ЛАБОРАТОРНАЯ РАБОТА №5 МОДЕЛИРОВАНИЕ ПРОСТРАНСТВЕННЫХ ОТНОШЕНИЙ И РАСЧЕТ КЛАСТЕРИЗАЦИИ ГЕОДАННЫХ

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

Задачи работы: 1) выполнить расчет усредненного центра данных и эллипса стандартного отклонения; 2) рассчитать индекс ближайшего соседства; 3) рассчитать пошаговую пространственную автокорреляцию; 4) рассчитать величину глобального индекса Морана; 5) рассчитать величину глобального индекса Getis-Ord G; 6) выполнить анализ горячих точек; 7) выполнить анализ кластеров и выбросов геоданных.

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

Геопространственная статистика – это комбинация методов традиционной статистики и геопространственных данных, под которыми подразумеваются данные (или как принято в статистике, переменные), имеющие пространственную привязку, для которых известны координаты их местоположения, которая оперирует не случайными величинами, а пространственными переменными. В отношении поиска закономерностей в пространственных данных геостатистика позволяет выполнять измерение пространственного распределения геоданных, анализ их структурных закономерностей, а также расчет кластеризации и ее анализ.

Одним из вариантов применения методов геопространственной статистики для целей землеустройства может стать анализ пространственного распределения агрохимических свойств почв земель сельскохозяйственного назначения. При выполнении такого анализа возможно:

- выявить и математически оценить пространственное распределение агрохимических показателей почвы;
- изучить пространственную автокорреляцию данных и определить местоположения в области исследования с аномальными значениями;
- оценить кластеризацию данных об агрохимических свойствах почвы и определить местоположения кластеров в пространстве;
- выполнить визуализацию кластеров путем построения карты локального индикатора пространственной ассоциативности;

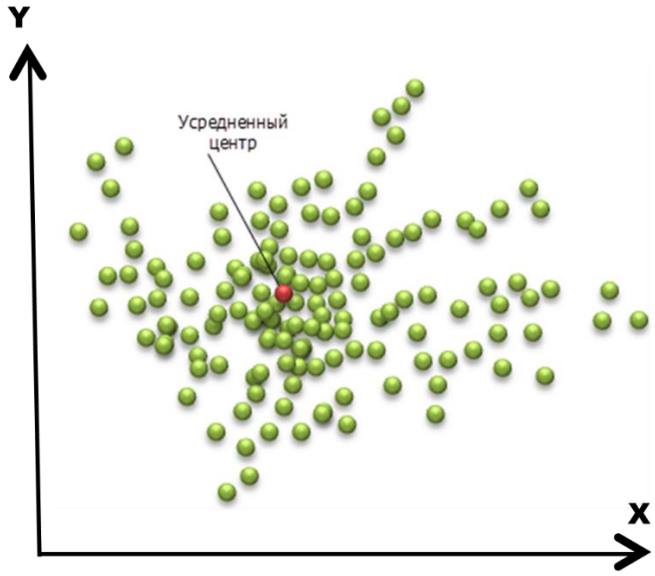
- установить наиболее четкие границы между плодородными и мало плодородными землями.

Еще одним прикладным аспектом применения методов геопространственной статистики является выделение однородных по агрофизическим и агрохимическим свойствам почв и агротехнологическим характеристикам участков пахотных земель для дальнейшего внедрения системы точного земледелия.

Ход выполнения работы:

1. Расчет усредненного центра данных и эллипса стандартного отклонения

Усредненный центр распределения данных определяет географический центр (или центр концентрации) для набора объектов и рассчитывается либо по атрибутам значений, либо по значениям координат x и y. В случае наличия данных с большим разбросом минимальных и максимальных значений альтернативой усредненному центру является медианный центр. Данная опция может оказаться полезной в случае поиска массива земель со сходными параметрами – агрофизическими, агрохимическими, физико-химическими или любыми другими (рис. 1).



$$\bar{X}w = \frac{\sum x_i * w_i}{\sum w_i} \qquad \bar{Y}w = \frac{\sum y_i * w_i}{\sum w_i}$$

где: $\bar{X}w$ и $\bar{Y}w$ – координаты усредненного центра распределения данных;

x_i и y_i – координаты каждой точки, для которой известно значение атрибута;

w_i – значение весового атрибута в точке данных с координатами x_i и y_i .

Рис. 1. **Сущность усредненного центра данных**

Используя возможности модуля «Пространственная статистика» набора инструментов ArcToolBox, в частности опцию «Усредненный центр», определить усредненный центр исходных данных по атрибутивному полю, указанному в индивидуальном задании, выполнив настройки, как показано на рис. 2. Атрибутивное поле, указанное в задании, необходимо указать в поле диалогового окна «Поле измерений» (в данном примере это атрибутивное поле «калий»).

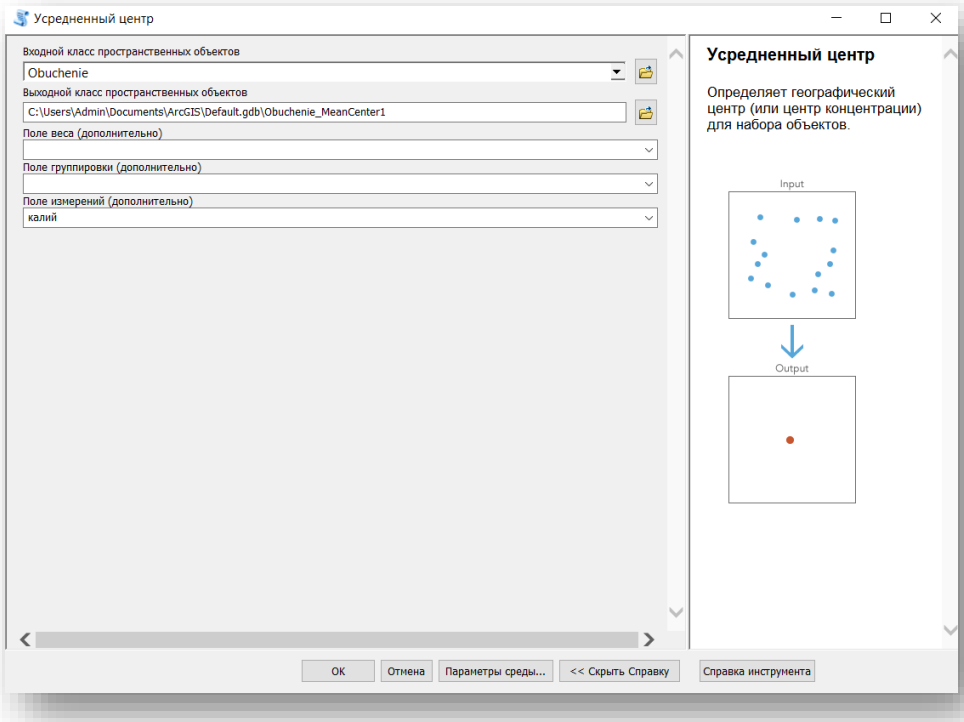


Рис. 2. **Диалоговое окно настроек опции «Усредненный центр»**

В результате применения данного инструмента будет построен усредненный центр данных о содержании подвижного калия в пределах землепользования сельскохозяйственного предприятия (рис. 3).

