ЛАБОРАТОРНАЯ РАБОТА №1 ИССЛЕДОВАНИЕ ГЕОПРОСТРАНСТВЕННЫХ ДАННЫХ С ПОМОЩЬЮ НАБОРА ГРАФИКОВ ДЛЯ ИССЛЕДОВАТЕЛЬСКОГО АНАЛИЗА (ESDA)

Цель работы: освоить методику выполнения исследовательского анализа геопространственных данных с использованием функциональных возможностей модуля Geostatistical Analyst.

Задачи работы: 1) выполнить построение гистограммы распределения геопространственных данных; 2) построить график нормальности распределения геоданных; 3) выполнить анализ тренда данных; 4) оценить стационарность геоданных и построить карту Вороного; 5) построить облако вариограммы геоданных.

Исходные данные для выполнения работы: шейп-файл точечных объектов – данных о содержании гумуса, фосфора, калия и pH почвы в пределах территории сельскохозяйственного предприятия.

Набор графиков для исследовательского анализа пространственных данных (ESDA) позволяет выполнить комплексный статистический анализ геоданных. В состав инструментов для выполнения геостатистического анализа входят следующие (рис. 1).

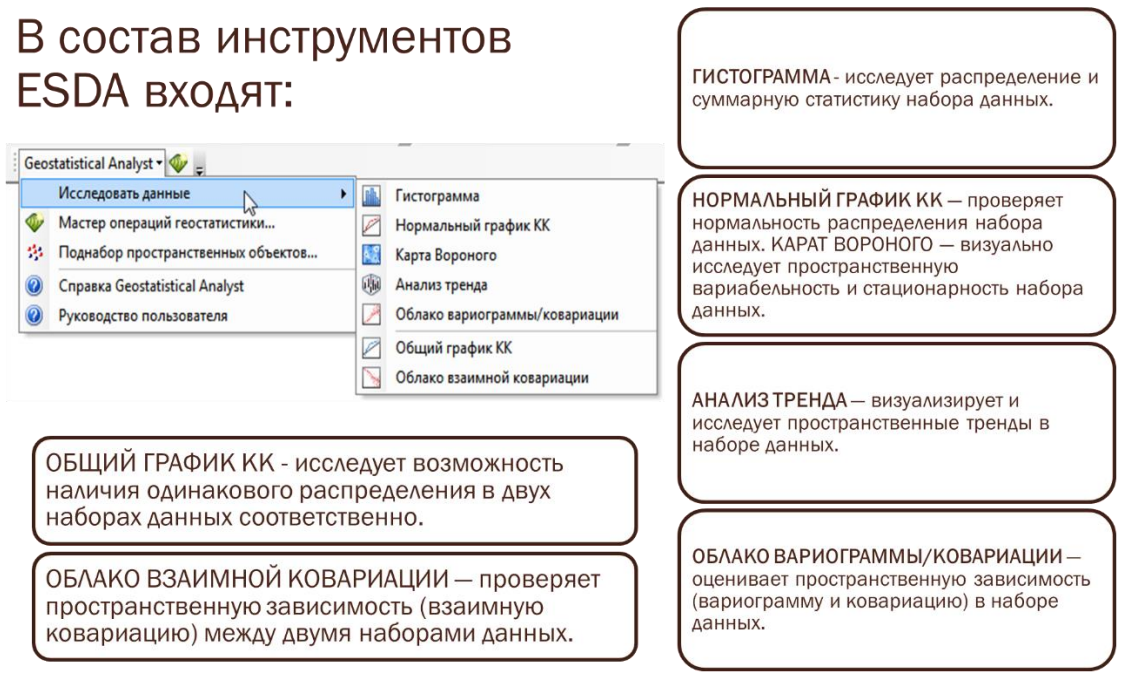


Рис. 1. **Инструменты для выполнения геостатистического анализа**

Ход выполнения работы:

Загрузить файл с исходными данными в таблицу содержания нового рабочего проекта.

1. Создание и анализ гистограммы распределения пространственных данных. Используя возможности модуля «Геостатистический анализ», в частности опцию «Исследовать данные» и инструмент «Гистограмма», построить гистограмму распределения значений содержания гумуса (рис. 2).

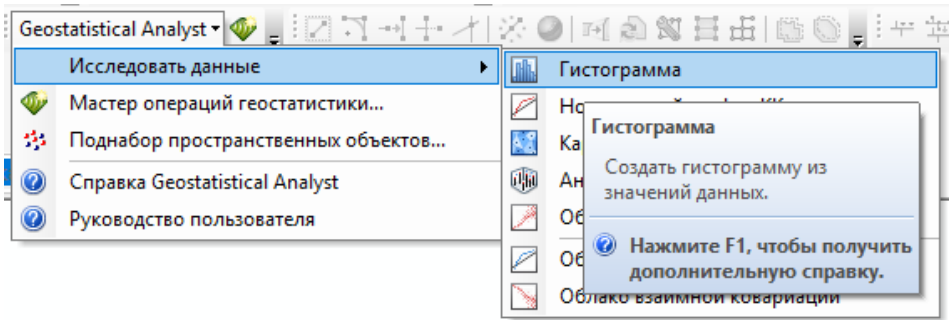


Рис. 2. **Диалоговое окно выбора опции «Гистограмма»**

В результате применения данного инструмента будет создана гистограмма распределения данных (рис. 3).

