ЛАБОРАТОРНАЯ РАБОТА №5

Исследование данных мониторинга земель с помощью набора графиков для исследовательского анализа (ESDA)

Цель работы: Изучить возможности исследовательского анализа данных заложенных в современных ГИС.

Задание. Построить исследовательские графики на основе имеющихся пространственных данных в среде ГИС ArcGIS

Вводные данные

Исследовательский анализ пространственных данных (Exploratory Spatial Data Analysis, ESDA) - это набор вычислительных методов и методов визуализации, используемых для анализа пространственных данных путем:

применения классической непространственной описательной статистики, обеспечивающей динамическую связь с картами ГИС и пространственными объектами;

выявления пространственных взаимодействий, отношений и закономерностей с использованием матрицы пространственных весов (определяемой соответствующим методом концептуализации), проверкой гипотез и различных показателей.

Методы и инструменты ESDA используются для:

описания и обобщения распределения пространственных данных;

визуализации пространственного распределения;

изучения пространственной автокорреляции (пространственных отношений и связей);

выявления пространственных выбросов;

выявления кластеров;

определения «горячих» и «холодных» точек.

Описательная статистика - это набор статистических процедур, обобщающих основные характеристики распределения посредством вычисления и построения:

распределения частот;

центра, рассеяния и формы (среднее, медиана и стандартное отклонение);

стандартной ошибки;

процентилей и квартилей;

выбросов;

коробчатых диаграмм;

нормального графика КК.

Статистика выводов - это раздел статистики, специализирующийся на анализе образцов для формирования выводов обо всей генеральной совокупности.

Пространственная статистика использует статистические методы для анализа пространственных данных, количественной оценки пространственных процессов, выявления скрытых или неожиданных закономерностей и моделирования данных в географическом контексте. Пространственная статистика в значительной степени основана на статистике выводов и проверке гипотез для анализа закономерностей, что позволяет точнее моделировать пространственные явления. В отличие от непространственных методов, пространственная статистика в своих математических формулах использует пространственные свойства, такие как местоположение, расстояние, площадь, длина и близость. Пространственная статистика количественно определяет и дополнительно отображает то, что человеческий глаз и разум интуитивно видят при чтении карты, отражающей пространственное расположение, распределение, процессы или тенденции .

Назначение описательной статистики и ESDA

Обычно первой задачей любого анализа является описание набора данных. Это позволяет быстро понять, как изменяются данные, и определить возможные ошибки (например, неприемлемые значения, пропуски [пустые ячейки] или выбросы [оценки, чрезмерно выделяющиеся на общем фоне]). Для описания набора данных мы используем описательную статистику (или «сводную статистику»). Вот типичные вопросы, на которые может ответить описательная статистика в географическом контексте: какова величина среднего дохода в данной области? Какой процент людей имеет высшее образование в данной области? Сколько клиентов конкретного кафе живут в пределах 10 минут ходьбы? Каковы их покупательная способность и величина стандартного отклонения в их доходах?

Описательная статистика применяется для вычисления конкретных характеристик (например, среднего или стандартного отклонения), что позволяет получить представление о распределении данных. Однако эти характеристики не отражают связей между результатами и пространственными объектами на карте. Главная особенность инструментов ESDA: они динамически связаны с картами в среде ГИС. Например, когда на диаграмме рассеяния выбирается некоторая точка, на соответствующей карте выделяется пространственный объект. Аналогично, при выборе пространственных объектов на карте выделяются соответствующие точки/области/полосы на графиках. В основе исследовательского анализа пространственных данных лежит понятие пространственной автокорреляции (см. главу 4), согласно которой более близкие пространственные объекты имеют более похожие значения (в одном или нескольких атрибутах). То есть ESDA предлагает более мощные возможности, обнаруживая закономерности в данных посредством картирования и статистической проверки гипотез (см. главу 3).

Сила исследовательского анализа пространственных данных (ESDA) покоится на двух основных его характеристиках (Dall'erba 2009; Haining et al. 2010, стр. 209):

извлекает знания благодаря возможностям интеллектуального анализа данных - информация, которую несут значения атрибутов, имеет отношение к местоположению. Это особенно важно в отсутствие теоретической основы, например в междисциплинарных областях социальных наук;

использует широкий спектр графических методов в сочетании с картированием, что делает анализ более доступным для людей, не привыкших к построению моделей.

Описательная статистика всегда используется в комбинации с инструментами ESDA. Иногда эти две области имеют размытые границы, по крайней мере это особенно верно в отношении простых инструментов. По этой причине в одних книгах гистограммы, диаграммы разброса или коробчатые диаграммы относятся авторами к описательной статистике, а в других к исследовательскому анализу пространственных данных. Различия между ними не имеют большого значения, пока человек понимает, как работает каждый инструмент. По сути, единственное отличие простых инструментов ESDA (например, гистограмм, диаграмм рассеяния, коробчатых диаграмм) состоит в возможности связывать графики с пространственными объектами, что способствует их использованию в исследовательском анализе. Вообще говоря, простые инструменты ESDA можно использовать до этапа моделирования, а расширенные инструменты ESDA - на этапе построения модели для выявления пространственных отношений и скрытых закономерностей в пространственных данных.

