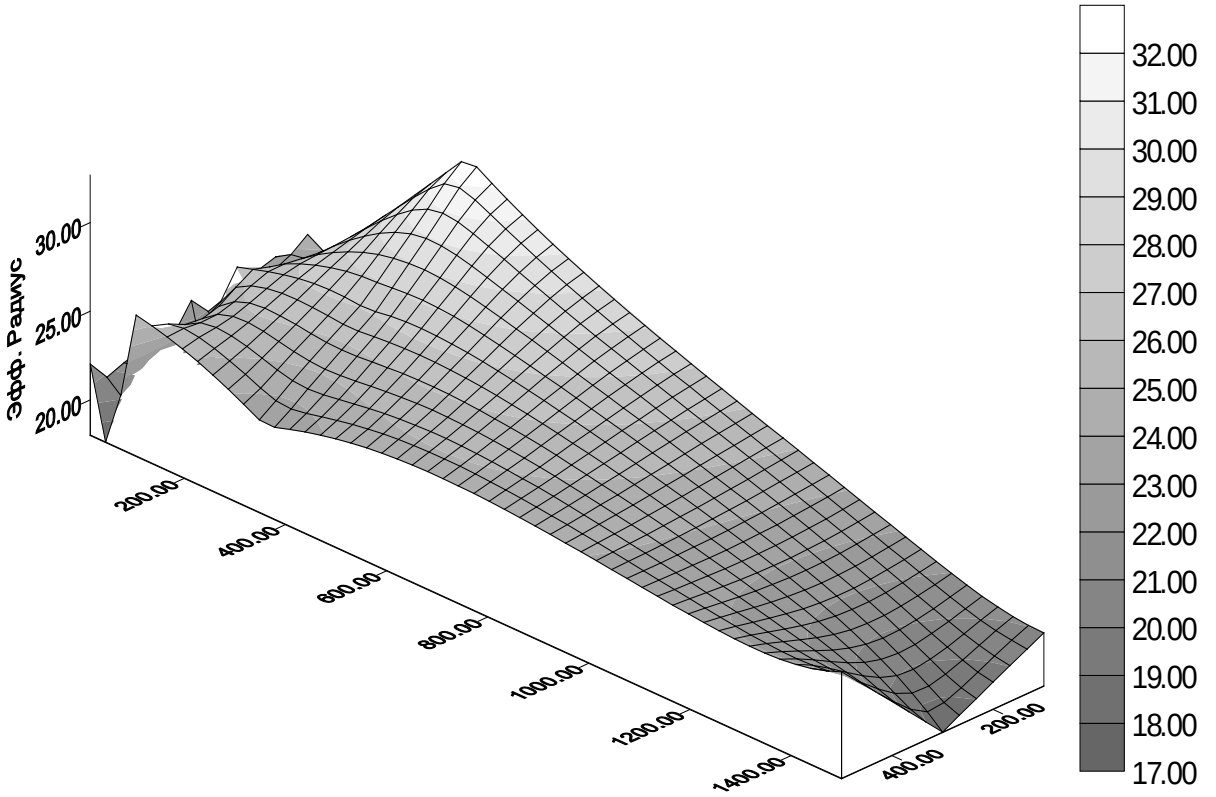


17 Распределение RMSE в зависимости от числа связей и количества данных для тренировки