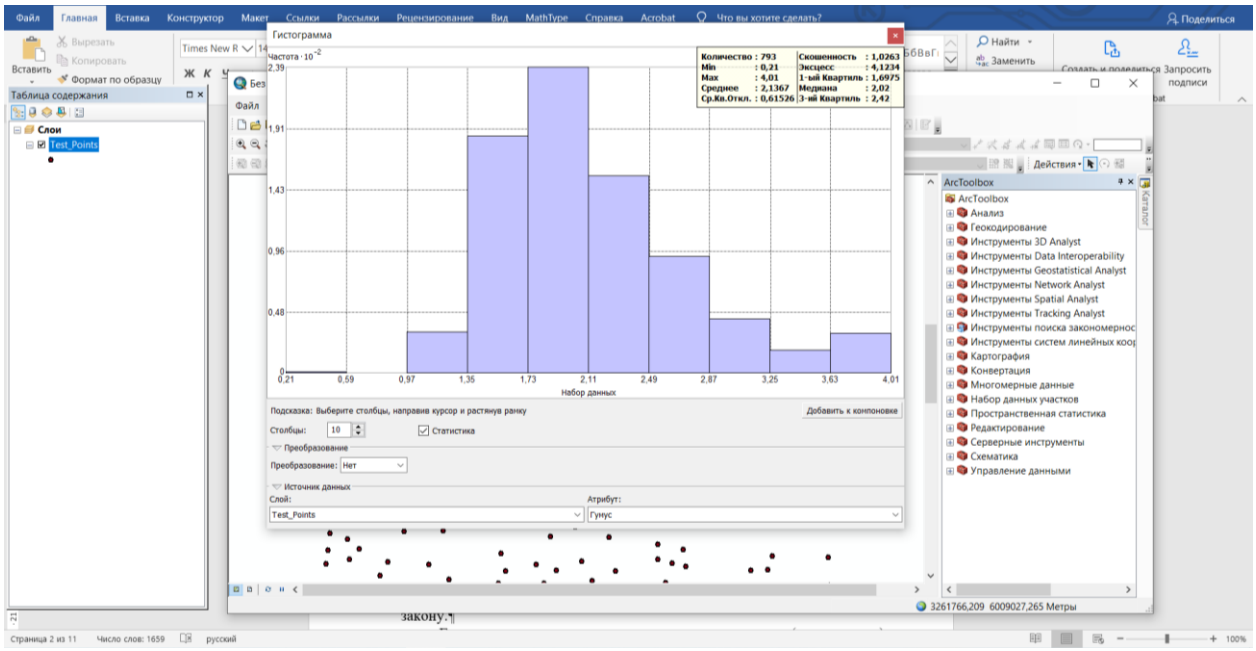


Рис. 3. **Гистограмма распределения геоданных о содержании гумуса в почве**

С помощью гистограммы можно исследовать форму распределения данных. Просматривая статистику среднего значения и медианы, можно

определить расположение центра распределения данных. В идеале распределение данных должно иметь колоколообразную форму, что свидетельствует о его нормальности. О нормальном распределении данных свидетельствует и значение медианы, которое должно быть максимально приближено к среднему значению. В приведенном примере среднее и медиана примерно равны, что является аргументом в пользу того, что данные могут быть распределены по нормальному закону.

Гистограмма данных о содержании гумуса в почве говорит о том, что распределение данных является одновершинным (с одной выпуклостью) и смещено вправо. Левый хвост распределения показывает относительно малое количество точек выборки с небольшими значениями содержания гумуса.

Если распределение геоданных имеет несколько пиков (экстремумов), то есть данные распределены асимметрично, к ним применяется логарифмическое преобразование, которое приближает распределение к нормальному (рис. 4).

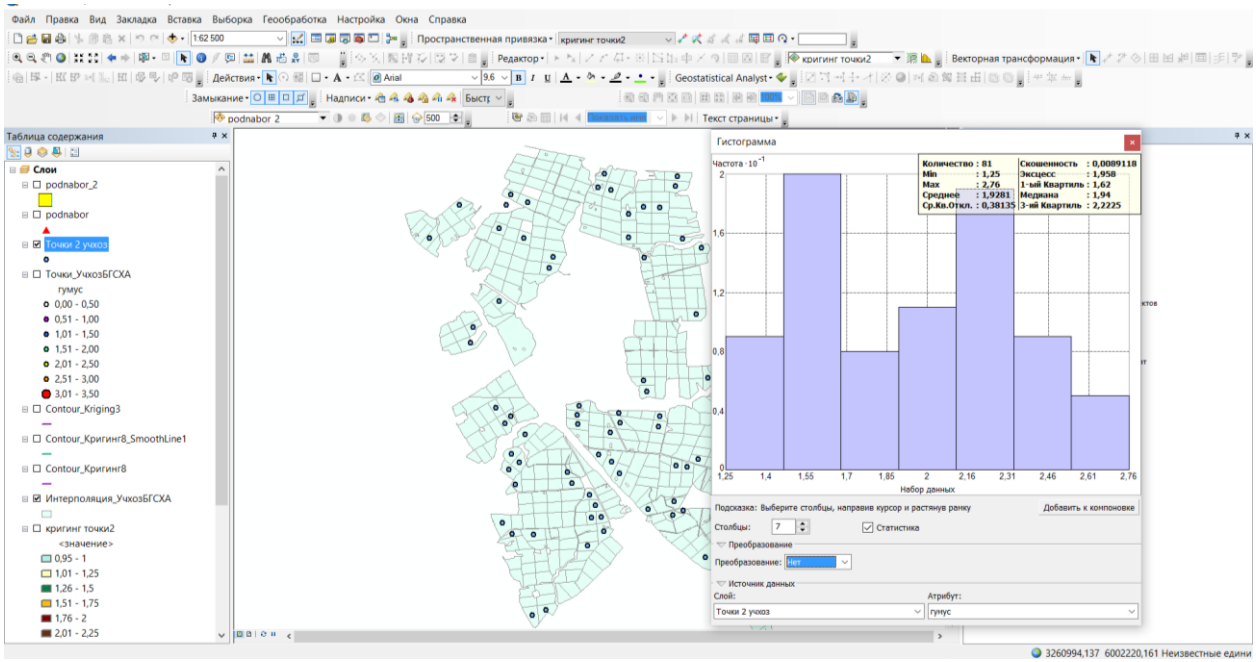


Рис. 4. **Гистограмма распределения геоданных с несколькими экстремумами**

Для выполнения преобразования следует воспользоваться выпадающим списком рядом с кнопкой «Преобразование» (рис. 5).