Назначение пространственной статистики

Пространственную статистику можно рассматривать как часть коллекции методов пространственного анализа, таких как исследовательский анализ пространственных данных (ESDA), пространственный анализ точечных закономерностей, пространственная кластеризация и пространственная эконометрика. Пространственная статистика в основном используется для:

анализа географического распределения с помощью центрографических измерений (см. главу 3). По аналогии с описательной статистикой, есть возможность исследовать географическое распределение, чтобы найти средний центр и стандартное расстояние. Вычисления в пространственной статистике производятся на основе местоположения каждого объекта; в этом основное их отличие от однородной описательной статистики, которая рассматривает исключительно непространственные атрибуты пространственных объектов. Даже притом что пространственная статистика, опирающаяся на анализ географического распределения, позволяет осуществлять взвешивание с использованием значений атрибутов, ее результаты относятся к пространственным измерениям. Обычно она использует пространственные объекты, являющиеся точками и полигонами (центроиды);

анализа пространственных закономерностей. Пространственная статистика может использоваться для анализа закономерностей распределения объектов в пространстве. Когда в вычислениях участвуют точечные объекты, анализ называется анализом точечных массивов. В ходе такого анализа определяется, является ли расположение точек случайным, сгруппированным (кластеризованным) или рассредоточенным. Анализ пространственных закономерностей, в котором исследуются значения пространственных атрибутов (признаков), является частью анализа пространственной автокорреляции, рассматриваемого в главе 4;

определения пространственной автокорреляции, горячих точек и выбросов;

выполнения пространственной кластеризации;

моделирования пространственных отношений. Пространственная статистика также может использоваться для выявления ассоциаций и отношений между атрибутами и пространством; в качестве примеров можно назвать методы пространственной регрессии и пространственные эконометрические модели;

анализа пространственно-непрерывных переменных, таких как температура, уровень загрязнения, мощность почвенно-растительного слоя и т. д. Вообще говоря, пространственный статистический анализ пространственно-непрерывных (полевых) переменных называется геостатистикой Геостатистика сосредоточивается на описании пространственных вариаций в наборе наблюдаемых значений и их интерполяции/экстраполяции .

Пространственная статистика основана на статистических понятиях, но включает определение местоположения в терминах географических координат, расстояния и площади. Она дополняет классические статистические измерения и процедуры и предлагает расширенные возможности анализа данных. В географическом анализе классическая и пространственная статистики используются вместе, дополняя друг друга. Однако между классической и пространственной статистиками есть принципиальная разница. В классической статистике мы делаем следующее предположение относительно выборки: это набор независимых наблюдений, которые следуют определенному распределению, обычно нормальному. В пространственной статистике, напротив, из-за присущей пространственной зависимости и существования (обычно) пространственной автокорреляции основное внимание уделяется методам выявления и описания этих зависимостей и корреляций. Иначе говоря, в классической статистике наблюдения должны быть независимыми, тогда как в пространственной статистике между наблюдениями обычно имеется пространственная зависимость. Классическая статистика должна быть изменена, чтобы учесть это условие.

Выполнение работы

Используемые инструменты ArcGIS: Choropleth map (Фоновая картограмма), Histogram (Гистограмма), Normal 00 plot (Нормальный график КК), Boxplots (Коробчатая диаграмма), Z-score (Z-оценки).

5.1. Создание фоновой картограммы доходов.

Перейдите в папку, где вы сохранили набор данных, сопровождающий книгу, и щелкните на файле Lab2_SimpleESDA.mxd

Главное меню > **File** (Файл) > **Save As** (Сохранить как) > My_Lab2_SimpleES- SA.mxd

В папку I:\BookLabs\Lab2\Output

TOC (Таблица содержания) **> RC** (щелчок правой кнопкой) на слое **City > Properties** (Свойства) > **TAB** = **Symbology** (Символы) > **Quantities** (Количество) > **Graduated colors** (Градуированные цвета)

Value (Значение) = **Income** (рис. 5.1)

Color Ramp (Цветовая схема) = от желтого до коричневого

Classes (Классов) = 4 > щелкните на кнопке **Classify** (Классифицировать) > **Break Values** (Граничные значения) > 15000 > **Enter** > 20000 > **Enter** > 25000 > **Enter** > 40000 > OK

RC Label (Подпись) > **Format Labels** (Формат подписей) > **Numeric** (Числовой) > **Rounding** (Округление) > **Number of decimal places** (Число десятичных знаков) = 2 > OK

TOC (Таблица содержания) > **RC** на **City** > **Properties** (Свойства) > **Save As Layer File** (Сохранить как файл слоя > **Name** (Имя) = Income.lyr В папку I:\BookLabs\Labl\Output

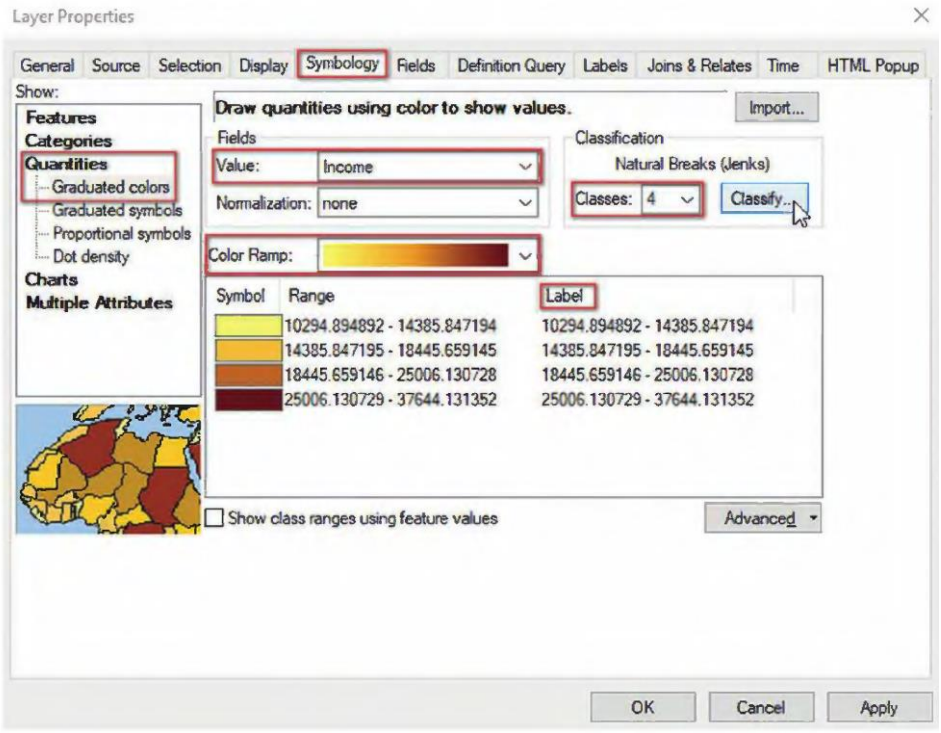

Рис. 5.1. Диалог Layer properties (Свойства слоя) с настройками картограммы для отображения доходов

Совет: сохранение слоя **City** в файле позволяет сохранить представление дохода. Имейте в виду, что после добавления слоя в таблицу содержимого он получает имя исходного шейп-файла (например, **City,** как в

данном примере), а не имя, под которым он был сохранен (например, Citylncome.lyr).

