|  |  |
| --- | --- |
|  | **Министерство науки и высшего образования Российской Федерации**  **Федеральное государственное бюджетное образовательное учреждение**  **высшего образования**  **«Московский государственный технический университет**  **имени Н.Э. Баумана**  **(национальный исследовательский университет)»**  **(МГТУ им. Н.Э. Баумана)** |

ФАКУЛЬТЕТ \_\_\_\_\_ИНФОРМАТИКА И СИСТЕМЫ УПРАВЛЕНИЯ\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

КАФЕДРА \_\_\_\_\_СИСТЕМЫ ОБРАБОТКИ ИНФОРМАЦИИ И УПРАВЛЕНИЯ\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

***К КУРСОВОЙ РАБОТЕ***

***НА ТЕМУ:***

*Разведочный анализ данных для выбранного датасета*

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

Студент \_\_ИУ5-14М\_\_\_\_\_\_\_ **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_**Н.В. Журавлев**\_\_\_\_\_\_**

(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель курсовой работы **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_**В.Ю. Строганов**\_\_\_\_**

(Подпись, дата) (И.О.Фамилия)

*2023 г.*

**Министерство науки и высшего образования Российской Федерации**

**Федеральное государственное бюджетное образовательное учреждение**

**высшего образования**

**«Московский государственный технический университет имени Н.Э. Баумана**

**(национальный исследовательский университет)»**

**(МГТУ им. Н.Э. Баумана)**

УТВЕРЖДАЮ

Заведующий кафедрой \_\_\_ИУ5\_\_\_\_

(Индекс)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_В.И. Терехов\_\_

(И.О.Фамилия)

«\_04\_» \_\_\_\_сентября\_\_\_\_\_\_\_ 2023 г.

**ЗАДАНИЕ**

**на выполнение курсовой работы**

по дисциплине \_\_\_\_Многомерный анализ данных в системах искусственного интеллекта \_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Студент группы \_ИУ5-14М\_\_

Журавлев Николай Вадимович\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Фамилия, имя, отчество)

Тема курсовой работы Разведочный анализ данных для выбранного датасета\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Направленность КР (учебная, исследовательская, практическая, производственная, др.)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_УЧЕБНАЯ\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Источник тематики (кафедра, предприятие, НИР) \_\_\_\_\_КАФЕДРА\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

График выполнения работы: 25% к \_\_\_ нед., 50% к \_\_\_ нед., 75% к \_\_ нед., 100% к \_\_\_ нед.

***Задание*** \_Произвести разведочный анализ данных выбранного датасета\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

***Оформление курсовой работы:***

Расчетно-пояснительная записка на \_12\_\_ листах формата А4.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Дата выдачи задания «\_04\_» \_\_\_\_сентября\_\_\_\_\_\_\_ 2023 г.

**Руководитель курсовой работы**  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_В.Ю. Строганов\_\_\_\_\_

(Подпись, дата) (И.О.Фамилия)

**Студент \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_**Н.В.Журавлев\_\_**\_\_\_\_\_**

(Подпись, дата) (И.О.Фамилия)

Примечание: Задание оформляется в двух экземплярах: один выдается студенту, второй хранится на кафедре.

## Оглавление

[Оглавление 2](#_Toc156029133)

[Цель работы 3](#_Toc156029134)

[Описание анализируемого датасета 3](#_Toc156029135)

[Корреляционный анализ 4](#_Toc156029136)

[Регрессионного анализа — линейная регрессия 5](#_Toc156029137)

[Факторный анализ 7](#_Toc156029138)

[Дисперсионный анализ 8](#_Toc156029139)

[Кластерный анализ 9](#_Toc156029140)

[Дискриминантный анализ 9](#_Toc156029141)

[Заключение 11](#_Toc156029142)

[Список литературы 12](#_Toc156029143)

## Цель работы

Целью данной курсовой работы является апробация методов разведочного анализа данных и прогнозирования:

* корреляционный анализ
* дисперсионный анализ
* регрессионный анализ
* факторный анализ
* кластерный анализ
* дискриминантный анализ

## Описание анализируемого датасета

В качестве исследуемого набора выбран Estate\_valuation.csv – набор рыночных данных по оценке недвижимости, собранный в Синдианском округе, город Нью-Тайбэй, Тайвань. Набор содержит:

1. X1 transaction date – дата заключения сделки; значения - дата в формате год, прошедшая часть (например, 2013,250 = март 2013 г., 2013,500 = июнь 2013)
2. X2 house age – возраст дома; значения – любые численные
3. X3 distance to the nearest MRT – метров до ближайшей станции метро; значения - любые численные
4. X4 number of convenience stores - количество магазинов повседневного спроса в шаговой доступности; значения - любые численные
5. X5 latitude – широта, на которой расположен дом; значения - любые численные
6. X6 longitude – долгота, на которой расположен дом; значения - любые численные
7. Y house price of unit area - стоимость дома за единицу площади; значения - любые численные

## Корреляционный анализ

Корреляционный анализ - это статистический инструмент, позволяющий установить связь между двумя или более различными переменными, а также оценить, насколько сильна взаимосвязь между этими переменными. Суть корреляционного анализа заключается в вычислении коэффициента корреляции, который показывает степень линейной зависимости между двумя переменными.

Если результат корреляционного анализа положительный, то взаимосвязь двух переменных прямо пропорциональная. Это означает, что при увеличении одной переменной, вторая будет также увеличиваться. Как правило, такой результат принято называть “позитивной корреляцией”. Если результат корреляционного анализа отрицательный, то взаимосвязь двух переменных обратно пропорциональная. Это означает, что при увеличении одной переменной, вторая будет уменьшаться. Такой эффект называется “отрицательной корреляцией”. Если результат анализа стремится к нулю, то взаимосвязь между двумя переменными отсутствует.

В данной курсовой работе использовался корреляционный анализ Спирмена.

Листинг 1

|  |
| --- |
| df = pd.read\_csv('csvwn.csv', encoding='latin-1', sep=';')  numeric = ['X1', 'X2', 'X3', 'X4', 'X5', 'X6', 'Y']  sns.pairplot(df[numeric])  print(df[numeric].corr(method='spearman', numeric\_only=False)) |

В результате получаем значения показанные на рис.1.

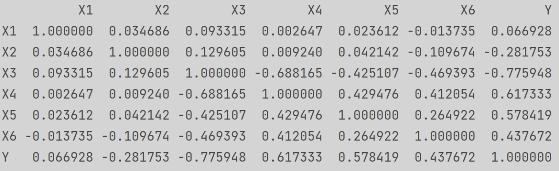


Рисунок 1

Интерпретация значений коэффициента корреляции:

1. до 0,5 – слабая корреляции
2. от 0,5 до 0,7 – средняя корреляции
3. от 0,7 до 0,9 – высокая корреляции
4. от 0,9 - очень высокая корреляция

Исходя из полученных результатов можно сделать следующие выводы:

1. Стоимость дома за единицу площади в значительной степени зависит от количества магазинов повседневного спроса в шаговой доступности (X4 number of convenience stores)
2. Стоимость дома за единицу площади так же сильно зависит от метров до ближайшей станции метро (X3 distance to the nearest MRT)
3. В меньшей, но значительной степени стоимость дома за единицу площади зависит от широты и долготы, на которых расположен дом (X5 latitude, X6 longitude)

Остальные зависимости не рассматриваются из-за за того, что рассматривается столбец Y house price of unit area.

## Регрессионного анализа — линейная регрессия

Для регрессионного анализа выдвигаем гипотезу о том, какие факторы влияют на стоимость дома за единицу площади. Гипотеза для регрессионного анализа сформулирована следующим образом: "Стоимость дома за единицу площади зависит от возраста дома, метров до ближайшей станции метро, количество магазинов повседневного спроса в шаговой доступности, широты, на которой расположен дом, долготы, на которой расположен дом".