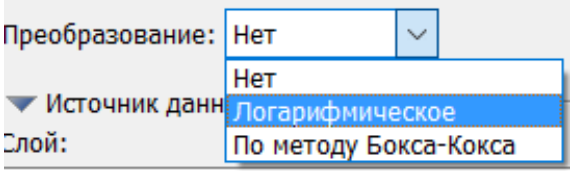


Рис. 5. **Опции преобразования геоданных**

Логарифмическое преобразование часто используется, когда данные смещены в положительном или отрицательном направлении и присутствует мало очень больших или слишком малых значений.

Преобразование по методу арксинуса может быть использовано для данных, которые представляют относительное содержание или проценты.

На рисунке 6 показана гистограмма до и после выполнения логарифмического преобразования. После преобразования в гистограмме остался только один экстремум (рис. 6б).

а) б)

Рис. 6. **Гистограмма распределения геоданных до (а) и после (б) выполнения логарифмического преобразования**

Если у распределения есть длинный правый хвост больших значений, то у него *положительная асимметрия*, а если длинный левый хвост малых значений – то *отрицательная*. Среднее значение для распределений с положительной асимметрией больше, чем медиана, а для распределений с отрицательной асимметрией – наоборот.

Также на гистограмме можно выделить экстремальные значения и увидеть, как они расположены в пространстве на отображаемой карте. Следует помнить о необходимости выбора атрибута – поля со значениями из атрибутивной таблицы, данные из которого будут анализироваться. В данном примере это поле «гумус».

В данном инструменте доступна опция *удаления экстремально низких или высоких значений* из выборки. Для этого следует выделить столбцы гистограммы с нежелательными значениями, щёлкнув на кнопку мыши и протащив курсор по этим столбцам. Опорные точки в данном диапазоне будут выделены цветом. Далее следует щёлкнуть кнопку «Очистить выбранные объекты» на панели «Инструменты», чтобы удалить выбранные точки в проекте и на гистограмме (рис. 7).

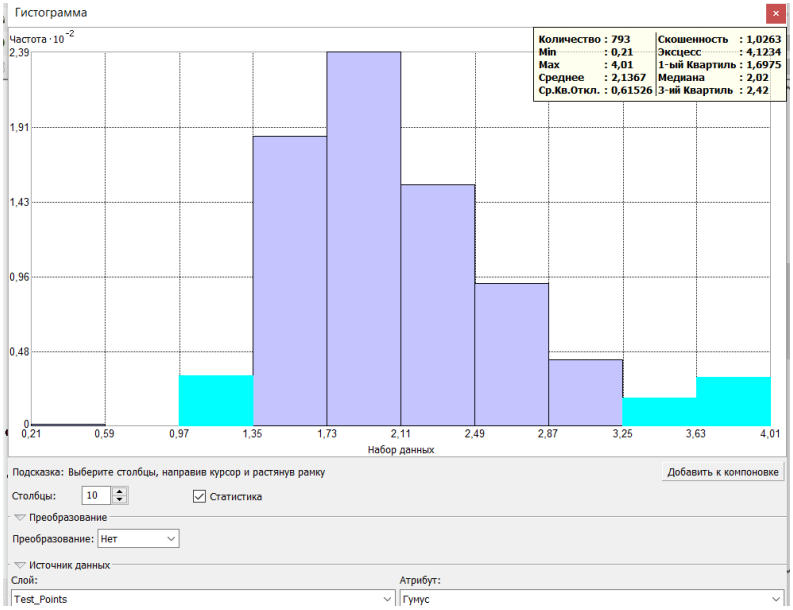


Рис. 7. **Гистограмма распределения геоданных с выбранными экстремальными значениями**

В таблице статистики присутствуют сведения о величине среднего, минимального и максимального значения в ряду данных, а также об их разбросе. Разброс или дисперсия точек вокруг среднего значения - еще одна важная характеристика отображаемого частотного распределения.

Дисперсия данных представляет собой среднеквадратическое отклонение всех значений от среднего. Поскольку в нее включаются квадраты разностей, вычисляемая дисперсия чувствительна к необычно высоким или низким значениям. Дисперсия оценивается суммированием квадратических отклонений от среднего и делением суммы на (N-1), где N – общее количество наблюдений или значений. *Стандартное отклонение* представляет собой квадратный корень из дисперсии и описывает разброс данных вокруг среднего. Чем меньше дисперсия и стандартное отклонение, тем гуще сконцентрирован кластер измерений вокруг среднего значения и тем ниже степень варьирования данных.

Эксцесс основан на размере хвостов распределения и представляет собой показатель вероятности того, что распределение будет создавать выпадающие значения. Эксцесс нормального распределения равен трем. Распределения с относительно толстыми хвостами называются островершинными (лептокуртическими), и у них эксцесс больше трех. Распределения с относительно тонкими хвостами называются плосковершинными (платикуртическими), и у них эксцесс меньше трех.

