Министерство образования и науки Российской Федерации

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ОБРАЗОВАНИЯ
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»

Кафедра Математической экономики

ПРОГРАММНЫЕ СРЕДСТВА ОБРАБОТКИ ГЕОКОДИРОВАННЫХ СТАТИСТИЧЕСКИХ ДАННЫХ

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

Студента 2 курса 247 группы

направления 09.04.03 – «Прикладная информатика»

Механико-математического факультета

Картошкина Андрея Геннадьевича

Научный руководитель
профессор, д. э. н., профессор В.А. Балаш

Зав. кафедрой
д.ф.-м.н., профессор С.И. Дудов

Саратов 2018

Введение. Актуальность исследования. Актуальность исследования обусловлена тем, что в настоящее время рынок ценных бумаг в Российской Федерации набирает обороты, и изучение методов моделирования их поведения становится отнюдь не последним по значимости. Используя методы моделирования поведения котировок любой интересующийся человек сможет принять решение как, в конечном счёте, о совершении манипуляций с ценными бумагами, так и в целом о желании войти на данный рынок. Зная и понимая его состояние и возможности, держатель ценных бумаг имеет возможность планировать свои расходы и доходы, лучше ориентироваться в сфере для прогнозирования своего финансового будущего и принятия важнейших рыночных решений. Моделирование поведения котировок финансовых инструментов способно помочь участникам рынка в определении его перспектив, динамики, а также более прибыльного направления деятельности. Помимо всего прочего, создание модели способно предупредить о ряде возможных кризисных явлений на различных видах рынков. Внедрение моделирования способно дать колоссальный эффект для экономики и финансовую выгоду для держателей ценных бумаг.

Целью исследования является разработка программного обеспечения для обработки геоданных.

В соответствии с данной целью поставлены и решены следующие **задачи**:

1. Рассказать о геоинформационных системах;
2. Охарактеризовать основные методы обработки геоданных;
3. Создать программный продукт, используя данные по макроэкономическим показателям по регионам Российской Федерации.

Объектом исследования является программный продукт, который даст возможность обрабатывать геоданных по макроэкономическим показателям.

Предметом исследования является обработка геоданных и их применение для анализа социально-экономической ситуации в регионах.

В первой части *«Географическая информационная система (ГИС)» описываются история и развитие ГИС, области её применения, а также рассказывается об актуальности развития данной системы.*

Географическая информационная система (ГИС) является системой, предназначенной для сбора, хранения, обработки, анализа, управления и представление пространственных или географических данных . Акроним ГИС иногда используется для обозначения научной дисциплины (GIScience), изучающей географические информационные системы и являющейся частью более широкой академической дисциплины - геоинформатики .

В общем, термин описывает любую информационную систему, которая объединяет, сохраняет, редактирует, анализирует, делится и отображает географическую информацию.

ГИС-приложения - это инструменты, которые позволяют пользователям создавать интерактивные запросы, анализировать пространственную информацию, редактировать данные на картах и представлять результаты всех этих операций. Географическая информатика - это наука, лежащая в основе географических концепций, приложений и систем. Впервые термин «географическая информационная система» был использован Роджером Томлинсоном в 1968 году в его статье «Географическая информационная система для регионального планирования». Томлинсон также признан «отцом ГИС».

В 1854 году Джон Сноу определил источник вспышки холеры в Лондоне путем маркировки точек на карте, отображающих места, где жили жертвы холеры. В результате он получил скопление точек в одном месте, рядом с которым он нашел источник воды. Это было одно из самых ранних успешных применений географической методологии в эпидемиологии. В то время, как основные элементы топографии существовали в картографии и ранее, карта Джона Сноу была уникальной тем, что использовала картографические методы не только для описания, но и для анализа кластеров географически зависимых явлений.

Современные технологии ГИС используют цифровую информацию, для которой используются различные методы создания цифровых данных. Наиболее распространенным методом создания данных является оцифровка , когда карта или план съемки переносится в цифровой носитель посредством использования программы CAD и возможностей геолокации. Благодаря широкой доступности орто-исправленных образов (со спутников, самолетов, воздушных шаров Helikites и беспилотных летательных аппаратов) оцифровка

кадров становится основным средством, с помощью которого извлекаются географические данные.