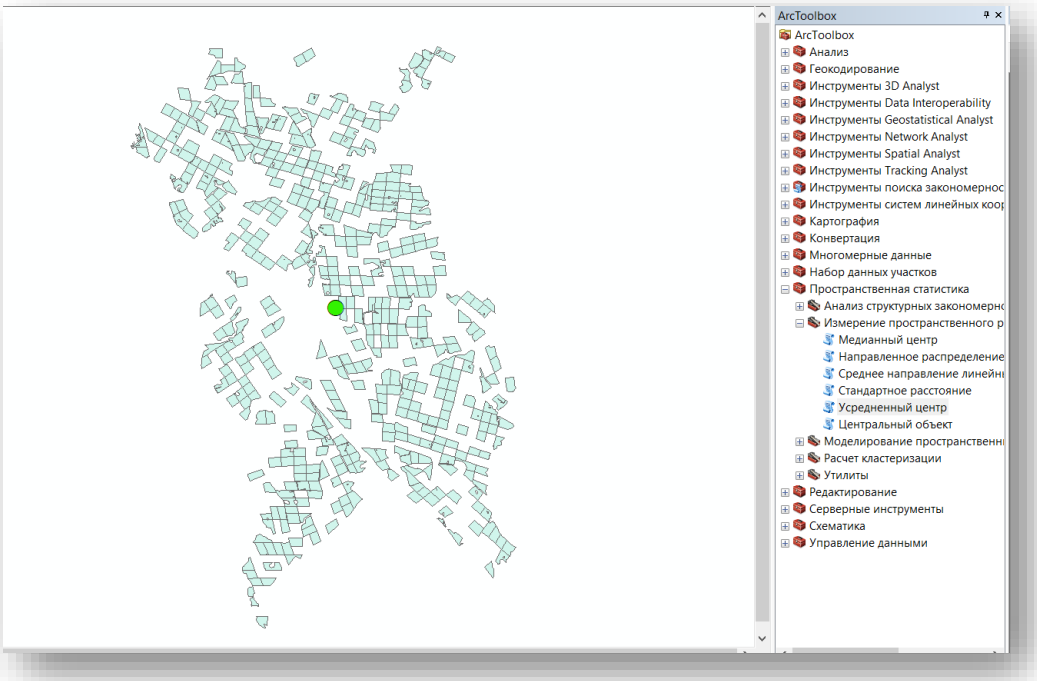
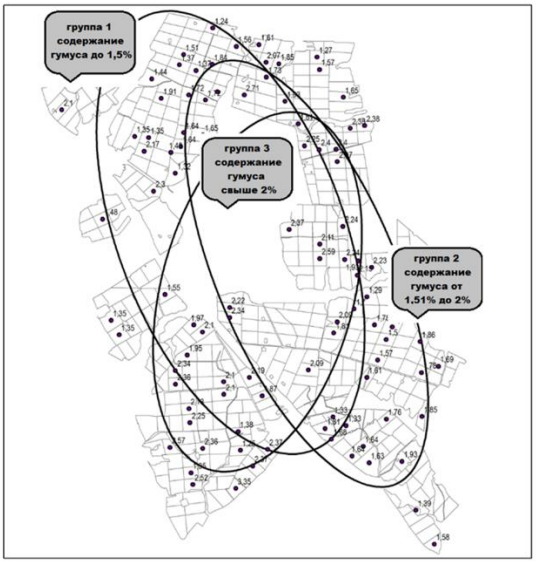


Рис. 3. **Результат реализации опции «Усредненный центр»**

Инструмент «направленное распределение» создает эллипс со стандартным отклонением для обобщения пространственных характеристик географических объектов: центральной тенденции, дисперсии и направленных тенденций. Эллипс стандартного отклонения является пространственной вариацией усредненного центра данных (рис. 4).



НАПРАВЛЕННОЕ
РАСПРЕДЕЛЕНИЕ ДАННЫХ

$$SDx_w = \sqrt{\frac{\sum w_i * (x_i - \bar{X}w)^2}{\sum w_i}}$$

$$SDy_w = \sqrt{\frac{\sum w_i * (y_i - \bar{Y}w)^2}{\sum w_i}}$$

где:
SDx_w и SDy_w – среднеквадратичное отклонение по координатам x и y от среднего значения выборки геопространственных данных.

13

Рис. 4. **Сущность эллипса направленного распределения**

Используя возможности модуля «Пространственная статистика» набора инструментов ArcToolBox, в частности опцию «Направленное распределение», определить эллипс стандартного отклонения исходных данных по атрибутивному полю, указанному в индивидуальном задании, выполнив настройки, как показано на рис. 5.

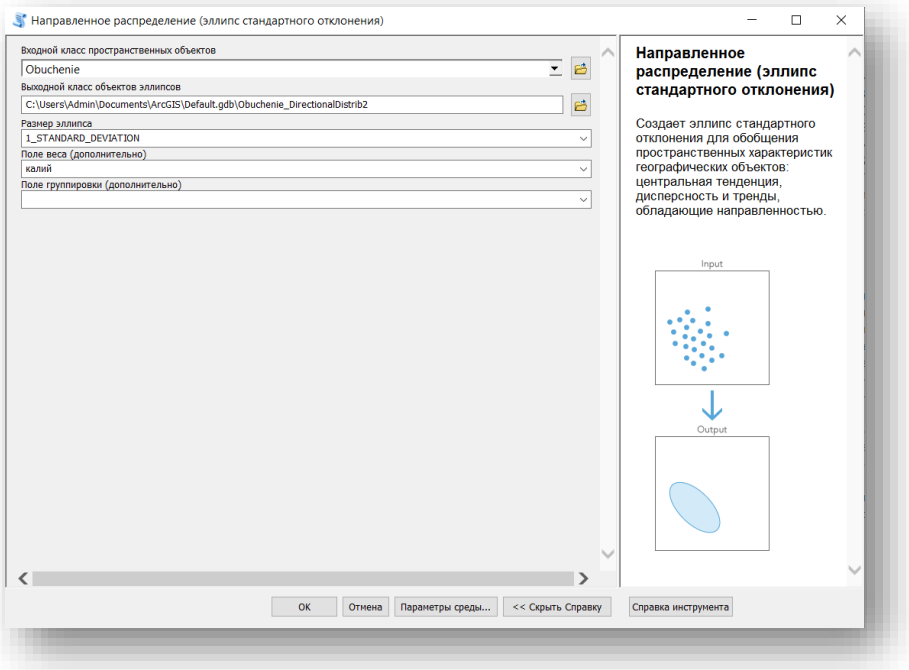


Рис. 5. **Диалоговое окно настроек опции «Направленное распределение»**

В результате применения данного инструмента будет построен эллипс стандартного отклонения содержании подвижного калия в пределах землепользования сельскохозяйственного предприятия (рис. 6).