Показателями, характеризующими распределение данных, являются *квартили*. Первая и третья квартили соответствуют кумулятивной пропорции 0,25 и 0,75. Если данные организованы в порядке возрастания, 25 процентов значений будут находиться ниже первой квартили, а еще 25 процентов - выше третьей квартили. Первая и третья квартили являются особыми случаями квантилей. Квантили вычисляются следующим образом: quantile = (i – 0,5) / N, где i – упорядоченное *i*-тое значение данных.

2. Создание и анализ нормального графика распределения пространственных данных. Используя опцию «Нормальный график КК» модуля «Геостатистический анализ» проверить нормальность распределения набора данных. Точки нормального графика КК дают представление об одномерной нормальности набора данных. Если данные распределены нормально, точки выстроятся на базовой линии, проходящей под углом 45 градусов. Если данные не распределены нормально, точки будут отклоняться от базовой линии. По результатам применения указанной опции будет построен график следующего вида (рис. 8).

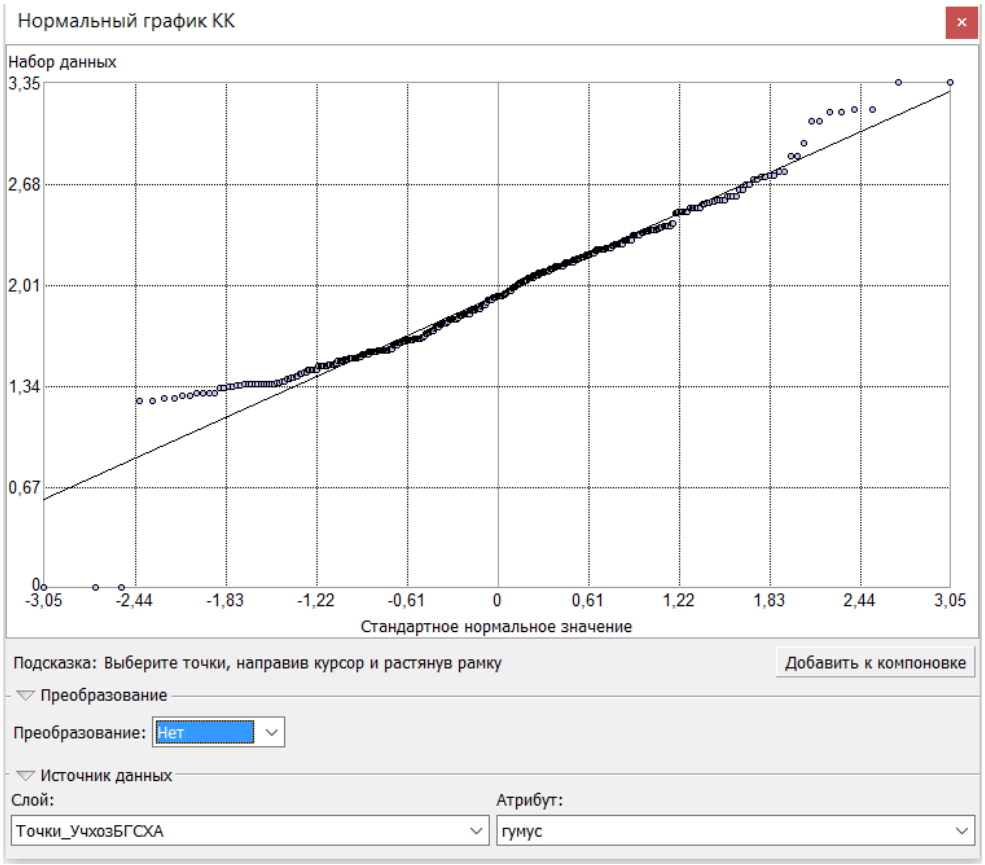


Рис. 8. **График нормальности распределения геоданных**

Значения квантилей стандартного нормального распределения нанесены на ось X нормального графика КК, а соответствующие значение квантилей набора данных – на ось Y. В данном примере точки значений расположены близко к 45-градусной базовой линии. Основное отклонение от этой линии возникает при высоких (более 3%) и низких (менее 0,5%) значениях содержания гумуса в почве рабочих участков.

Направив курсор на график и растянув рамку, можно выбрать точки со значениями, которые автоматически отобразятся и в рабочем проекте. Таким образом можно просмотреть, где именно находятся точки, данные в которых имеют слишком высокие или очень низкие значения (рис. 9).