ГИС использует пространственно-временное местоположение в качестве ключевой индексной переменной для любой другой информации. Подобно тому, как реляционная база данных, содержащая текст или цифры, может связывать множество разных таблиц с использованием общих индексных переменных (ключи), ГИС может относить другую несвязанную информацию, используя местоположение в качестве ключевой индексной переменной. Ключ - это местоположение или протяженность пространства-времени. Любая переменная, которая может быть расположена пространственно и, как правило, временно, может быть указана с использованием ГИС. Местоположения или экстенты в пространстве-времени Земли могут быть записаны как даты и время появления, а координаты x, y и z представляют собой, долготу , широту и высоту соответственно. Эти координаты ГИС могут представлять собой другие квантованные системы пространственно-временной привязки (например, номер кадра фильма, маркер дорожной мили, контрольный ориентир геодезиста, адрес здания, пересечение улиц, входные ворота, зондирование глубины воды, рисунок POS или CAD происхождение). Единицы, применяемые к записанным временным пространственным данным, могут широко варьироваться (даже при использовании точно таких же данных), но все земные пространственно-временные ссылки на местоположение и протяженность должны в идеале относиться друг к другу и, в конечном счете, к «реальному» физическому местоположению или степени в пространстве-времени.

С помощью точной пространственной информации можно анализировать, интерпретировать и представлять невероятное разнообразие реальных и прогнозируемых прошлых или будущих данных. Эта ключевая характеристика ГИС начала открывать новые направления научного исследования поведения моделей реальной информации, которые ранее не были систематически коррелированы.

Во второй части *«Методы обработки геоданных» описываются основные методы обработки географической информации.*

Методы обработки геоданных позволяют оценивать большое количество пространственных закономерностей. С помощью этих методов можно отве-

тить на вопросы "Наблюдается ли в пространственном наборе данных или связанным с ним значением пространственная кластеризация?"и "Усиливается ли кластеризация со временем или нет?".

Охарактеризуем некоторые из доступных методов:

1. Анализ кластеризации точечных объектов. Вычисляет индекс кластеризации точечных объектов на основе среднего расстояния от каждого объекта до ближайшего к нему соседнего объекта.
2. Анализ горячих точек. Определяет статистическую значимость "горячих"точек и "холодных"точек на основе индекса Getis-Ord Gi.
3. Кластеризация с высокими и низкими значениями. Измеряет степень кластеризации высоких или низких значений, используя расчет индекса Getis-Ord Gi.
4. Многовариантный пространственный кластерный анализ (Ripleys K Function). Определяет, проявляют ли пространственные объекты, или ассоциированные с ними значения, статистически значимую кластеризацию или дисперсию по диапазону расстояний.
5. Пространственная автокорреляция. Измеряет пространственную автокорреляцию на основе местоположений пространственных объектов и атрибутивных значений, используя статистику общего индекса Морана.

Анализ кластеризации точечных объектов измеряет расстояние между центром каждого объекта и местоположением центра его ближайшего соседа. Затем он усредняет все эти расстояния до ближайших соседей. Если среднее расстояние меньше среднего для гипотетического случайного распределения, считается, что такое распределение объектов кластеризуется. Если среднее расстояние меньше среднего для гипотетического случайного распределения, считается, что такое распределение объектов дисперсное. Соотношение среднего расстояния между соседними объектами рассчитывается как отношение наблюдаемого среднего расстояния к ожидаемому среднему расстоянию (ожидаемое среднее расстояние рассчитывается для гипотетического случайного распределения с тем же количеством объектов, покрываю-

щих ту же самую общую область). Индекс кластеризации точечных объектов (ANN) рассчитывается по следующей формуле:

$$ANN = (\frac{\overline{D_0}}{\overline{D_E}}) \tag{1}$$

где $\overline{D_0}$ - среднее наблюдаемое расстояние от каждого объекта до его ближайшего соседа:

$$\overline{D_0} = \frac{\sum_{i=1}^{n} d_i}{n} \tag{2}$$

а $\overline{D_E}$ - ожидаемое среднее расстояние от каждого объекта до его ближайшего соседа, которое расчитывается следующим образом:

$$\overline{D_E} = \frac{0.5}{\sqrt{n/A}} \tag{3}$$

В приведенный выражениях d_i равняется расстоянию между объектом i и его ближайшим соседом, n соответствует общему числу объектов, и A является площадью минимально прямоугольника, охватывающего все объекты.