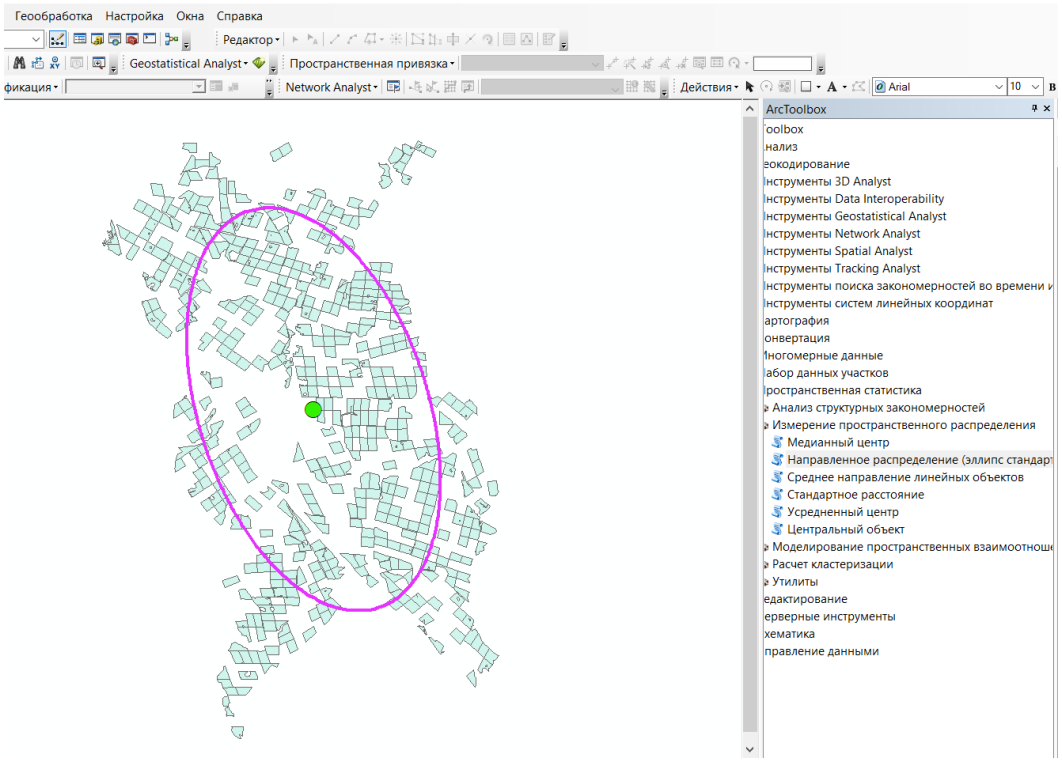


Рис. 6. **Результат реализации опции «Направленное распределение»**

Сравнивая размер, форму и перекрытие эллипсов для различных показателей, например агрохимических и физико-химических свойств почвы, можно получить дополнительную информацию об их пространственной взаимосвязи. Либо, как в представленном на слайде примере, где данные сгруппированы по какому-либо признаку, получить несколько эллипсов для каждой из групп и использовать полученную информацию для установления оптимального количества того либо иного показателя. Данную опцию можно использовать, например, для установления границ зоны с максимальной либо минимальной стоимостью недвижимости по комплексу параметров, для каждого из которых строится отдельный эллипс распределения, а искомой зоной станет та, где будет зафиксировано максимальное количество пересечений эллипсов.

2. Расчет индекса ближайшего соседства

Используя возможности модуля «Пространственная статистика», в частности опцию «Среднее ближайшее соседство», определить индекс ближайшего соседства для исходных данных по атрибутивному полю, указанному в индивидуальном задании, выполнив настройки, как показано на рис. 7.

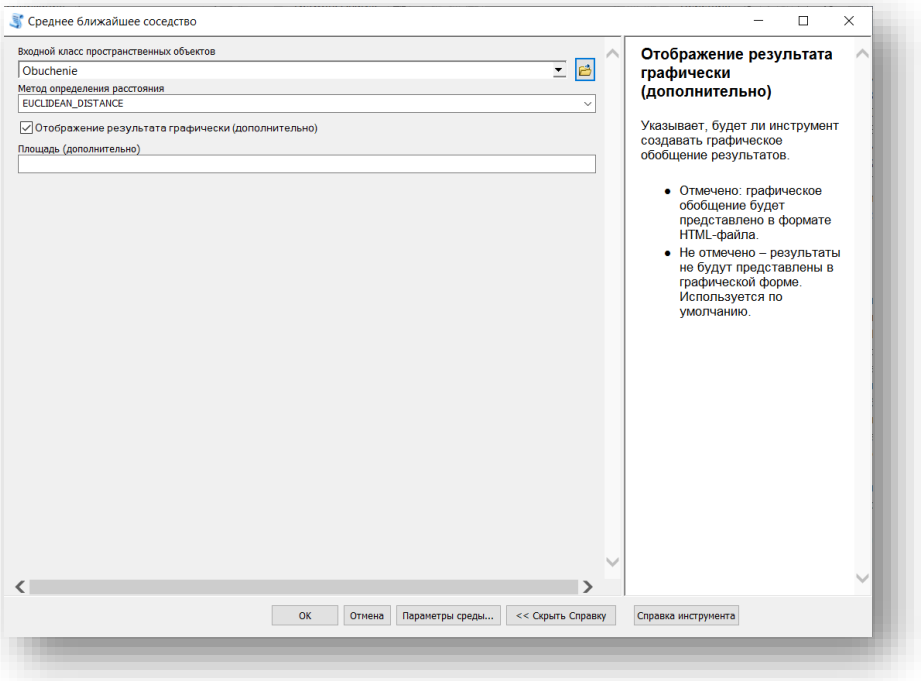


Рис. 7. **Диалоговое окно настроек опции «Среднее ближайшее соседство»**

По результатам реализации опции создается текстовый и графический отчет. На рис. 8 представлен пример текстового отчета.

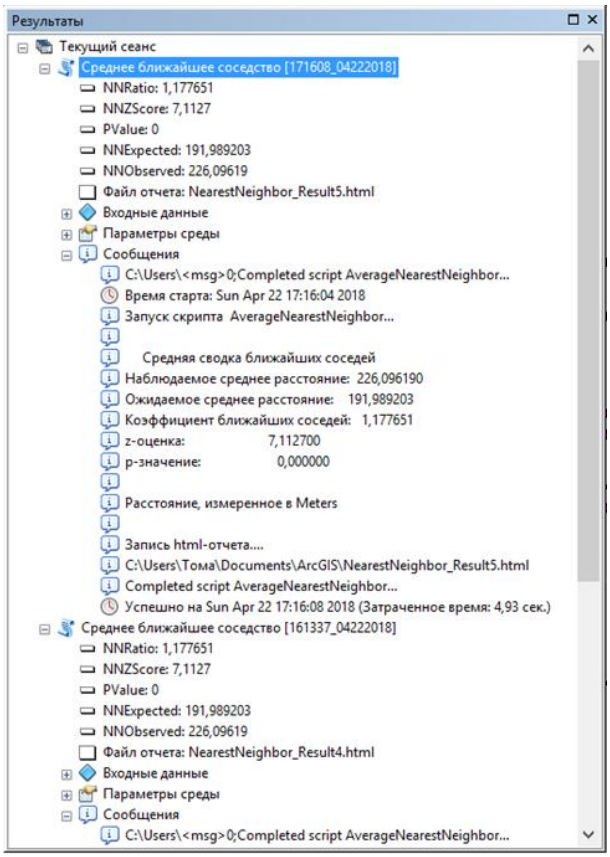
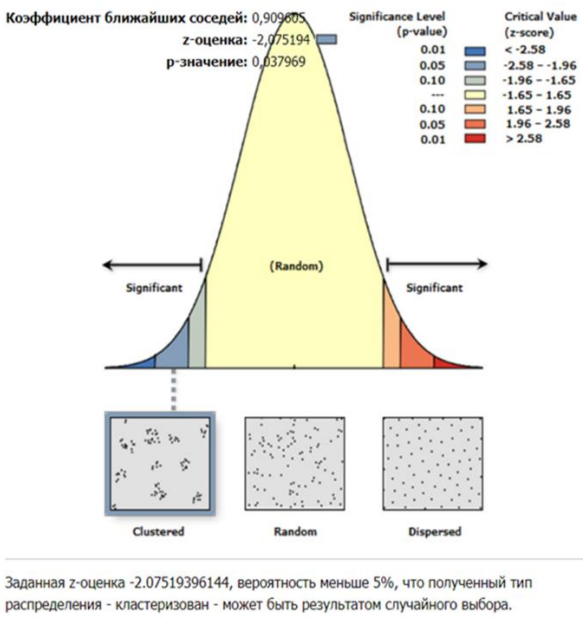
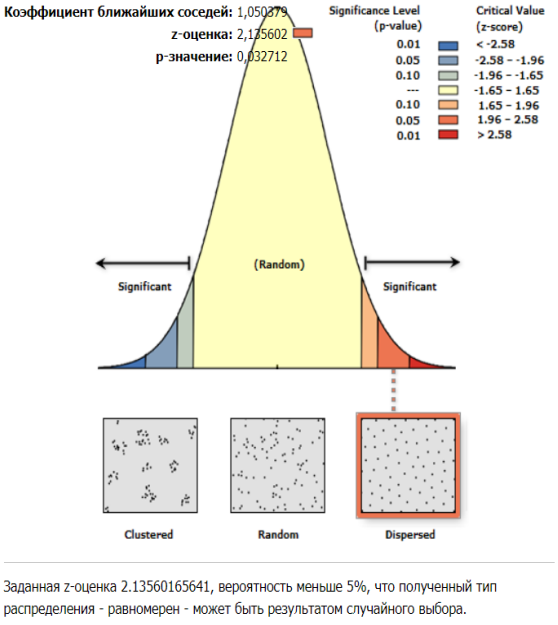