Интерпретация результатов: в результате действий, перечисленных выше, будет создано четыре группы округов со средним годовым доходом: (а) менее 15 000 в год (округа с низким доходом), (б) от 15 000 до 20 000 (округа со средним доходом), (в) от 20 000 до 25 000 (округа с доходом выше среднего) и (г) группа с доходом более 25 000 (округа с высоким доходом; см. рис. 5.2). На карте видно, что округа с высоким доходом расположены в центре (окрашены в темно-коричневый цвет). Большинство округов с уровнем доходов выше среднего или высоким расположены в деловой части города (полигон, очерченный красной границей). Низким доходом характеризуются округа на севере, западе и юге. Такое представление переменной «доход» помогает понять, как доход распределяется в пространстве (пространственное распределение значений) по городским округам. Чтобы проанализировать, как значения дохода распределяются по отношению к среднему (в данном случае не пространственно), можно построить гистограмму частотного распределения.

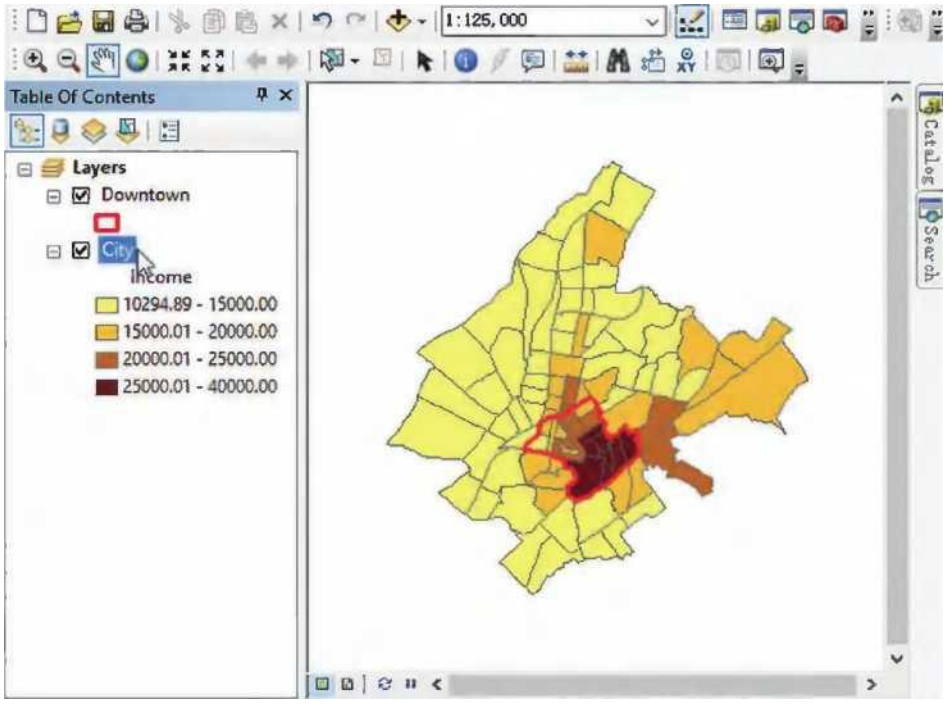

Рис. 5.2. Картограмма распределения доходов, классифицированных в четыре группы

Главное меню > **Selection** (Выборка) > **Select By Attributes** (Выбрать по атрибуту)

Layer (Слой) = **City**

SELECT * FROM City Where: “Income” >= 20000 **OK**

TOC (Таблица содержания) > **RC** на **City** > **Data** (Данные) > **Export Data** (Экспортировать данные) > Output feature class (Выходной класс объектов) =

I:\ BookLabs\Lab2\Output\Cl_HighIncome.shp

Главное меню > **Selection** (Выборка) > **Clear Selected Features** (Очистить выбранные объекты)

Добавьте шейп-файл в окно просмотра данных для проверки вывода, но потом, перед тем как продолжить упражнение, удалите его.

Стр. создание гистограммы.

Главное меню > **Customize** (Настройка) > **Extensions** (Дополнительные модули) > установите флажок **Geostatistical Analyst** (Геостатистический анализ) > **Close** (Закрыть) (если флажок **Geostatistical Analyst** (Геостатистический анализ) установлен, то не снимайте его)

Главное меню > **Customize** (Настройка) > **Toolbars** (Панели инструментов) > выберите **Geostatistical Analyst** (Геостатистический анализ В панели инструментов выберите **Geostatistical Analyst** (Геостатистический анализ; см. рис. 5.3) **Explore Data** (Исследовать данные) > **Histogram** (Гистограмма)

Layer (Слой) = **City** (см. рис. 2.20)