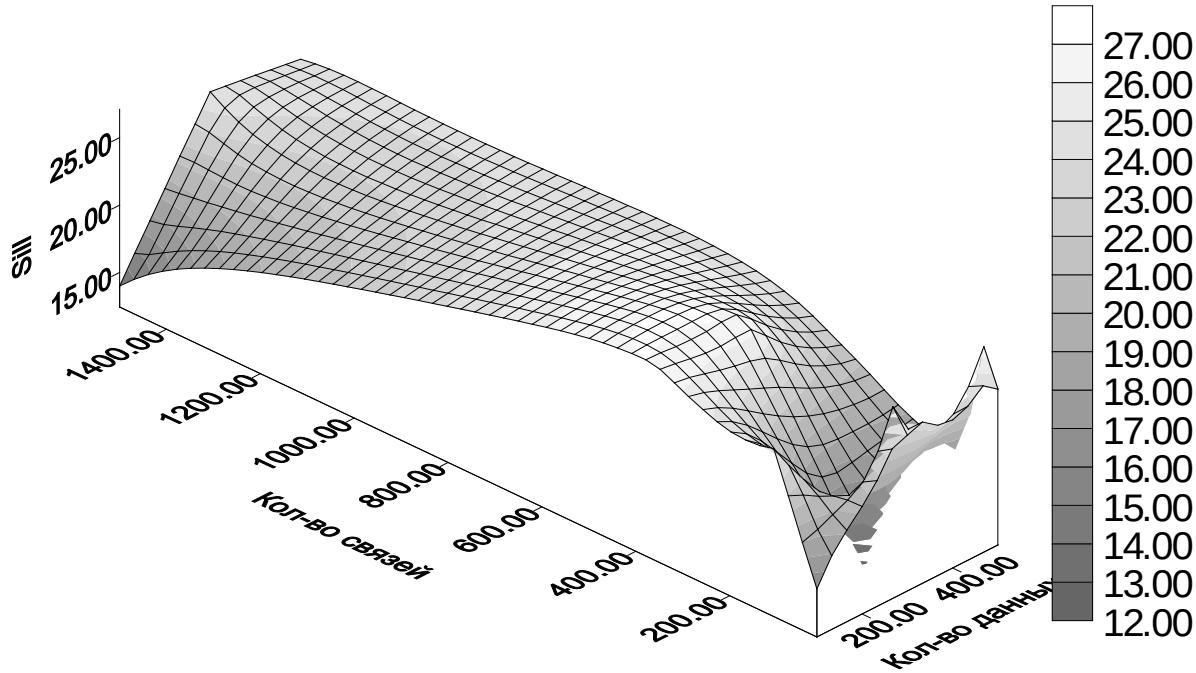


Для каждого набора тренировочных данных можно выделить диапазон, в которому должно принадлежать количество весов ИНС. На графиках пример такого диапазона обозначен стрелками. Использование ИНС с количеством весов вне этого диапазон ведет к увеличению RMSE.