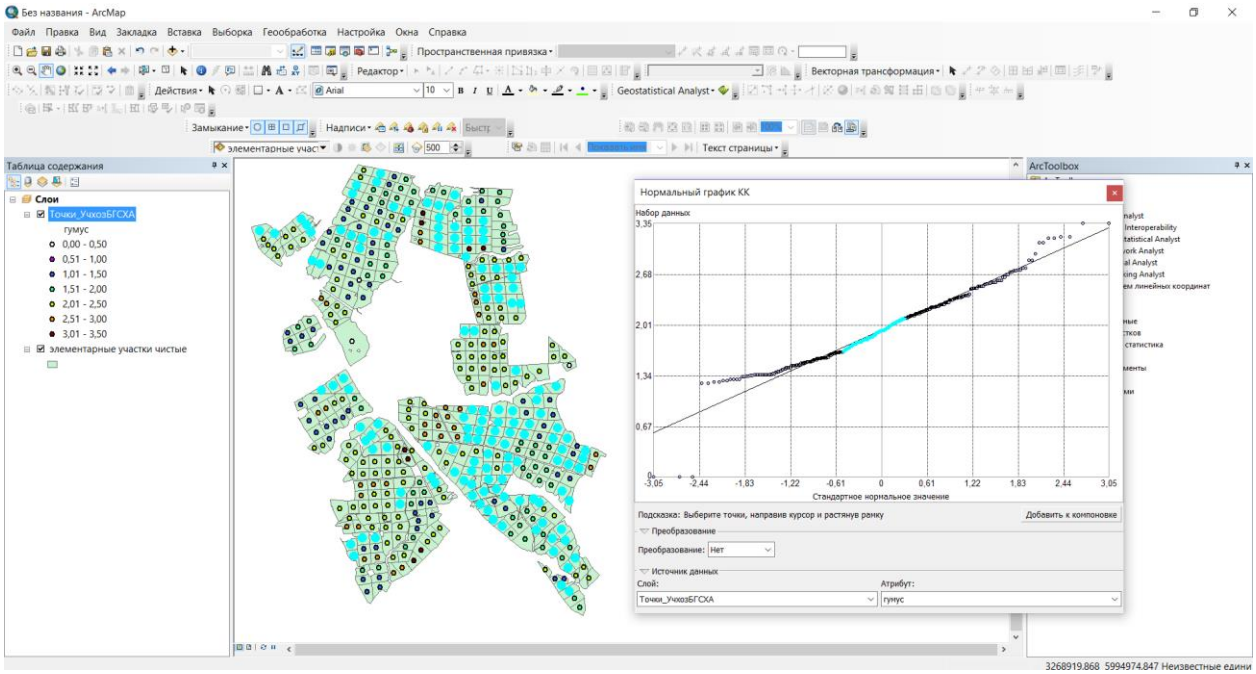
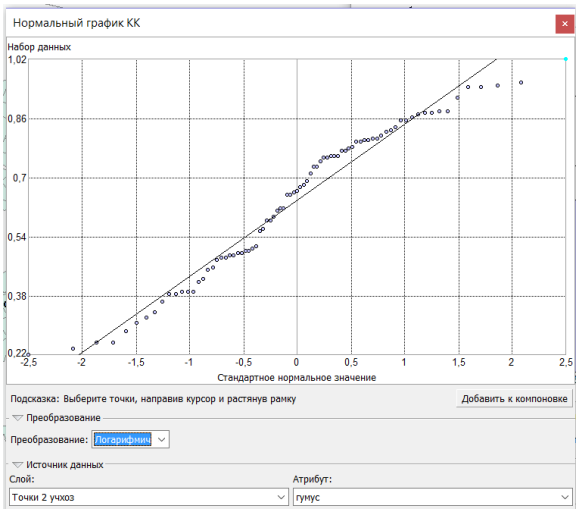


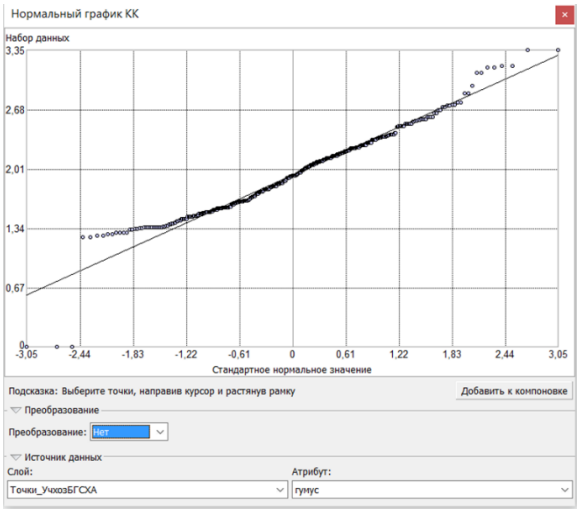
Рис. 9. **Рабочее окно проекта с отображением геоданных, выбранных на графике нормальности распределения**

Чтобы снять выделение с объектов, следует щелкнуть правой кнопкой мыши по названию слоя и выбрать путь: «Выборка» – «Очистить выбранные объекты».

Если отклонения данных от нормального графика слишком значительны, необходимо выполнить преобразование (рис. 10).



Нормальный график распределения без преобразования данных



Нормальный график распределения после логарифмического преобразования данных

Рис. 10. **График нормального распределения данных до и после выполнения преобразования**

3. Анализ тренда геопространственных данных. С помощью инструмента «Анализ тренда» можно выявлять тренды в наборе входных геоданных. Инструмент «Анализ тренда» позволяет отображать данные в трехмерной перспективе. Местоположения опорных точек наносятся на плоскость x, y. Над каждой точкой значение атрибута (в данном примере это содержание гумуса) отображается высотой отрезка (оси z).

Уникальной особенностью инструмента «Анализ тренда» является то, что значения проецируются на плоскости x, z и y, z в виде диаграмм рассеивания. Также это применимо к видам сбоку посредством трехмерных данных. Затем на проецируемых плоскостях выполняется подгон полиномов с помощью диаграмм рассеивания. Дополнительной особенностью является возможность вращения данных для отделения трендов направления. Инструмент также оснащен другими возможностями, позволяющими поворачивать и изменять перспективу всего изображения, размер и цвет точек и линий, удалять плоскости и точки, а также выбирать степень подгоняемого под диаграммы рассеивания полинома. Они реализуются через опцию «Опции графика». По умолчанию для отображения трендов данных инструмент выбирает полиномы второй степени, тем не менее можно оценить уровень соответствия данным полиномов первой и третьей степени.

Каждый вертикальный отрезок на графике анализа тренда представляет местоположение, а высота отрезка – значение каждого измерения концентрации гумуса. Точки данных проецируются на перпендикулярные плоскости - восток-запад (плоскость x, z) и север-юг (плоскость y, z) (рис. 11).