Рис. 8. **Текстовый отчет, сформированный по результатам выполнения поиска ближайшего соседства**

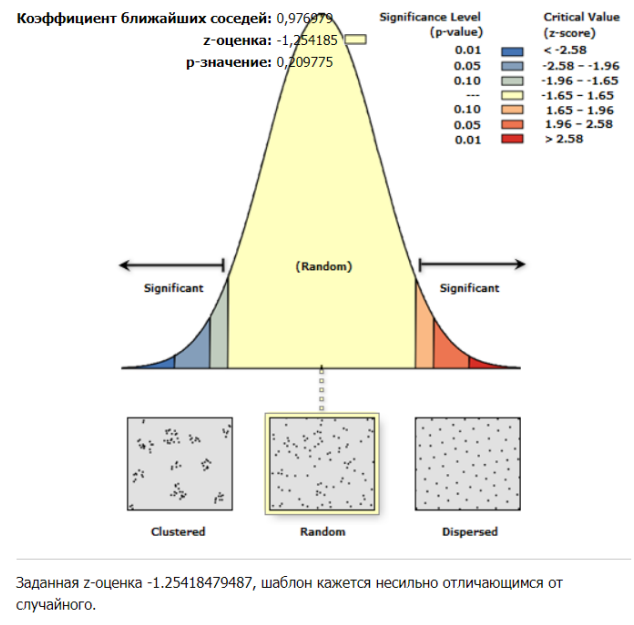
Индекс ближайшего соседства определяют, исходя из среднего расстояния от каждого объекта до ближайшего к нему соседнего объекта, при этом учитываются только координаты точек, а не значения атрибутов. Если индекс ближайшего соседства меньше единицы, данные распределены не случайно и в них имеются кластеризованные области (рис. 9а). Если данный показатель больше единицы, то данные распределены равномерно и явление кластеризации в них отсутствует (рис. 9б). Если индекс ближайшего соседства равен единице – распределение данных является случайным и нулевая гипотеза о том, что данные распределены случайно и пространственно не связаны не отвергается (рис. 9с).



а)



б)



В)

Рис. 9. **Пример графических отчетов, сформированных по результатам выполнения поиска ближайшего соседства**

3. Расчет пошаговой пространственной автокорреляции

Используя возможности модуля «Пространственная статистика», в частности опцию «Пошаговая пространственная автокорреляция», определить индекс ближайшего соседства для исходных данных по атрибутивному полю, указанному в индивидуальном задании, выполнив настройки, как показано на рис. 10.

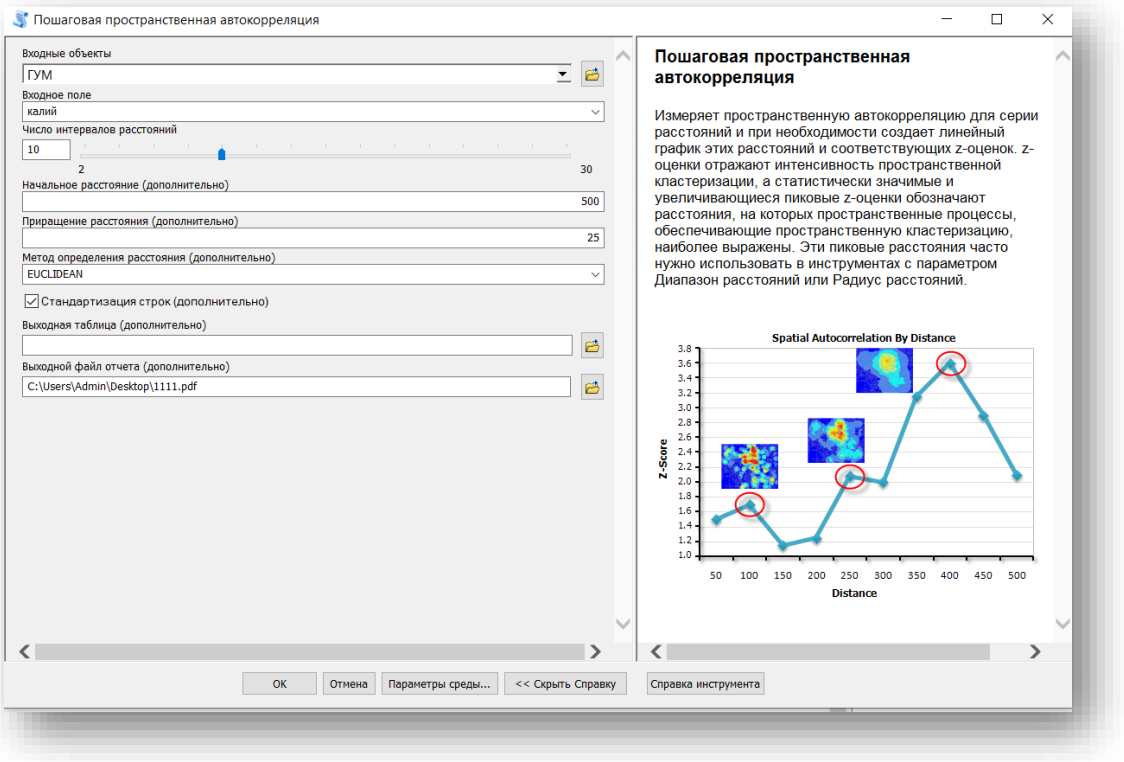


Рис. 10. **Диалоговое окно настроек опции «Пошаговая пространственная автокорреляция»**

При выполнении пошаговой пространственной автокорреляции выделяются десять интервалов расстояний, равномерно распределенных по всему экстенту. Для каждого интервала рассчитывается глобальный индекс Морана и интервал, для которого данный индекс будет наибольшим, рекомендуется как оптимальное расстояние для окрестности поиска. В результате получают граф, на котором отмечены минимальное и максимальное расстояния окрестности поиска ближайшего соседства. В данном примере это 600 и 1300 м (рис. 11).

Полученные результаты целесообразно использовать при подборе параметров модели вариограммы при моделировании пространственного распределения свойств почвы посредством интерполяции по методу кригинга, а также при установлении величины лага. Лаг – это расстояние, которое выбирается для поиска пар точек при расчете моментов второго порядка (вариограммы, ковариации). Именно величину лага следует учитывать при подборе шага в процессе создания оптимальной мониторинговой сети наблюдений за качественным состоянием земель.