Attribute (Атрибут) = **Income**

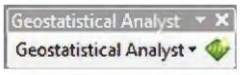

Рис 5.3 Панель инструментов геостатистического анализа

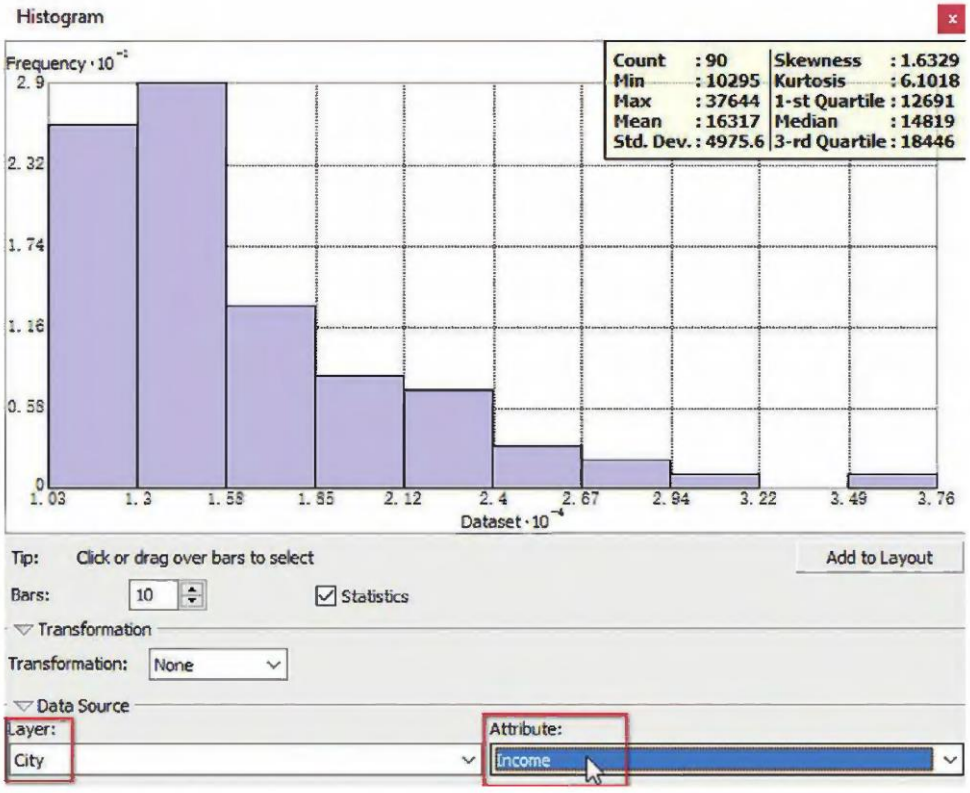

Рис. 5.4. Гистограмма частотного распределения доходов

Перетащите гистограмму в левый нижний угол карты, как показано на рис. 5.5.

TOC (Таблица содержания) > **RC** на **City** > **Open Attribute Table** (Открыть таблицу атрибутов)

Выберите столбик на гистограмме и посмотрите, какие городские округа на карте будут подсвечены.

В окне с гистограммой щелкните на кнопке **Add to Layout** (Добавить в компоновку) > вернитесь в **Data** View (Вид данных) > закройте гистограмму

Главное меню > **Selection** (Выборка) > **Clear Selected Features** (Очистить выбранные объекты) > закройте таблицу атрибутов для слоя **City** >

Главное меню > **File** (Файл) > **Save** (Сохранить)

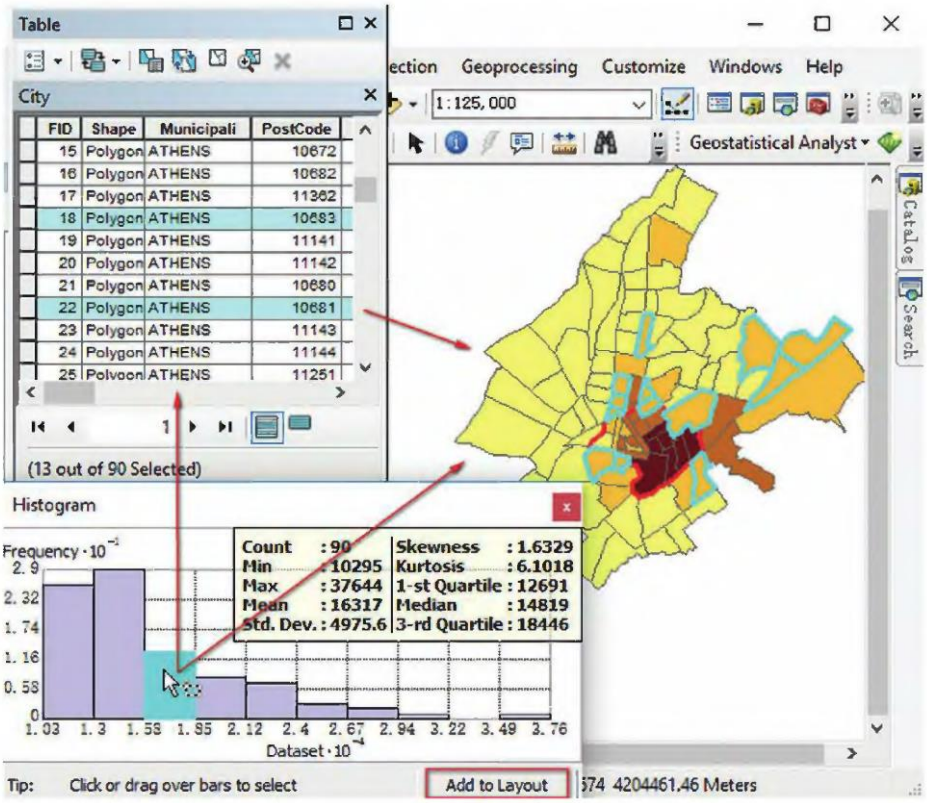

Рис. 5.5. Визуальные возможности инструментов ESDA.

При выборе столбика на гистограмме выделяются соответствующие полигоны на карте и строки в таблицеатрибутов City

Интерпретация результатов: гистограмма отображает основные описательные статистики дохода (например, среднее значение, стандартное отклонение, асимметрию, эксцесс, первый и третий квартили; см. рис. 2.20). Как можно видеть, распределение доходов имеет положительную асимметрию и значительно отличается от нормального распределения. Чтобы убедиться, что распределение доходов действительно отлично от нормального, необходимо построить нормальный график КК (см. рис. 5.6).