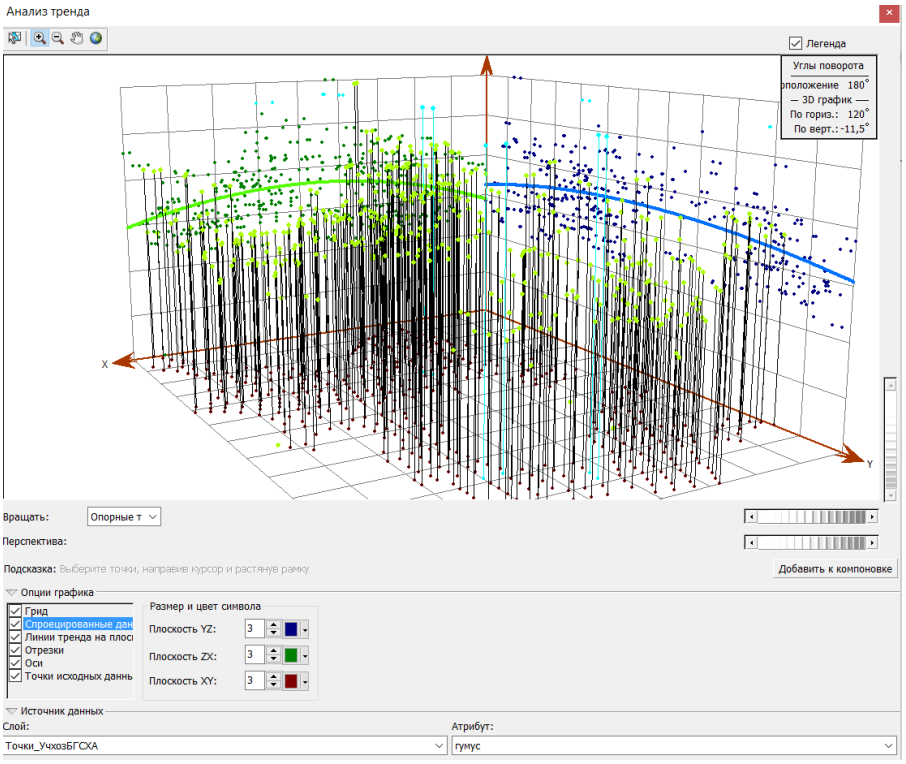


Рис. 11. **Результат анализа тренда геоданных**

Линия наилучшего соответствия (полином) проведена через проецируемые точки, показывая тренды в определенных направлениях. Если бы линия была ровная, это означало бы, что трендов нет. Однако светло-зеленая линия на приведенном выше рисунке начинается с низких значений, незначительно растет по направлению к центру оси x, а затем снижается. Аналогично синяя линия незначительно растет в северном направлении и также постепенно снижается. Это говорит о незначительно выраженном тренде, начиная с центра области данных во всех направлениях. Поскольку тренд имеет U-образную форму, оптимально использовать полином второго порядка в качестве глобальной модели тренда.

4. Создание карты Вороного. Карты Вороного строятся из серии полигонов, сформированных вокруг местоположения каждой точки выборки геопространственных данных. Полигоны Вороного созданы так, что каждое местоположение в пределах полигона находится ближе к выбранной точке этого полигона, чем любая другая выбранная точка. После создания полигонов, соседи выбранной точки определяются также, как и любая другая выбранная точка, чей полигон граничит с выбранной точкой (рис. 12).

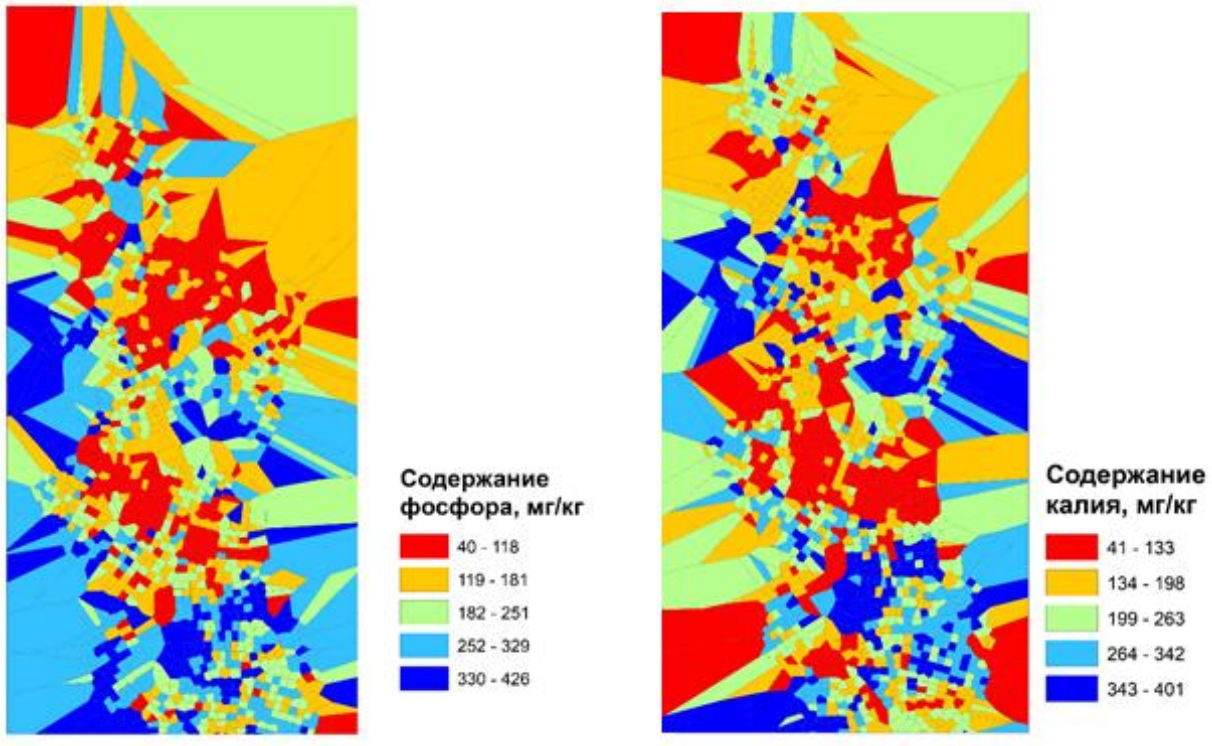


Рисунок 12. **Карты Вороного, иллюстрирующие неоднородность содержания подвижных фосфора и калия в почве**