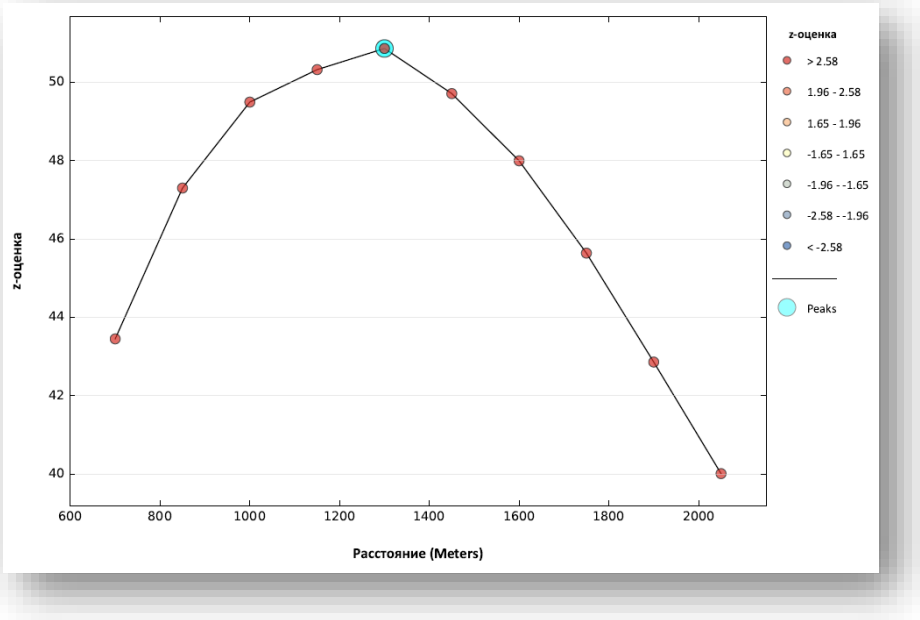
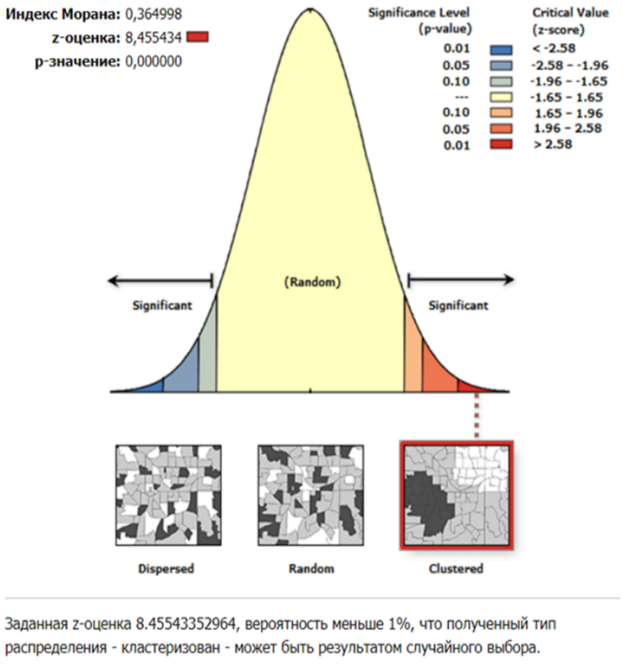


Рис. 11. **Графический отчет о результатах выполнении пошаговой пространственной автокорреляции**

4. Расчет величины глобального индекса Морана

Глобальный индекс Морана (рис. 12) позволяет установить факт наличия или отсутствия пространственной автокорреляции геоданных. Иными словами, он позволяет определить, существует ли в пределах исследуемой области кластеризация данных.



$$I = \frac{n \sum_{i=1}^{n} \sum_{j=i}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{[\sum_{i=1}^{n} \sum_{j=i}^{n} w_{ij}][\sum_{i=1}^{n} (y_i - \bar{y})^2]}$$

n – количество единиц в выборке;
w_{ji} – вес пространственной связи между i-й и j-й единицей выборки;
y_i – атрибутивное значение для i-й единицы выборки;
$\bar{y}$ – выборочное среднее значение атрибута.

Рис. 12. **Сущность глобального индекса Морана**

Используя возможности модуля «Пространственная статистика», в частности опцию «Пространственная автокорреляция», определить величину

глобального индекса Морана для исходных данных по атрибутивному полю, указанному в индивидуальном задании, выполнив настройки, как показано на рис. 13.

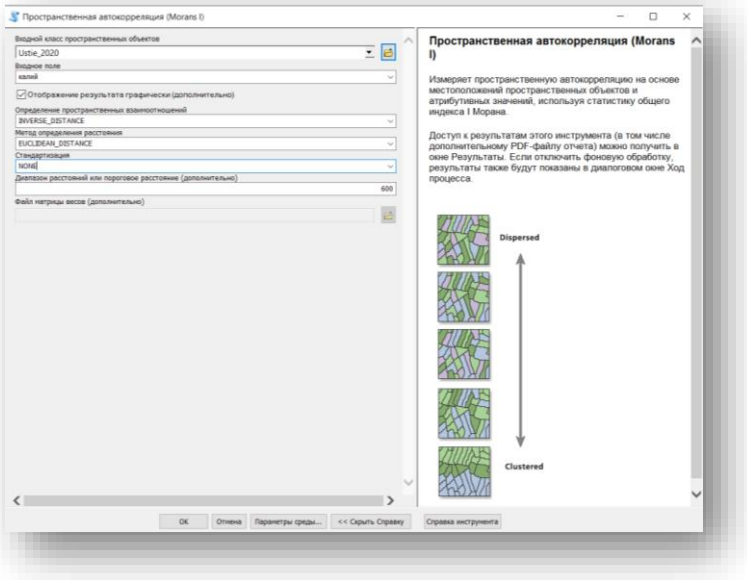
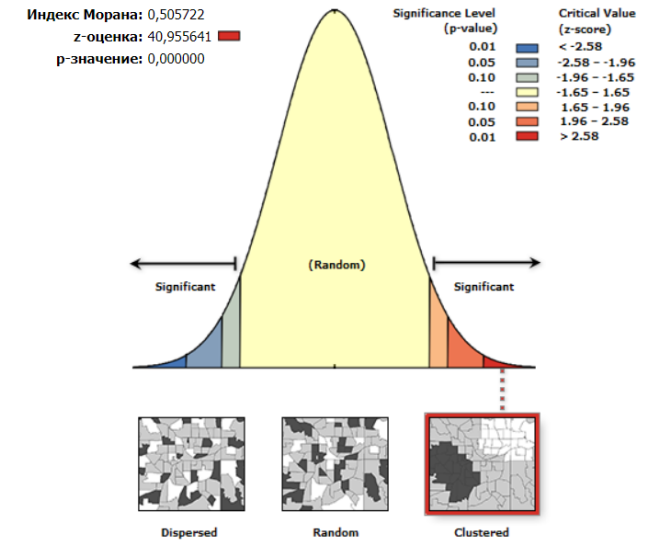


Рис. 13. **Диалоговое окно настроек опции «Пространственная автокорреляция»**

В результате реализации данной опции создается графический отчет с результатами расчетов (рис. 14).



Заданная z-оценка 40.9556414005, вероятность меньше 1%, что полученный тип распределения - кластеризован - может быть результатом случайного выбора.

Сводка глобального I Морана

Индекс Морана:	0,505722
Ожидаемый индекс:	-0,000775
Дисперсия:	0,000153
z-оценка:	40,955641
p-значение:	0,000000

Рис. 14. **Графический отчет о результатах расчета величины глобального индекса Морана**

Если величина глобального индекса Морана больше 1, то существует детерминированная прямая зависимость – группировка схожих (низких или высоких) значений геоданных, то есть их кластеризация. Если величина данного индекса 0, то данные распределены абсолютно случайно. Величина индекса меньше 1 означает детерминированную обратную зависимость – идеальное перемешивание низких и высоких значений, напоминающее шахматную доску, что свидетельствует о равномерном распределении геоданных.