Гистограмма связана с полигонами в шейп-файле и таблицей атрибутов слоя, что является одним из преимуществ использования инструментов ESDA в пространственном анализе. Если выбрать столбик на гистограмме, то автоматически будут выделены соответствующие полигоны на карте и строки в таблице атрибутов (см. рис. 5.5). Аналогично, если выбрать полигон на карте

или строку в таблице атрибутов, выделится соответствующий столбик на гистограмме. На получившейся гистограмме можно видеть отдельно стоящий столбик справа, который следует дополнительно проанализировать, потому что такое его расположение может говорить о наличии выбросов (округов с чрезвычайно большими значениями дохода)

Выброс > Среднее + 2,5 **x** Стандартное отклонение = 16 317 + 2,5 **x** 4975.6 = 28 756.5,
или меньше, чем
Выброс > Среднее - 2,5 **x** Стандартное отклонение = 3878.

Отсортировав таблицу атрибутов слоя **City** (Город) по столбцу Income (Доход), можно увидеть, существуют ли округа с доходами выше или ниже этих значений. Фактически у нас имеется два округа с доходами выше 28 756,5, которые можно отметить как выбросы.
Альтернативный способ проверки наличия выбросов - построение коробчатых диаграмм и визуализация z-значений.

5.2. Создание нормального графика КК для определения соответствия или несоответствия распределения доходов нормальному распределению.

Toolbar (Панель инструментов) = **Geostatistical Analyst** (Геостатистический анализ) > **Geostatistical Analyst** > **Explore Data** (Исследовать данные) > **Normal** 00 **Plot** (Нормальный график КК)

Выберите слой **City** (см. рис. 5.6)

Attribute (Атрибут) = **Income**

Выберите точку вверху справа на графике и посмотрите, какие полигоны выделятся (см. рис. 5.7).

В нормальном графике КК щелкните на кнопке Add **to Layout** (Добавить в компоновку) > перетащите график, расположив его рядом с гистограммой > вернитесь в **Data** View (Вид данных) > закройте гистограмму

Главное меню > **Selection** (Выборка) > **Clear Selected Features** (Очистить выбранные объекты)

Интерпретация результатов: нормальный график КК показывает, что значения дохода отклоняются от прямой линии нормального распределения (см. рис. 5.6). То есть можно утверждать, что распределение переменной дохода отличается от нормального. Выбрав точку в правом верхнем углу графика, можно увидеть, какие округа значительно отклоняются от линии ожидаемого значения дохода, если бы распределение было нормальным (см. рис. 5.7). Эти округа принадлежат к группе с высоким доходом и сконцентрированы в деловом районе города, что является интересным открытием с точки зрения пространственного анализа

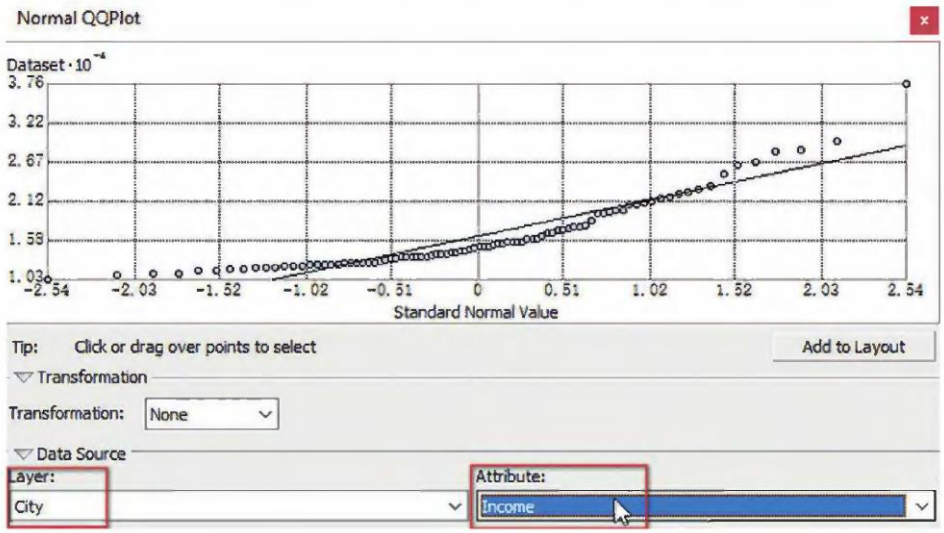

Рис 5.6. Нормальный график КК для переменной дохода

Округа с экстремально высокими значениями дохода (пространственные выбросы) можно определить с помощью стандартного отклонения, например выбросами можно считать наблюдения, отстоящие от среднего более чем на 2,5 стандартных отклонения (см. раздел 2.2.7). Стандартное отклонение и среднее значение дохода составляют 4975,6 и 16 317 соответственно (см. рис. 2.21). То есть значение дохода считается выбросом, если оно больше, чем
Выброс > Среднее + 2,5 **x** Стандартное отклонение =16317 + 2,5 **x**4975.6 = 28 756.5