Карты Вороного позволяют идентифицировать глобальные и локальные выпадающие значения и оценить стационарность геопространственных данных. Под глобальными выпадающими значениями подразумевают точки с очень высоким или очень низким значением по сравнению со всеми

значениями в наборе данных, а под локальными выпадающими значениями – измеренные опорные точки, которые имеют значение в пределах нормы для всего набора данных, но если посмотреть на окрестные точки, то это значение будет чрезвычайно высоким или низким по сравнению с ними. В данном примере в пределах исследуемой территории присутствуют точки с локальными выпадающими значениями.

Выпадающие значения важно идентифицировать поскольку они могут быть как реальными аномалиями в явлении, так и причиной неправильного измерения значения. Если выпадающее значение является фактической аномалией в явлении, то оно может быть самой показательной точкой в исследовании и осмыслении явления. В случае, если выпадающие значения вызваны ошибками во время ввода очевидно неправильных данных, они должны быть исправлены или удалены перед созданием интерполяционной поверхности.

5. Создание облака вариограммы. Облако вариограммы/ковариации позволяет проанализировать пространственную автокорреляцию между измеренными точками выборки. В общем случае предполагается, что объекты, расположенные близко друг к другу, более схожи, чем удаленные друг от друга. Облако вариограммы/ковариации позволяет проверить эту взаимосвязь. Для этого значение вариограммы, которое представляет собой квадрат разницы между значениями каждой пары местоположений, наносится на график по оси y, а по оси x откладывается расстояние между точками измерений в каждой паре (рис. 13).

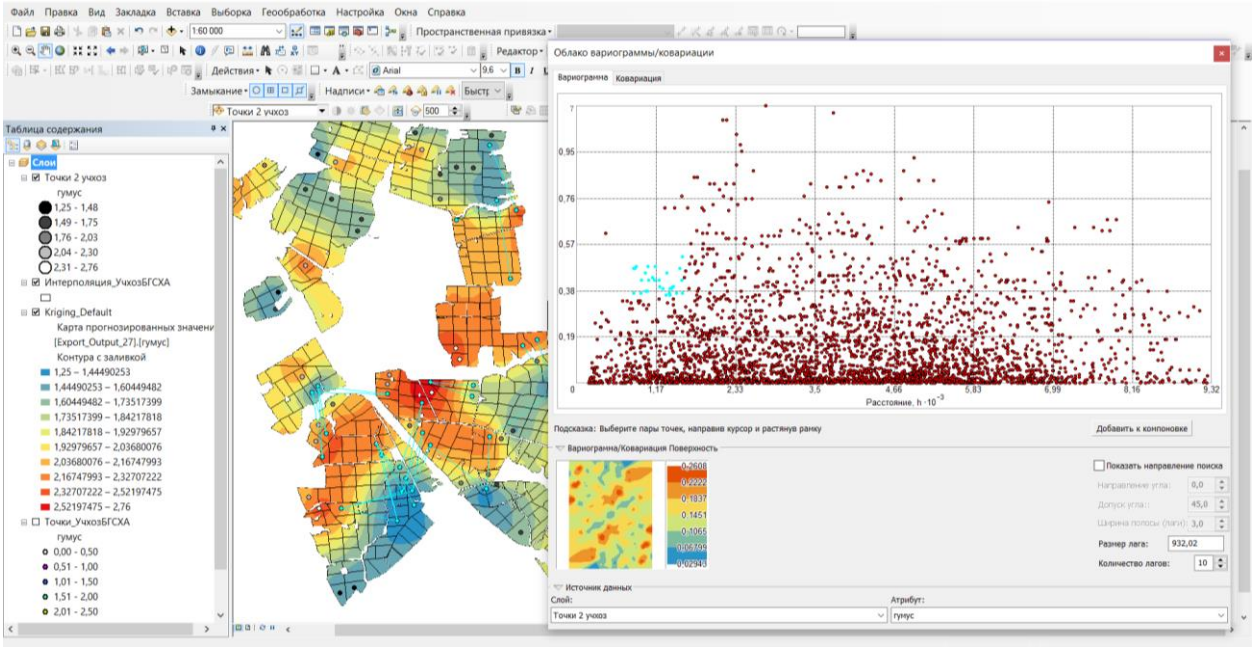


Рис. 13. **Облако вариограммы геоданных**

Каждая красная точка в облаке вариограммы/ковариации представляет пару местоположений. Исходя из того, что местоположения, близкие друг к

другу, должны быть более схожи, на графике вариограммы ближайшим местоположениям (в крайней левой области по оси x) должны соответствовать невысокие значения вариограммы (низкие значения по оси y). По мере увеличения расстояния между парами местоположений (вправо по оси x) значения вариограммы должны также расти (вверх по оси y). Однако по достижении определенного расстояния облако перестает меняться. Это свидетельствует о том, что значения в парах точек, расположенных друг от друга дальше этого расстояния, больше не коррелированы.

Если при рассмотрении вариограммы окажется, что некоторые местоположения данных, близкие друг к другу (около нуля по оси x), имеют более высокие значения вариограммы (вверх по оси y), чем ожидалось, то следует изучить эти пары местоположений на предмет точности данных. Направив курсор и растянув рамку, можно выбрать точки в пределах облака вариограммы и просмотреть их расположение на карте (рис. 14).