5. Расчет величины глобального индекса Getis-Ord G

Глобальный индекс Getis-Ord G оценивает общую структуру и тренд данных и наиболее эффективен, когда данные распределены достаточно равномерно, но необходимо найти неожиданные всплески высоких значений в пространстве. Глобальный индекс Getis-Ord G – это статистический показатель, который означает, что результаты анализа интерпретируются в контексте нулевой гипотезы. Нулевая гипотеза для данного статистического показателя утверждает, что нет пространственной кластеризации в значениях объектов (рис. 15).

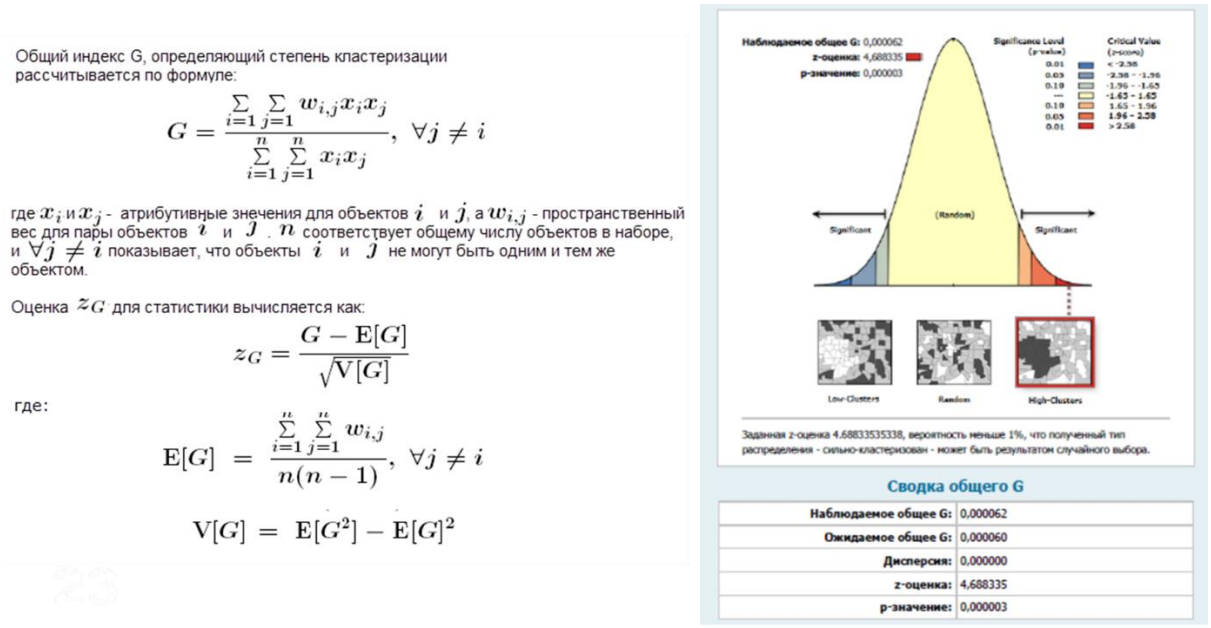


Рис. 15. **Сущность глобального индекса Getis-Ord G**

Используя возможности модуля «Пространственная статистика», в частности опцию «Высокая/низкая кластеризация», определить величину глобального индекса Getis-Ord G для исходных данных по атрибутивному полю, указанному в индивидуальном задании, выполнив настройки, как показано на рис. 16.

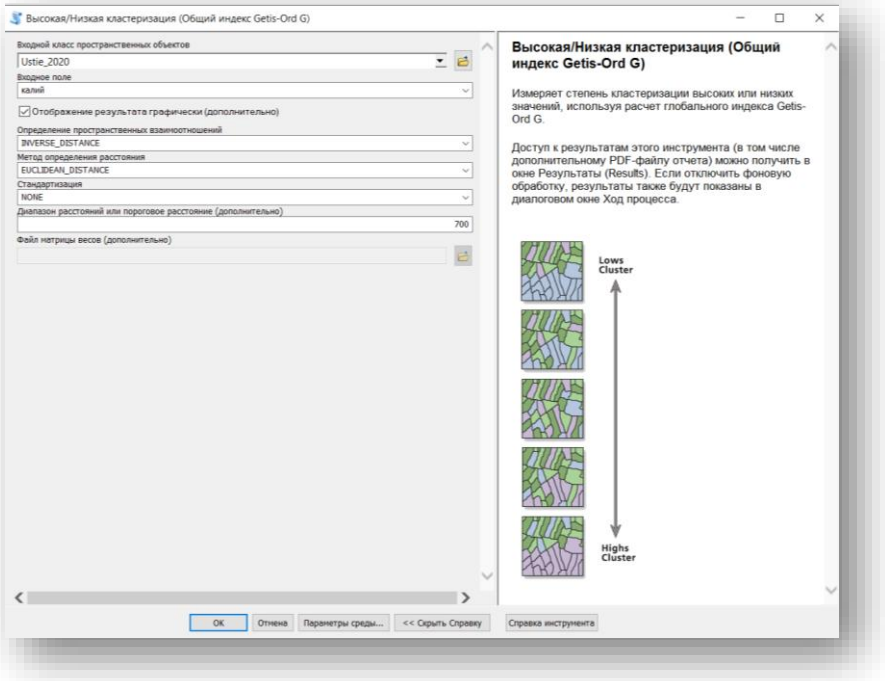
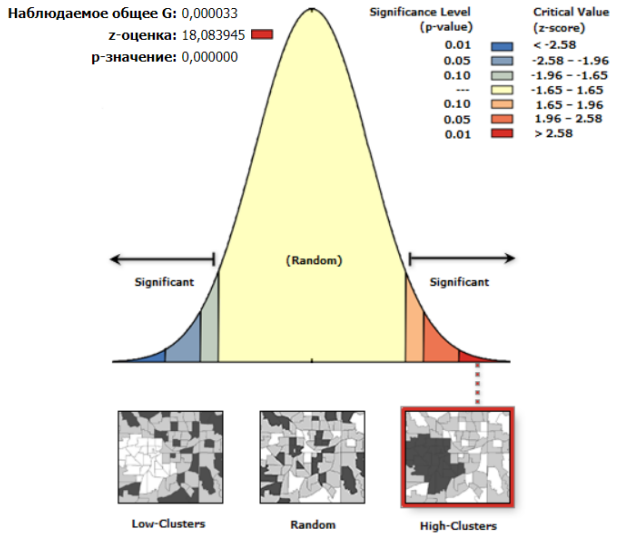


Рис. 16. **Диалоговое окно настроек опции «Высокая/низкая кластеризация»**

В результате реализации данной опции создается графический отчет с результатами расчетов (рис. 17).



Заданная z-оценка 18.0839447355, вероятность меньше 1%, что полученный тип распределения - сильно-кластеризован - может быть результатом случайного выбора.