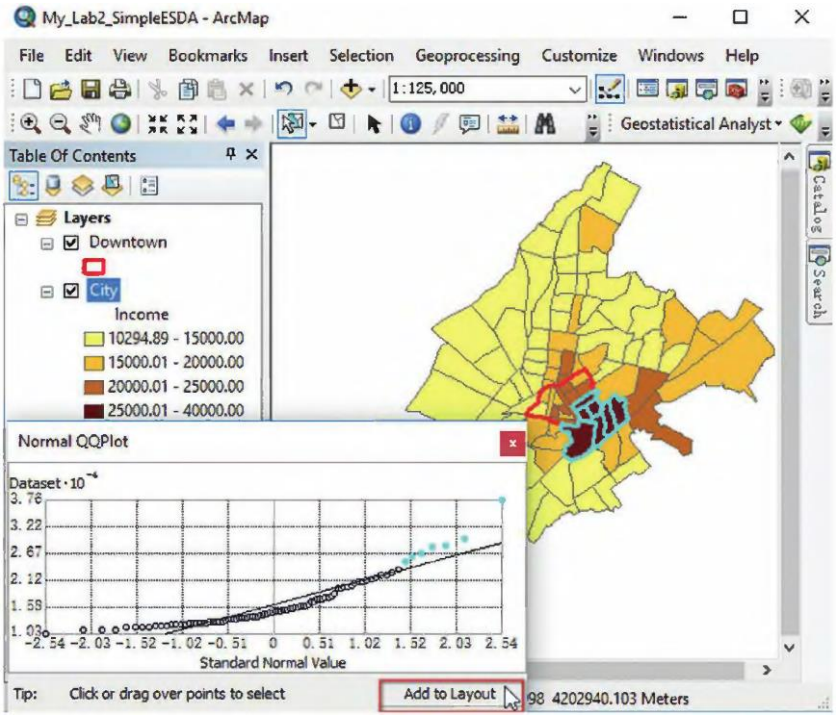

Рис. 5.7. Выявление округов с высоким доходом (соответствуют крайней правой точке на нормальном графике КК), которые существенно отклоняются от линии нормального распределения

5.3. Создание коробчатой диаграммы.

Главное меню > **View** (Вид) > **Graphs** (Диаграммы) > **Create Graph** (Построить диаграмму) >

Graph type (Тип диаграммы) = **Box Plot** (Коробчатая) (см. рис. 5.8) **Layer/Table** (Слой/Таблица) = **City**

Value field (Поле значений) = **Income** > **Next** (Далее) > **Title** (Заголовок) = **Income** > **Finish** (Готово)

Выберите экстремальный выброс на диаграмме (отмечен символом *) и посмотрите, какие округа выделятся на карте (см. рис. 5.9).

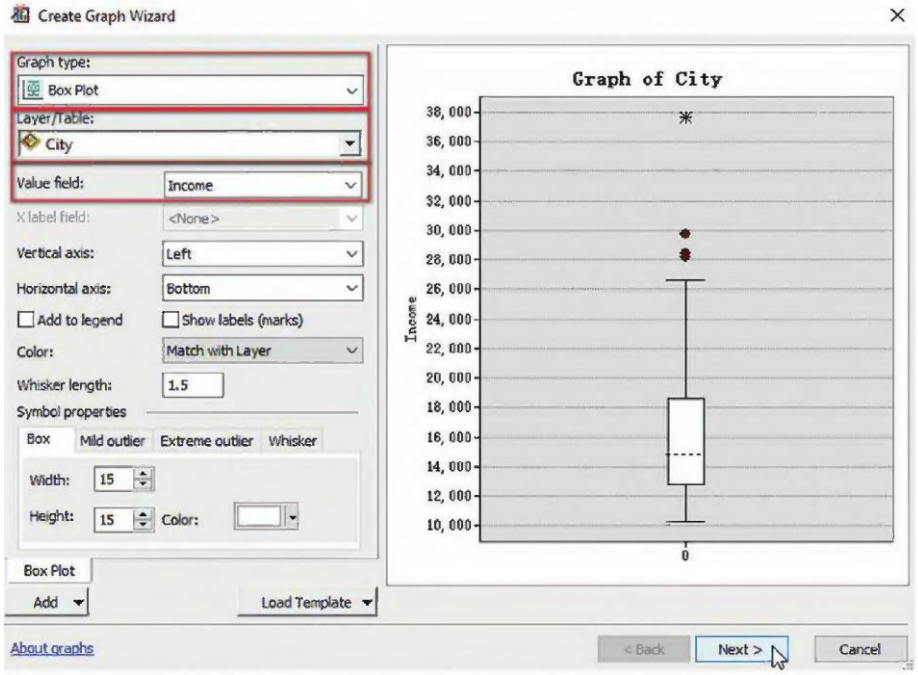

Рис. 5.8. Диалог создания коробчатой диаграммы

RC на коробчатой диаграмме > **Add to Layout** (Добавить в компоновку) > перетащите график в свободное место > вернитесь в **Data View** (Вид данных) > закройте диаграмму > главное меню > **Selection** (Выборка) > **Clear Selected Features** (Очистить выбранные объекты)

Интерпретация результатов: на коробчатой диаграмме можно видеть три умеренных выброса (обозначены точкой) и один экстремальный выброс. В верхней части диаграммы находятся умеренные и экстремальные выбросы, соответствующие высоким значениям дохода. Выбросы, соответствующие низкому доходу, не наблюдаются. Выбрав точку экстремального выброса на диаграмме, можно выделить соответствующий округ в центре города на карте (см. рис. 5.9). Чтобы решить, как поступить с экстремальным выбросом, сначала проверим, не является ли это наблюдение следствием человеческой ошибки. В данном случае значение верное. Затем оценим, насколько такое значение является разумным, опираясь на (а) наши знания в конкретной области (если таковые имеются) и (б) здравый смысл. Как мы знаем, в большинстве регионов мира имеет место неравенство доходов, поэтому районы со значительно более высоким среднедушевым доходом не являются

редкостью. Наш набор данных наглядно демонстрирует это.