Z-оценка кластеризации точечных объектов рассчитывается следующим образом:

$$Z = \frac{\overline{D_0} - \overline{D_E}}{SE} \tag{4}$$

Где SE (среднеквадратичная ошибка) равна:

$$SE = \frac{0.26136}{\sqrt{n^2/A}} \tag{5}$$

Метод анализа горячих точек рассчитывает индекс Getis-Ord Gi для каждого объекта в наборе данных. Индекс Getis-Ord Gi - индекса пространственной автокорреляции, проверяющий кластеризованность данных в выбранной области. Данный индекс анализирует исследуемую область локально с учетом соседства точек. Итоговые z-оценки и p-значения говорят о том, в какой области пространства кластеризуются объекты с высокими или низкими значениями. Метод работает путем анализа каждого объекта в контексте соседних объектов. Объект с высоким значением интересен, но, возможно, не

является статистически существенной горячей точкой. Чтобы быть статистически существенной горячей точкой, объект должен иметь высокое значение и быть окружен другими объектами с также высокими значениями. Локальная сумма для объекта и его соседей сравнивается пропорционально с суммой всех объектов; когда локальная сумма очень отличается от ожидаемой локальной суммы, и когда это отличие является слишком большим, чтобы быть результатом случайного процесса, получается статистически значимая z-оценка.

Сначала требуется вычисли индекс G_i. Делается это по следующей формуле:

$$G_i = \frac{\sum_{j=1}^{n} w_{i,j}\, x_j - \overline{X} \sum_{j=1}^{n} w_{i,j}}{S\sqrt{\frac{n \sum_{j=1}^{n} w_{i,j}^2 \breve{} (\sum_{j=1}^{n} w_{i,j})^2}{n-1}}} \tag{6}$$

Где x_j - атрибутивное значение для объектов j, $w_{i,j}$ - пространственный вес между объектом i и j, n - общее число объектов.

Индекс G_i, возвращенный для каждого объекта в наборе данных, является z-оценкой. Для статистически значимых положительных z-оценок, чем больше z-оценка, тем более интенсивна кластеризация высоких значений (горячая точка). Для статистически значимых негативных z-оценок, чем меньше z-оценка, тем более интенсивна кластеризация низких значений (холодная точка).

Кластеризация с высокими и низкими значениями (High/Low Clustering) измеряет концентрацию высоких или низких значений в изучаемой области, используя расчет глобального индекса Getis-Ord Gi.

Общий индекс G, определяющий степень кластеризации рассчитывается по формуле:

$$G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j}\, x_i x_j}{\sum_{i=1}^{n} \sum_{j=1}^{n} x_i x_j}, \forall j \neq i \tag{7}$$

где x_i и x_j - атрибутивные значения для объектов i и j, а $w_{i,j}$ - пространственный вес для пары объектов i и j. n соответсвует общему числу объектов в наборе и $\forall j \neq i$ показывает, что объекты i и j не могут быть одним и тем же объектом. Z-значение вычисляется следующим образом:

$$ZG = \frac{G - E[G]}{\sqrt{V[G]}} \tag{8}$$

где:

$$\begin{aligned} E[G] &= \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j}}{n(n-1)}, \forall j \neq i \\ V[G] &= E[G^2] - E[G]^2 \end{aligned} \tag{9}$$

Пространственный кластерный анализ на основе множественных расстояний, основанный на K-функции Рипли – это еще один метод обработки геоданных в случайных точечных данных. Отличительной чертой этого метода от остальных является то, что он суммирует пространственную зависимость (кластеризация или дисперсия объектов) по всему диапазону расстояний. Во многих исследованиях по изучению пространственных закономерностей необходим выбор подходящего масштаба анализа. Например, часто необходимо определить диапазон расстояний или пороговое расстояние (Distance Band or Threshold Distance).