Сводка общего G

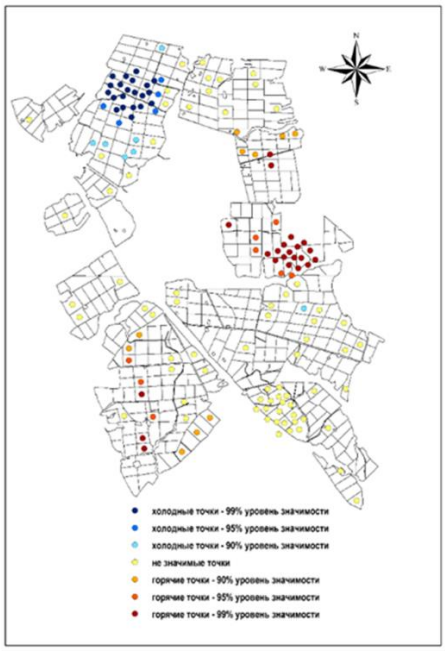
Наблюдаемое общее G:	0,000033
Ожидаемое общее G:	0,000030
Дисперсия:	0,000000
z-оценка:	18,083945
p-значение:	0,000000

Рис. 17. **Графический отчет о результатах расчета величины глобального индекса Getis-Ord G**

Если z-оценка при расчете индекса положительная, а наблюдаемый общий индекс G больше ожидаемого общего индекса G, имеет место кластеризация высоких атрибутивных значений в области изучения. Если z-оценка отрицательная, а наблюдаемый общий индекс G меньше ожидаемого общего индекса G, присутствует кластеризация низких атрибутивных значений в области изучения.

6. Выполнение анализа горячих точек

Целью анализа горячих точек является определение наличия у окрестности объекта статистически значимых отличий изучаемого атрибута от всей области значений. Если в окрестности объекта значение изучаемого атрибута выше, чем в изучаемой области, объект является «горячей точкой», если ниже – «холодной». Его выполняют посредством определения величины индекса Getis-OrdGi* – статистического показателя, рассчитываемого для каждого пространственного объекта в наборе данных. При расчете этого индекса учитываются не атрибутивные значения отдельных объектов, а атрибутивные значения их окрестностей, которые рассчитываются для каждого объекта и сравниваются со значениями в остальной области исследований (рис. 18).



$$Gi^* = \frac{\sum_{j=1}^{n} w_{i,j}x_j - \bar{X}\sum_{j=1}^{n} w_{i,j}}{S\sqrt{\frac{[n\sum_{j=1}^{n} w_{i,j}^2 - (\sum_{j=1}^{n} w_{i,j})^2]}{n-1}}}$$

где: x_j – атрибутивное значение объекта наблюдений;
j, $w_{i,j}$ – пространственный вес между объектами i и j;
n – общее число объектов;

$$\bar{X} = \frac{\sum_{j=1}^{n} x_i}{n}$$

где: $\bar{X}$ – выборочное среднее значение атрибута;

$$S = \sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - (\bar{X})^2}$$

где: S – стандартное отклонение от выборочного среднего значения.

Рис. 18. **Сущность анализа горячих точек**

Используя возможности модуля «Пространственная статистика», в частности опцию «Высокая/низкая кластеризация», определить величину глобального индекса Getis-Ord G для исходных данных по атрибутивному полю, указанному в индивидуальном задании, выполнив настройки, как показано на рис. 19.

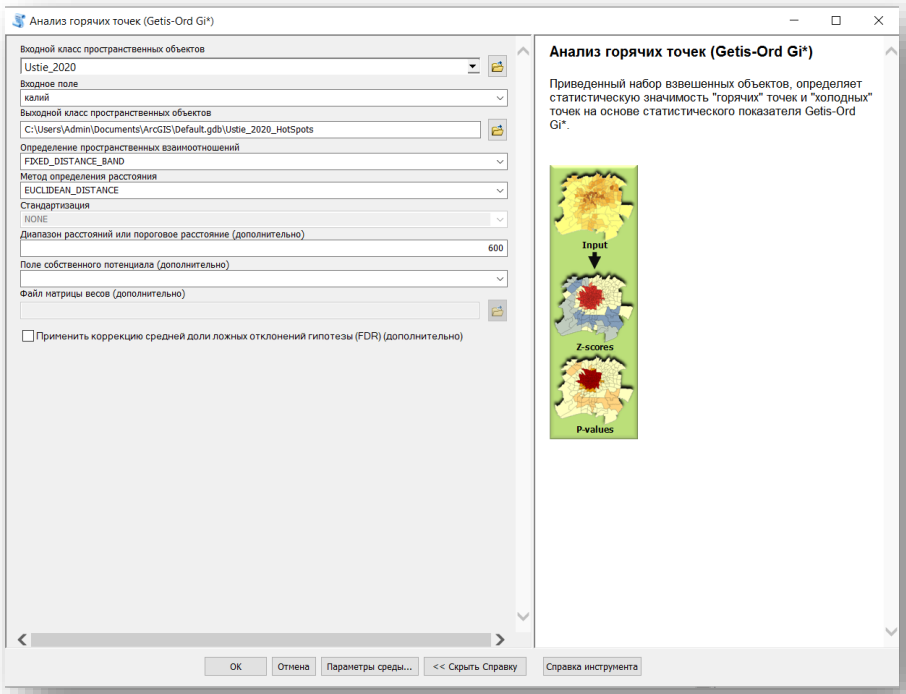


Рис. 19. **Диалоговое окно настроек опции «Анализ горячих точек»**

В результате реализации данной опции создается векторный слой с отображением статистически значимых кластеров высоких и низких значений (рис. 20).

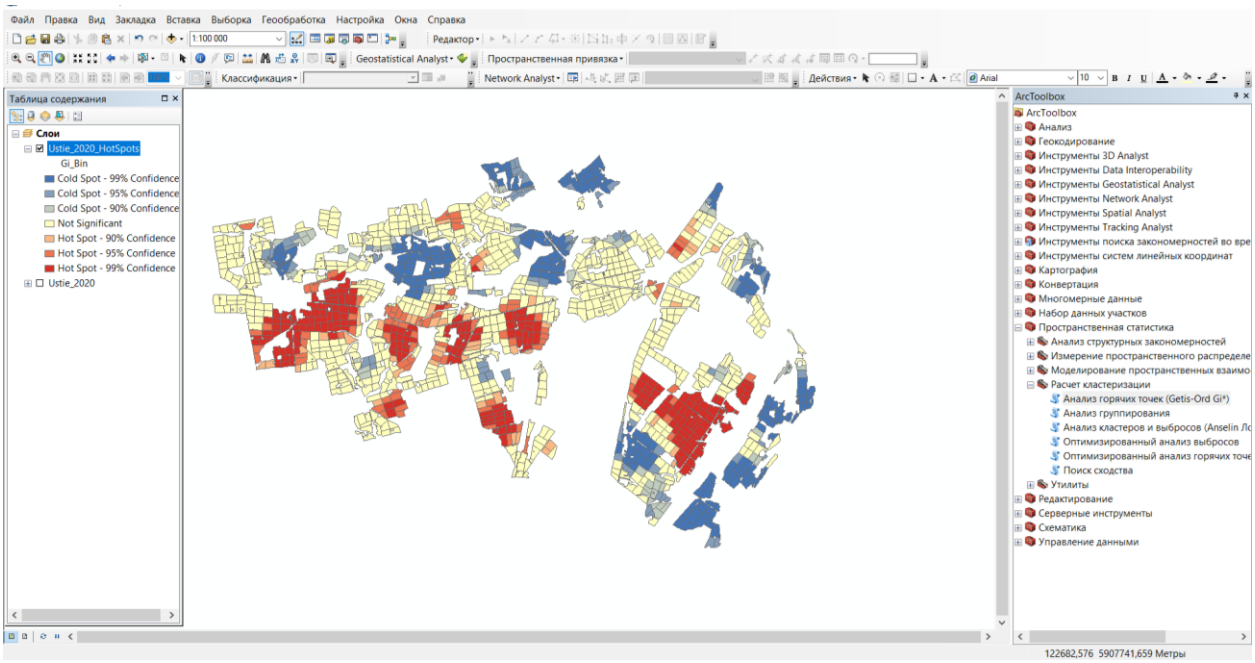
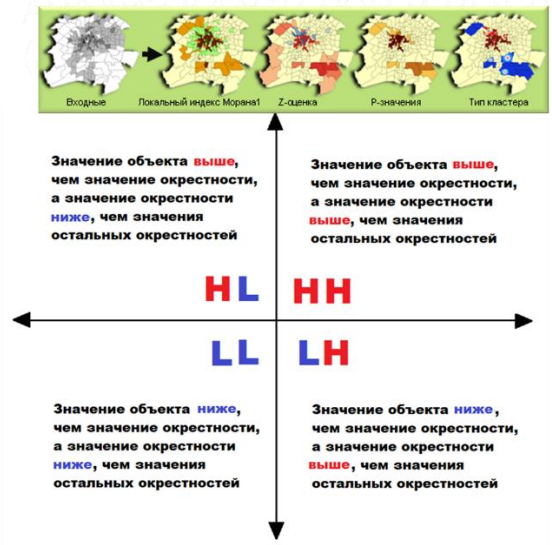