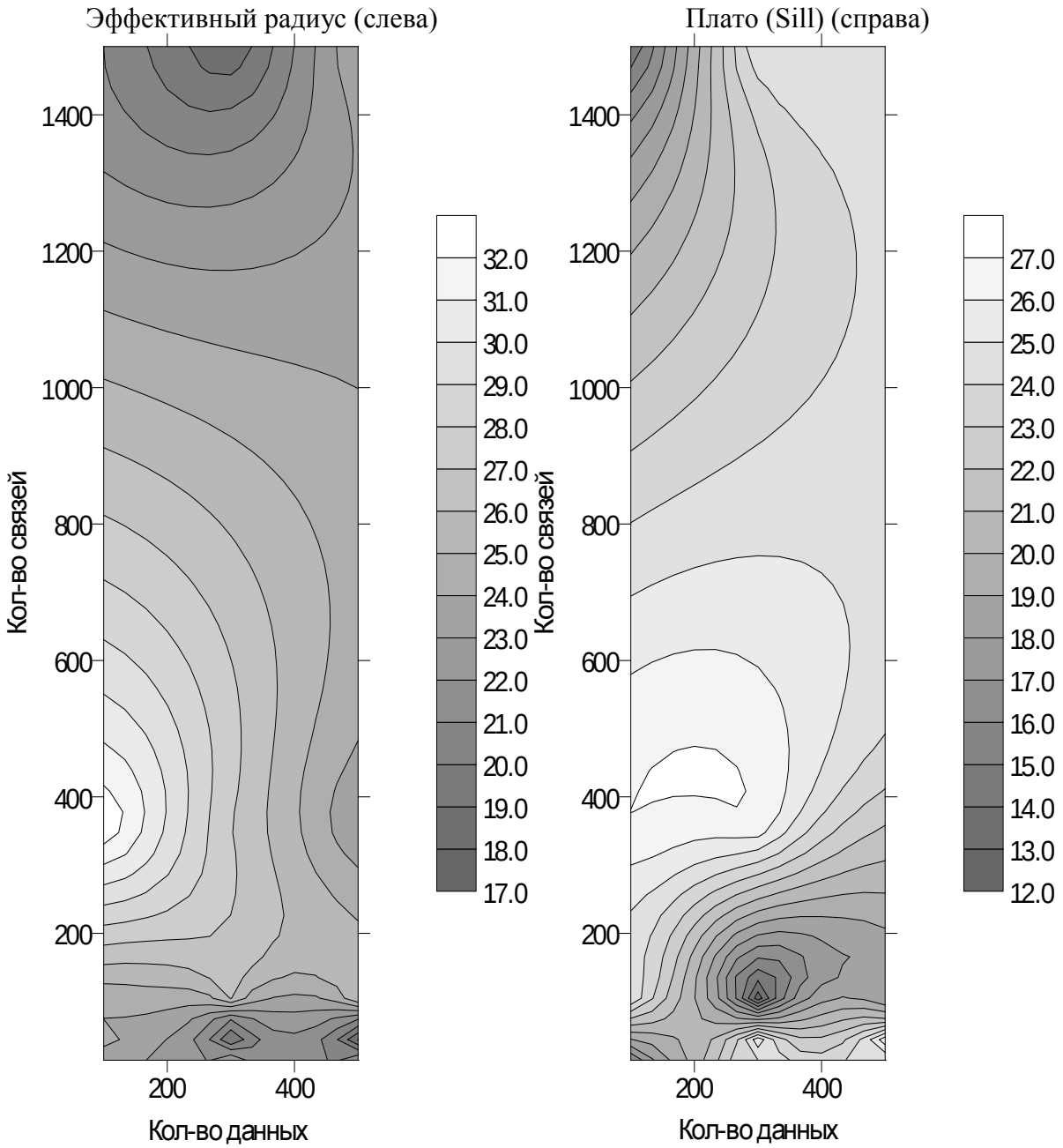
18 Зависимость эффективного радиуса вариограммы невязок от количества связей и количества данных для тренировки



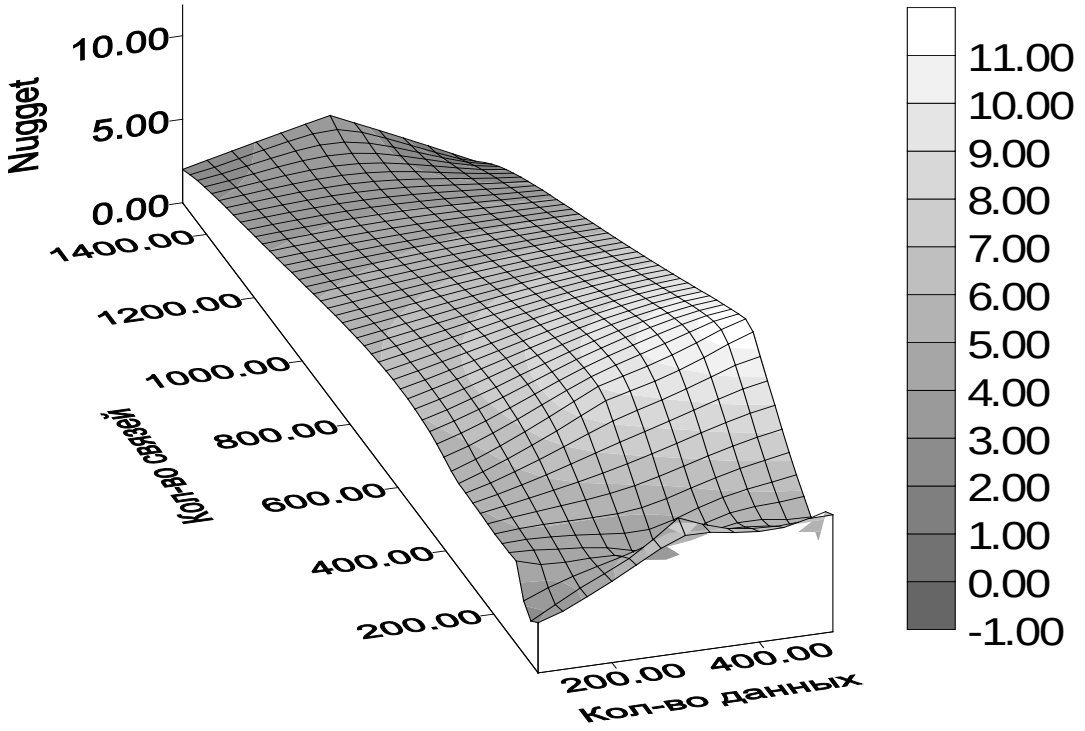
19 Зависимость плато (sill) вариограммы невязок от количества связей и количества данных для тренировки