При исследовании пространственных закономерностей на множественных расстояниях и пространственных масштабах, работают изменения закономерностей, часто отражающие превалирование определенных пространственных процессов. K-функция Рипли отражает, как центры пространственных кластеров или дисперсий изменяются при изменении размера соседей.

Существует несколько вариантов K-функции Рипли. Здесь представлена обычная трансформация K-функции, часто обозначаемая как $L(d)$:

$$L(d) = \sqrt{\frac{A\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} k_{i,j}}{\pi n(n-1)}} \tag{10}$$

где d - расстояние, n - общее количество объектов, A - значение площади для экстента всех объектов и $k_{i,j}$ - значение веса. Если не используется по-

правка по границам, вес равен 1, если расстояние между i и j меньше d, и равно 0 во всех остальных случаях.

Пространственная автокорреляция (Spatial Autocorrelation (Global Moran's I)) измеряет пространственную автокорреляцию, основанную одновременно на расположении объектов и их значениях. Исходя из предложенного набора объектов и связанных с ними атрибутов, метод пространственной автокорреляции оценивает, имеется ли кластеризация объектов или они распределены разбросанно, или случайно. Метод рассчитывает Морана, а также z-оценку и p-значение, чтобы оценить значимость индекса.

Индекс Морана для пространственной автокорреляции задается по следующей формуле:

$$I = \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} z_i z_j}{S_0 \sum_{i=1}^{n} z_i^2} \tag{11}$$

где z_i - отклонение атрибута для объекта i от его среднего значения $(x_i - \overline{X})$, $w_{i,j}$ - пространственный вес между объектами i и j, n - общее число объектов и S_0 - совокупность всех пространственных весов и вычисляется по формуле:

$$S_0 = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} \tag{12}$$

Z-оценка рассчитывается следующим образом:

$$Z_I = \frac{I - E[I]}{\sqrt{V[I]}} \tag{13}$$

где:

$$E[I] = -1/(n-1) V[I] = e[I^2] - E[I]^2 \tag{14}$$

В третьей части "Обработка геоданных с помощью языка R" *реализуется решение поставленной задачи с помощью языка R.*

Известно, что выявление, например, положительного пространственного эффекта для роста дохода на душу населения в регионах соответствует тому, что экономически растущий регион «тянет за собой» другие регионы. В то же время отрицательный пространственный эффект для роста дохода на душу населения соответствует тому, что растущий регион «забирает на себя» ресурсы и не дает расти другим регионам. Цель работы состоит в том, чтобы подтвердить теорию о связи макроэкономических показателях в соседних регионах с помощью обработки геокодированных данных.

Для выявления возможных пространственных эффектов были выделены три макроэкономических показателя:

1. Уровень дохода на душу населения в каждом регионе
2. Средняя продолжительность жизни по каждому региону
3. Количество врачей для каждого региона

Для выявления пространственных эффектов требуется сформированы диаграммы Морана и диаграммы Гири, которые смогут покаывают взаимосвязь между макроэкономическими показателями.

Для того, чтобы построить требуемые диагаммы необходимо понять, какие регионы граничат между собой. Для этого с помощью языка R графически показываются связи между субъектами РФ.

Зная соседей по каждому региону можно понять, влияет ли соседтсво более сильных регионов на общую ситуацию в округе. Для этого формируется диагруммы Морана и Гири по среденей продолжительности жизни в зависимости от дохода, а также количество врачей в расчете на душу населения.

По результатам исследования видно, что есть группы регионов, у которых самые высокие доходы, самая высокая продолжительность жизни, а также больше всего врачей на душу населения. Также есть группы, у которых обратная ситуация и все макроэкономические показатели находятся на более низком уровне.

Заключение. Проведённое исследование показало актуальность темы обработки геоданных. В особенности это важно для нашей страны, так как несмотря на большое расстояние между регионами присутвует связь между макроэкономическими показателями. Таким образом, даже улучшая уровень

жизни одного региона, а соответственно и его соседей, ситуация по округу в целом может не измениться.

Также, можно заключить, что не смотря на развитие центральных регионов, ситуция по стране в целом не меняется, а лишь усугубляется. И если картина не поменяется. то Россия рискует превратится в страну состоящей из одной Москвы и прилигающих к ней территорий.