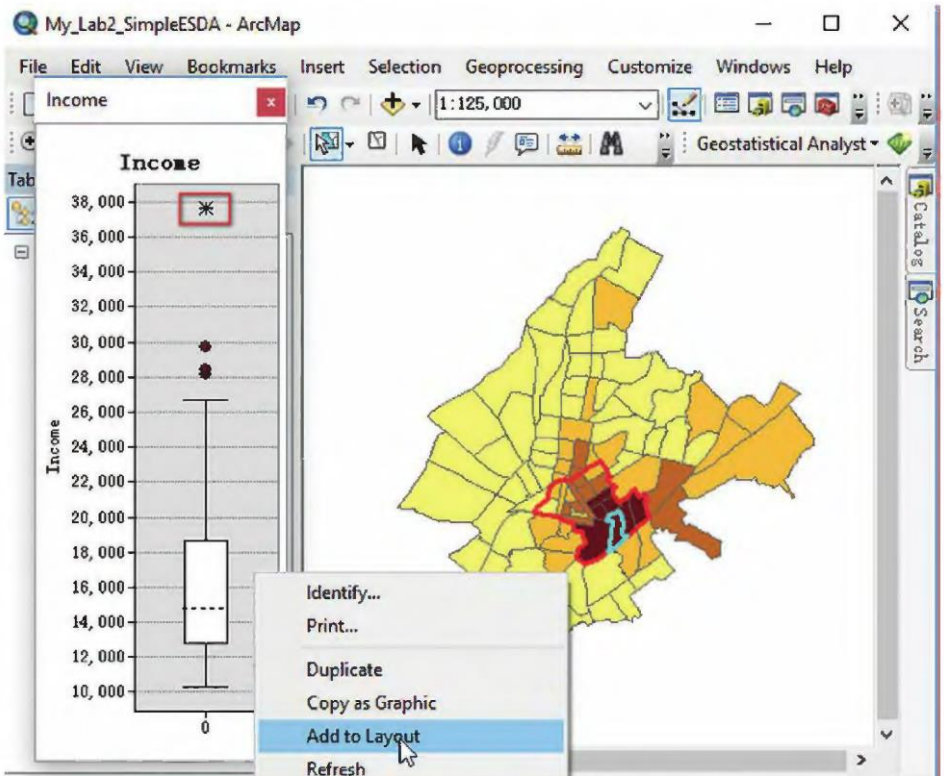

Рис. 5.9. Выберите экстремальный выброс на коробчатой диаграмме, чтобы отыскать соответствующий ему округ на карте.

Все графики сохраняются в представлении компоновки в .mxd

Отметим, что область, в которой сосредоточены округа с высоким доходом (включая выбросы), находится рядом с центральной городской площадью (Конституционная площадь), где расположен национальный парламент. Площадь окружена множеством красивых неоклассических зданий, рядом находятся городской парк, археологический памятник Акрополь и жилой район, который исторически привлекал слои населения с более высокими доходами. Кажется вполне разумным, что именно в этой области сосредоточены округа, где проживают люди с высокими и чрезвычайно высокими доходами. По этой причине и в контексте конкретного проекта (анализ рынка кафе) мы должны оставить выбросы в наборе данных, потому что они несут ценную для нас информацию. Однако в другом контексте может быть предпочтительнее удалить выбросы. Например, если бы нас интересовал социальный анализ округов со средним и низким доходами, то удаление экстремального выброса позволило бы получить более реалистичные результаты. Например, средний доход по городу уменьшится, что лучше отражает реальное экономическое положение большинства граждан. Стандартное отклонение и другие описательные статистики (например, доверительные интервалы, квартили) тоже получат другие значения. Иначе говоря, удаление выброса позволит точнее отразить социально-экономический профиль жителей. С точки зрения социального анализа наличие выбросов в нашем тематическом исследовании свидетельствует о

большом неравенстве доходов, что является важным открытием.

5.4. вычисление и отображение z-оценок доходов; выявление выбросов.

Другой способ анализа распределения доходов - вычислить и отобразить z-оценку. В этом случае отображаются отклонения дохода в каждом округе от среднего значения, и появляется возможность выявить округа со схожими или разными значениями среднего дохода, что также позволяет выявлять пространственные выбросы.

ТОС (Таблица содержания) > **RC** на **City** > **Open Attribute Table** (Открыть таблицу атрибутов)

Щелкните на кнопке **Table Options** (Опции таблицы) > **Add Field** (Добавить поле) >

Name (Имя) = **IncZScore Type**
(Тип) = **Float**

Precision (Точность) = **5** (число цифр по обе стороны от десятичной запятой) **Scale** (Масштаб) = 3 (число цифр после запятой)

ОК

RC на столбце **IncZScore > Field Calculator** (Калькулятор поля) > (щелкните на кнопке **YES** (ДА), если появится всплывающее сообщение, указывающее, что вычисление выполняется вне сеанса правки) > в поле **IncZScore** введите **([Income]** -16317)/4975.6 **> ОК >** закройте таблицу (см. рис. 5.10)

См. упражнение 1.1: среднее = 16 317, стандартное отклонение = 4975.6

ArcToolbox > **Spatial Statistics** (Пространственная статистика) > **Rendering** (Отображение) > **ZScore Rendering** (Отображение z-оценок; см. рис. 5.11) **Input Feature Class** (Класс входных объектов) = **City Field to Render** (Поле для отображения) = **IncZScore**

Output Layer File (Выходной файл слоя) = I:\BookLabs\Lab2\Output\Inc-ZScore.lyr

OK

Далее определите округа, где наблюдаются выбросы, т. е. со значением IncZScore больше 2,5.

Главное меню > **Selection** (Выборка) > **Select By Attributes** (Выбрать по атрибуту) >

Layer (Слой) = **IncZScore** (см. Рис.5.12)

Method (Метод) = **Create a new selection** (Создать новую выборку)

DC на «IncZScore»