Рис. 20. **Рабочее окно проекта с отображением результатов выполнения анализа горячих точек**

7. Выполнение анализа кластеров и выбросов

Анализ кластеров и выбросов выполняют посредством определения величины локального индекса Морана. Данный вид геопространственного анализа позволяет идентифицировать концентрации высоких значений, концентрации низких значений и пространственные выбросы геопространственных данных (рис. 21).



$$Ii = \frac{x_i \overline{X}}{S_i^2}$$

$$\sum_{j=1, j\neq i}^{n} w_{i,j*}(x_j - \overline{X})$$

где: x_j – атрибутивное значение объекта наблюдений;

$\overline{X}$ – выборочное среднее значение атрибута;

$w_{i,j}$ – пространственный вес между объектами i и j;

$$S_i^2 = \frac{\sum_{j=1, j\neq i}^{n} (x_i - \overline{X})^2}{n-1}$$

где: n – общее число объектов.

Рис. 21. **Сущность анализа кластеров и выбросов**

Используя возможности модуля «Пространственная статистика», в частности опцию «Анализ кластеров и выбросов», определить величину глобального индекса Getis-Ord G для исходных данных по атрибутивному полю, указанному в индивидуальном задании, выполнив настройки, как показано на рис. 22.

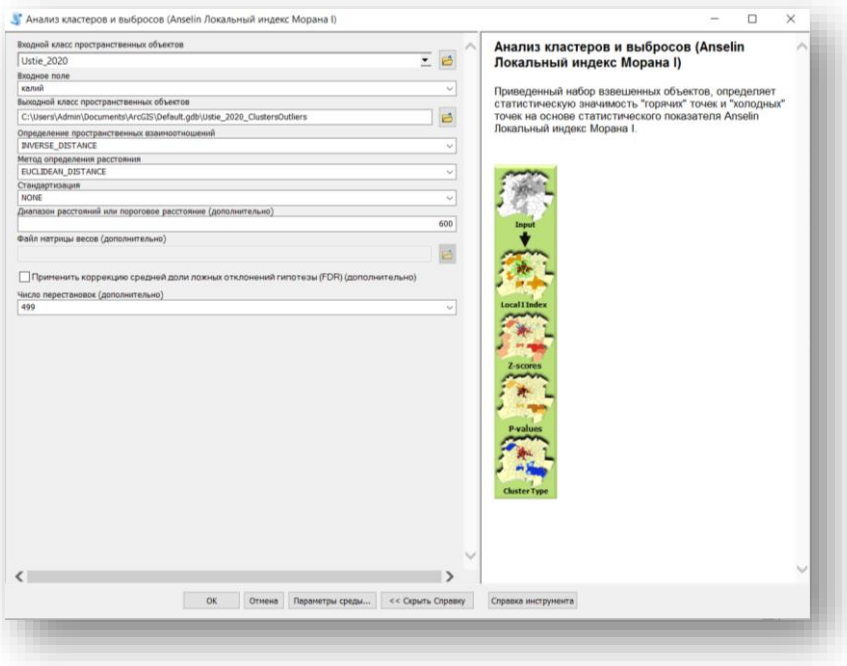


Рис. 22. **Диалоговое окно настроек опции «Анализ кластеров и выбросов»**

В результате реализации данной опции создается векторный слой с отображением статистически значимых кластеров высоких и низких значений, а также выбросов высоких и низких значений (рис. 23).

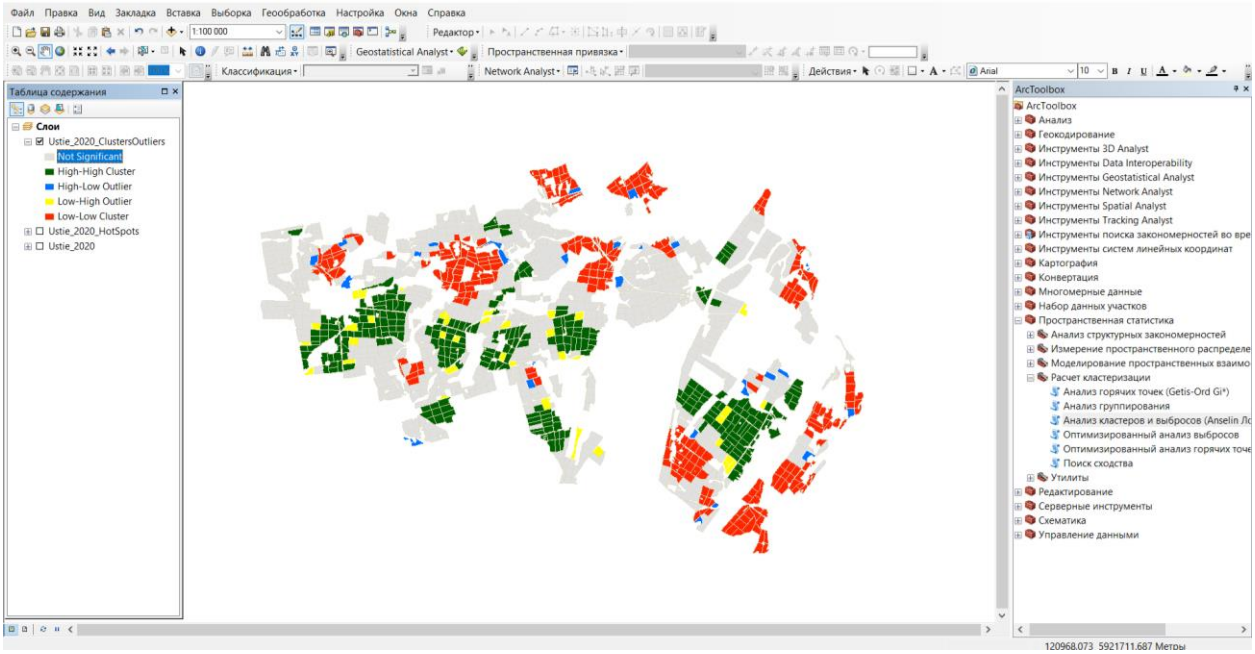


Рис. 23. **Рабочее окно проекта с отображением результатов выполнения анализа кластеров и выбросов**

Если в результате выполнения кластерного анализа установлено наличие кластеров типа HL и LH, то это свидетельствует о том, что в пределах исследуемой области фиксируются пространственные выбросы высоких (HL-кластеры) и низких (LH-кластеры) значений.

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.