Листинг 2

|  |
| --- |
| X = df[['X1', 'X2', 'X3', 'X4', 'X5', 'X6']]  Y = df['Y']  X = sm.add\_constant(X, prepend=False)  model = OLS(Y, X)  res = model.fit()  print(res.summary()) |

После выполнения получается результат представленный на рис.2.

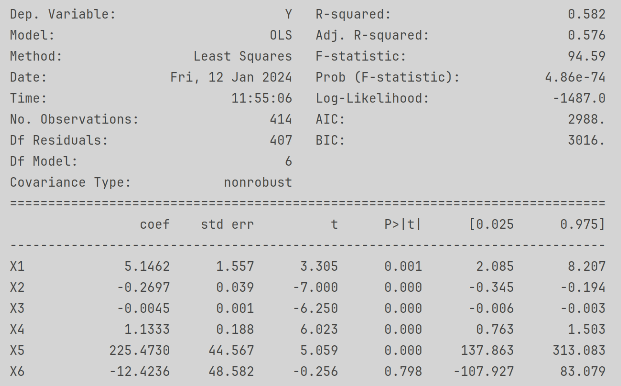


Рисунок 2

Значение R-squared всегда находится в диапазоне от 0 до 1, где 0 означает отсутствие связи между переменными, а 1 означает полную линейную связь. В нашем случае R-squared: 0.582, что означает среднюю линейную связь.

Столбец X6 может отбросить, так как коэффициент значимости p\_value выше 0.05.

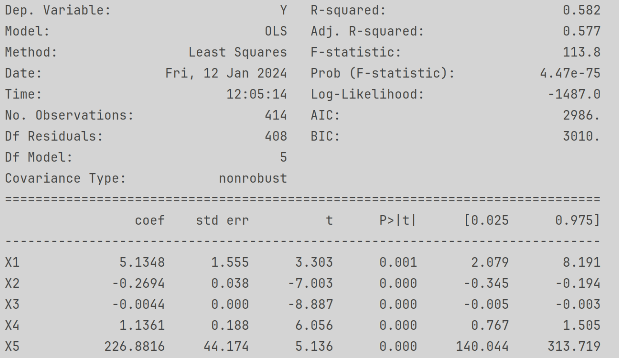


Рисунок 3

После удаления X6 коэффициент детерминации не изменился. Это означает, что X6 не влияет на изменение зависимой переменной.

## Факторный анализ

Факторный анализ - многомерный метод, применяемый для изучения взаимосвязей между значениями переменных. Предполагается, что известные переменные зависят от меньшего количества неизвестных переменных и случайной ошибки. Факторный анализ позволяет решить две важные проблемы исследователя: описать объект измерения всесторонне и в то же время компактно. С помощью факторного анализа возможно выявление скрытых переменных факторов, отвечающих за наличие линейных статистических корреляций между наблюдаемыми переменными.

Листинг 3

|  |
| --- |
| for col in df.columns:  df[col] = (df[col] - df[col].mean()) / df[col].std()  fa = Factor(df, n\_factor=3, method='pa')  res = fa.fit() |