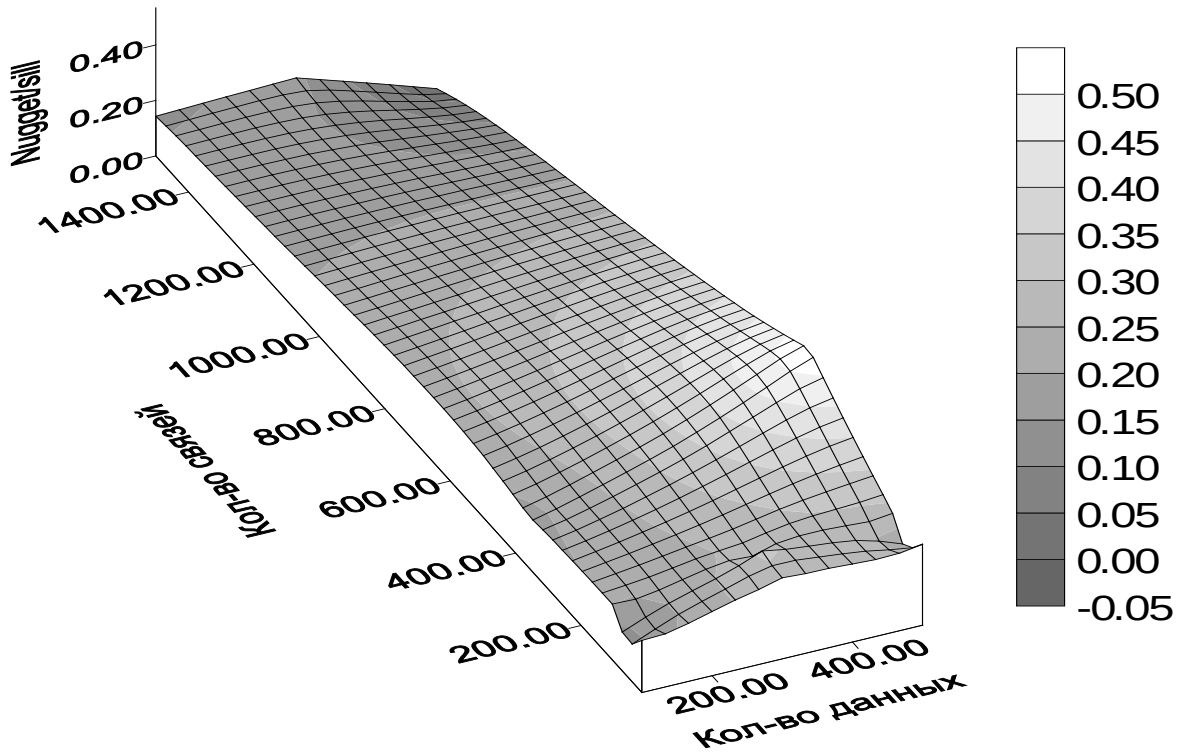
20 Распределение параметров моделей вариограмм в зависимости от числа связей и количества данных для тренировки



21 Зависимость наггета (nugget) вариограммы невязок от количества связей и количества данных для тренировки



22 Зависимость отношения Nugget/Sill вариограммы невязок от количества связей и данных для тренировки



23 Распределение параметров моделей вариограмм в зависимости от числа связей и количества данных для тренировки