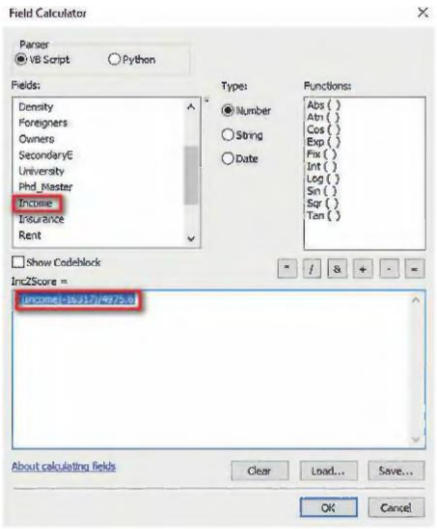

Рис. 5.10. Вычисление z-оценки для поля Income

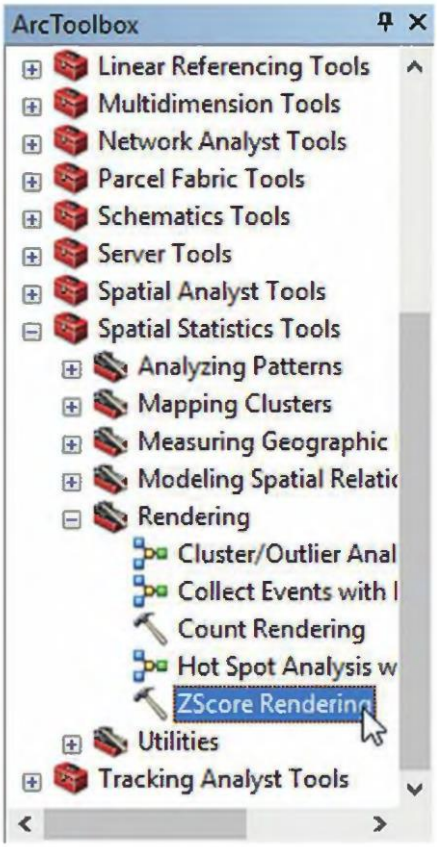

Рис. 5.11. Выбор z-оценки для отображения

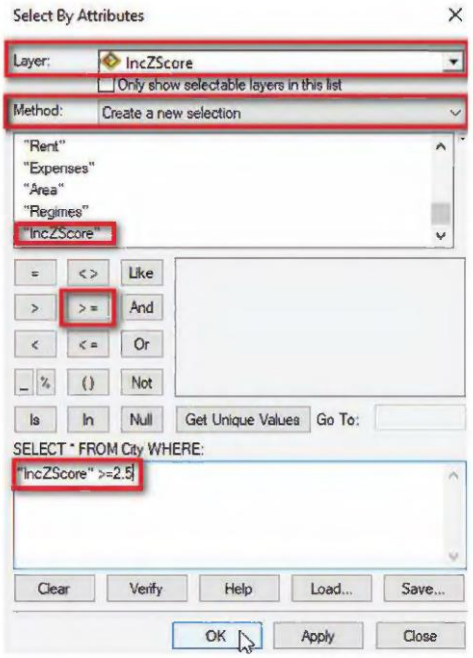

Рис 5. 12 Диалог выбора атрибутов

Перейдите в поле **SELECT * FROM City WHERE** и введите: **“lncZScore”>=2.5 OK**

Главное меню > **Selection** (Выборка) > **Clear Selected Features** (Очистить выбранные объекты)

Главное меню > **File** (Файл) > **Save** (Сохранить)

Интерпретация результатов: на карте (рис. 5.13) показаны z-оценки дохода для всех округов. Чем выше или ниже z-оценка, тем больше разница между средним доходом в округе и средним доходом во всей исследуемой области. Средний годовой доход в округах, выделенных красным и сосредоточенных в центре, отличается от среднего дохода по городу более чем на два стандартных отклонения. Разница более чем в 2,5 стандартных отклонения может говорить о потенциальных пространственных выбросах. Исходя из этого определения, два округа с высоким уровнем дохода можно обозначить как пространственные выбросы (выделены на карте голубым контуром). Выбросов в сторону крайне низких доходов не выявлено. Очевидно, что разные определения выбросов приводят к несколько разным результатам (см. рис. 2.25). Сколько и какие выбросы в конечном итоге будут оставлены в наборе данных, зависит от анализа

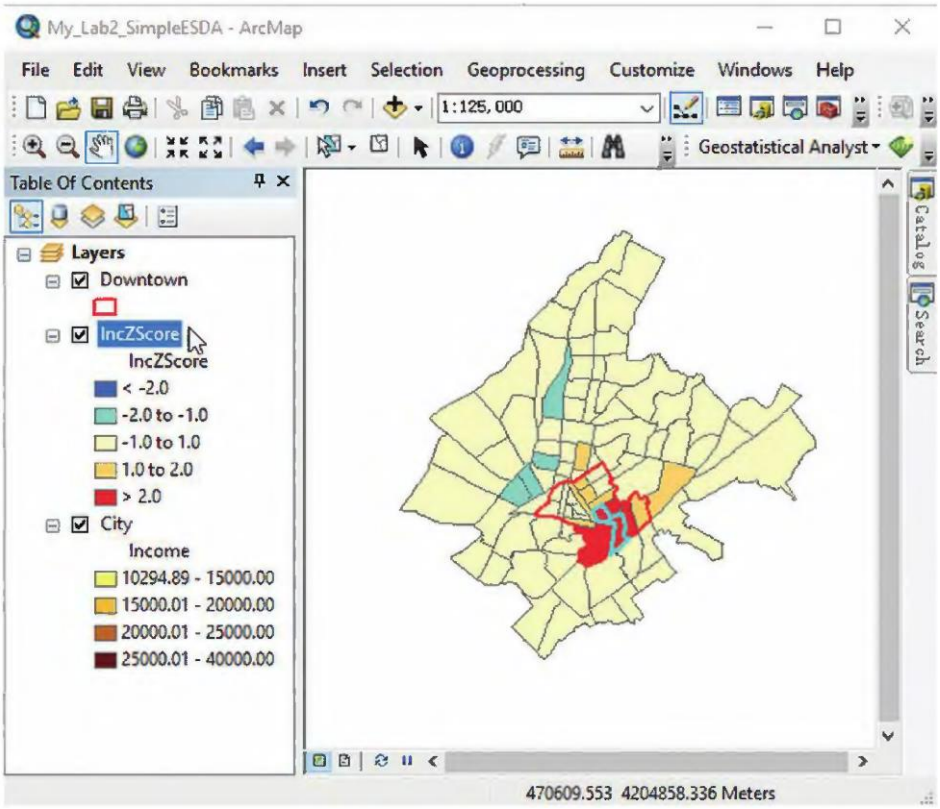

Рис. 5.13. Отображение z-оценок дохода.У двух округов величина z-оценки превысила 2.5

Индивидуальное задание. Используя рассмотренный подход выполнить анализ результатов кадастровой оценки сельскохозяйственных земель сельскохозяйственного предприятия согласно выданным преподавателем данным