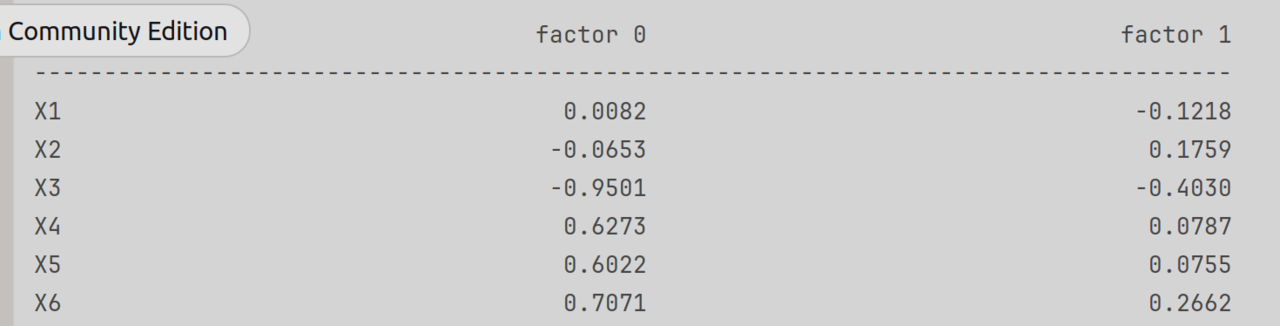


Рисунок 4

В результате по таблице факторных нагрузок можно увидеть, что переменные Х3, X4, X5, X6 коррелируют с фактором 1, и X3, с фактором 2.

## Дисперсионный анализ

В проведенном дисперсионном анализе используется гипотеза о равенстве средних значений в нескольких группах. В данном случае мы предполагаем, что средние значения измеряемых переменных одинаковы для всех групп или обработок. Формально, нулевая гипотеза (H0) в дисперсионном анализе звучит как "Средние значения всех групп равны".

Листинг 4

|  |
| --- |
| data = np.array(df[numeric])  f\_value, p\_value = stats.f\_oneway(\*data)  print("F-значение:", f\_value)  print("p-значение:", p\_value) |

После проведения дисперсионного анализа, и получения значений F-значение и p-значения, мы можем принять решение об отклонении или не отклонении нулевой гипотезы в зависимости от уровня значимости (обычно 0.05).

Для проведения дисперсионного анализа используется функцию f\_oneway из библиотеки scipy.stats, которая возвращает значения F-значения и p-значения.

F-значение: 0.26574880853703486

p-значение: 0.9999999999999999

Так как p-значение больше 0.05, то нулевую гипотезу не отвергаем.

## Кластерный анализ

Главное назначение кластерного анализа – разбиение множества исследуемых объектов и признаков на однородные в соответствующем понимании группы или кластеры. Это означает, что решается задача классификации данных и выявления соответствующей структуры в ней. Для проведения кластерного анализа, используем метод k-средних из библиотеки scikit-learn для Python.

Листинг 5

|  |
| --- |
| df1 = pd.DataFrame(df[numeric])  scaler = StandardScaler()  scaled\_data = scaler.fit\_transform(df1)  kmeans = KMeans(n\_clusters=6)  kmeans.fit(scaled\_data)  df1['Cluster'] = kmeans.labels\_  print(df1) |

В результате получаем распределение по кластерам:

X1 X2 X3 X4 X5 X6 Y Cluster

0 2012.917 32.0 84.87882 10 24.98298 121.54024 37.9 4

1 2012.917 19.5 306.59470 9 24.98034 121.53951 42.2 0

2 2013.583 13.3 561.98450 5 24.98746 121.54391 47.3 2

3 2013.500 13.3 561.98450 5 24.98746 121.54391 54.8 2

4 2012.833 5.0 390.56840 5 24.97937 121.54245 43.1 0

.. ... ... ... .. ... ... ... ...

409 2013.000 13.7 4082.01500 0 24.94155 121.50381 15.4 3

410 2012.667 5.6 90.45606 9 24.97433 121.54310 50.0 0

411 2013.250 18.8 390.96960 7 24.97923 121.53986 40.6 2

412 2013.000 8.1 104.81010 5 24.96674 121.54067 52.5 0

413 2013.500 6.5 90.45606 9 24.97433 121.54310 63.9 2

## Дискриминантный анализ

Дискриминантный анализ (Discriminant Analysis) - это статистический метод, для решения задач распознавания образов, который используется для принятия решения о том, какие переменные разделяют возникающие наборы данных (так называемые «группы»). В отличие от кластерного анализа в дискриминантном анализе группы известны заранее. Основная цель дискриминантного анализа состоит в том, чтобы определить, какие переменные наилучшим образом разделяют эти группы.