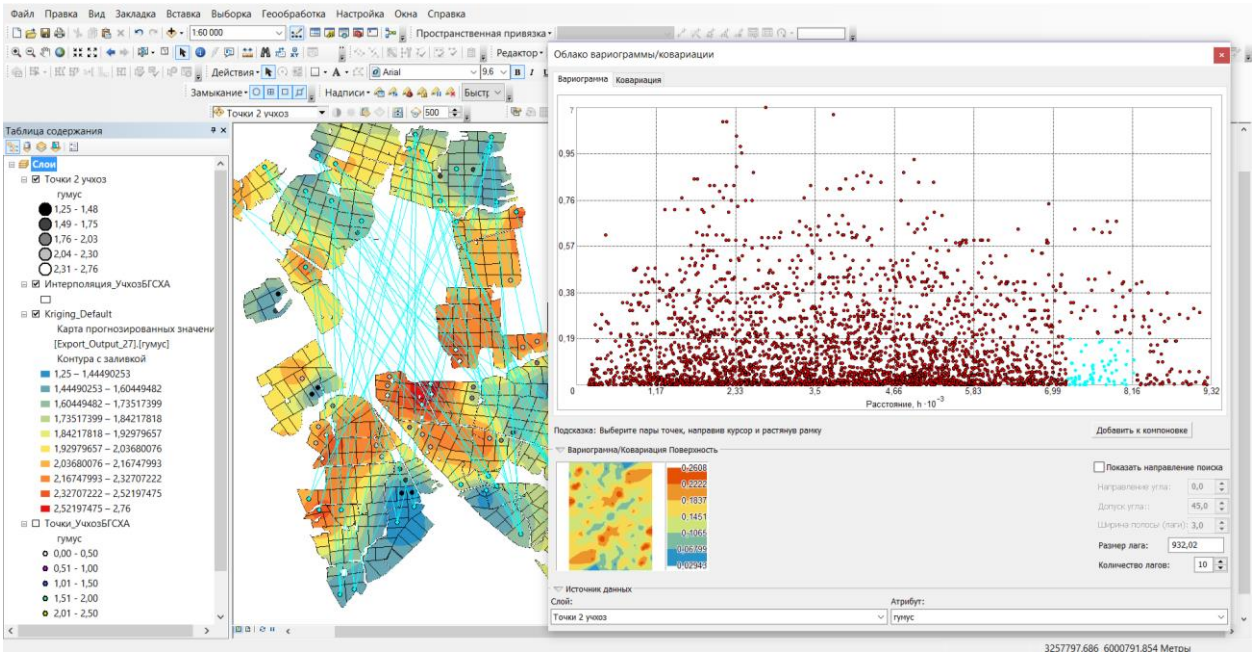


Рис. 14. **Результат выбора точек интереса в пределах облака вариограммы**

Чтобы проанализировать влияние направления на облако вариограммы, можно использовать инструмент «Направление поиска». Для этого следует установить флажок «Показывать направление поиска», щёлкнуть на указателе направления и переместите его на любой угол.

Направление по указателю определяет, какие пары местоположений данных будут нанесены на вариограмму. Например, если указатель ориентирован в направлении юго-восток, на вариограмму будут нанесены только пары точек данных, расположенных к юго-востоку друг от друга. Это позволяет исключить лишние пары и проанализировать влияния направления на данные (рис. 15)

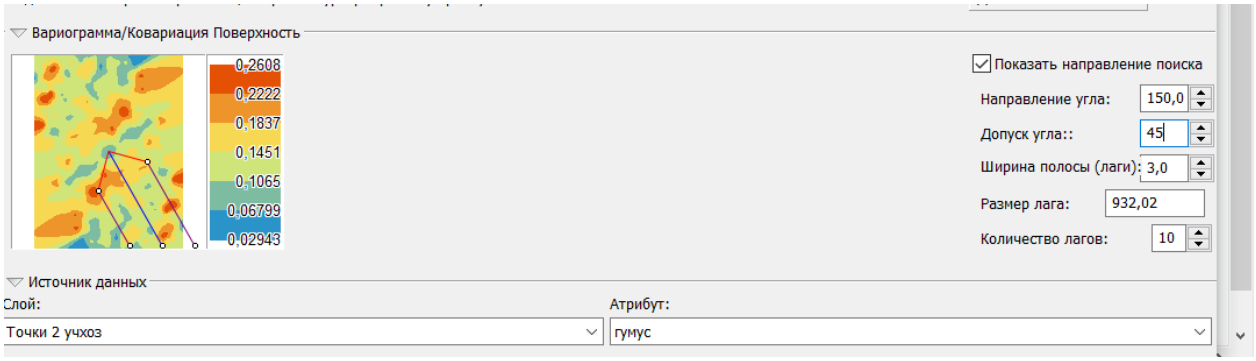


Рис. 15. **Пример указателя направления**

Построение вариограмм помогает определить, какие методы целесообразно использовать для пространственного моделирования – детерминированные или геостатистические.

В целом исследовательский анализ геопространственных данных позволяет проверить их на соответствие требованиям того или иного метода интерполяции и лучше понять феномен исследования, в результате чего можно принять более взвешенное и обоснованное решение при выборе адекватной модели интерполяции.

ЛИТЕРАТУРНЫЕ И ИНФОРМАЦИОННЫЕ ИСТОЧНИКИ:

1. Геостатистика: теория и практика / В. В. Демьянов, Е. А. Савельева; под ред. Р. В. Арутюняна; Ин-т проблем безопасного развития атомной энергетики РАН. – М.: Наука, 2010. – 327 с.