Суть дискриминантного анализа заключается в поиске линейной комбинации признаков, которая наилучшим образом разделяет группы объектов, которая называется дискриминантной функцией. После того как она найдена, её используют для предсказания принадлежности новых объектов к определенным группам.

Листинг 6

|  |
| --- |
| data = {  'transaction\_date': df['X1'],  'house\_price\_of\_unit\_area': df['Y'],  'house\_age': df['X2'],  'distance\_to\_the\_nearest\_MRT': df['X3'],  'number\_of\_convenience\_stores': df['X4'],  'latitude': df['X5'],  'longitude': df['X6']  }  df2 = pd.DataFrame(data)  X = df2[['house\_age', 'distance\_to\_the\_nearest\_MRT', 'number\_of\_convenience\_stores']]  y = df2[['house\_price\_of\_unit\_area']]  y = y.astype('int')  model = LinearDiscriminantAnalysis()  model.fit(X, y)  new\_data = [[19.5, 306.5947, 9]]  predicted\_class = model.predict(new\_data)  print(f'Предсказание: {predicted\_class[0]}') |

Для проверки предсказания нового значения использовалось вторая строка из датасета. Из дискриминантного анализа, мы получаем предсказание для цены равным 44. При сравнении с изначальным датасетом, где результатом является 42.2, можно увидеть, что это является близко к полученному результату.

## Заключение

В ходе выполнения данной курсовой работы были проанализированы данные датасета рыночных данных по оценке недвижимости. Из результатов корреляционного анализа можно сделать следующие выводы:

1. Стоимость дома за единицу площади в значительной степени зависит от количества магазинов повседневного спроса в шаговой доступности (X4 number of convenience stores)
2. Стоимость дома за единицу площади так же сильно зависит от метров до ближайшей станции метро (X3 distance to the nearest MRT)
3. В меньшей, но значительной степени стоимость дома за единицу площади зависит от широты и долготы, на которых расположен дом (X5 latitude, X6 longitude)

Результат регрессионного анализа – нахождения значения R-squared равным 0.582, что означает среднюю линейную связь.

В результате факторного анализа по таблице факторных нагрузок можно увидеть, что переменные Х3, X4, X5, X6 коррелируют с фактором 1, и X3, с фактором 2.

Результаты кластерного анализа показали разделение данных на 6 кластеров.

Результаты дисперсионного анализа показали, что мы не отвергаем нулевую гипотезу: нет оснований считать средние значения различными.

Из дискриминантного анализа, мы получаем предсказание цены для второй строки датасета равным 44. При сравнении с изначальным датасетом, где результатом является 42.2, можно увидеть, что это является близко к полученному результату.

## Список литературы

1. Афонин П.Н., Афонин Д.Н. Статистический анализ с применением современных программных средств : учебное пособие / Афонин П.Н., Афонин Д.Н; – СПб.: ИЦ «Интермедия», 2017. – 100 c.
2. Боровиков В.П. Statistica: искусство анализа данных на компьютере. Для профессионалов. СПб.: Питер, 2003. – 688 с.
3. Боровиков В.П., Ивченко Г.И. Прогнозирование в системе Statistica в среде Windows. Основы теории и интенсивная практика на компьютере. М., 2006. – 368 с
4. Вуколов Э.А. Основы статистического анализа. Практикум по статистическим методам и исследованию операций с использованием пакетов Statistica и Excel. М.: Форум, 2010. – 464 с.
5. Глинский В.В., Ионин В.Г. Статистический анализ. Учебное пособие. Новосибирск: Инфра-М, 2002. – 240 с
6. Симчера В.М. Методы многомерного анализа статистических данных. М.: Финансы и статистика. – 2008. – 400 с.
7. Тихомиров Н.П., Тихомирова Т.М., Ушмаев О.С. Методы эконометрики и многомерного статистического анализа. М.: Экономика, 2011. – 640 с.
8. Халафян А.А. Statistica 6. Статистический анализ данных. СПб.: Бином-Пресс, 2010. – 528 с.