

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ**

**ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ**

**«ДОНСКОЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»**

**(ДГТУ)**

Кафедра «Организация перевозок и дорожного движения»

**МЕТОДИЧЕСКИЕ УКАЗАНИЯ**

**ДЛЯ ПРАКТИЧЕСКИХ ЗАНЯТИЙ ПО ДИСЦИПЛИНЕ**

**«ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ НА ТРАНСПОРТЕ»**

Направление подготовки 23.04.01 «Технология транспортных процессов»

Программа: «Транспортная логистика», «Организация перевозок на автомобильном транспорте», «Интеллектуальные транспортные системы»

Ростов-на-Дону

ДГТУ

2023

УДК 656

Составитель:

Методические указания для практических занятий по дисциплине «Интеллектуальный анализ данных на транспорте» –

Ростов-на-Дону: Донской гос. техн. ун-т, 2023. – 25 с.

УДК 656

Методические указания предназначены для магистрантов очной и заочной формы обучения по направлению подготовки 23.04.01 «Технология транспортных процессов», программы «Транспортная логистика», «Организация перевозок на автомобильном транспорте», «Интеллектуальные транспортные системы». Приведены примеры и задачи, связанные с подготовкой и анализом транспортных данных.

Печатается по решению редакционно-издательского совета Донского государственного технического университета

Ответственный за выпуск зав. кафедрой «Организации перевозок и дорожного движения» д-р техн. наук, профессор В.В. Зырянов

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

В печать \_\_\_.\_\_\_.2023 г.

Формат 60×84/16. Объем \_\_\_ усл. п. л.

Тираж \_\_\_ экз. Заказ № \_\_\_

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

Издательский центр ДГТУ

Адрес университета и полиграфического предприятия:

344000, г. Ростов-на-Дону, пл. Гагарина, 1

© Донской государственный

технический университет, 2023

**Введение**

Интеллектуальный анализ данных как одно из новых междисциплинарных направлений включает в себя различные способы работы с данными и в зависимости от предметной области имеет некоторую специфику применения этих способов.

Ставшее уже классическим data mining применимо не ко всем задачам технологии транспортных процессов. В частности, исследование транспортных потоков требует специфических подходов к очистке, агрегированию, преобразованию и последующему анализу данных о потоке.

Знания, полученные ранее при изучении дисциплины «Математические методы в транспортных исследованиях», в полной мере могут быть применены при интеллектуальном анализе данных о транспортных потоках.

Последовательное выполнение практических заданий по дисциплине «Интеллектуальный анализ данных на транспорте» позволит получить необходимые магистранту навыки научных исследований и компетенции.

Для выполнения всех заданий следует получать исходные данные у преподавателя , по ссылке или qr-коду из приложения 1. Результаты следует сохранять в файлах с именем, состоящим из номера группы, фамилии студента и названия, указанного в условии задачи, например «АМИТС11\_Иванов\_task1» или «AMITS\_Li\_task1». Все примеры следует выполнить и сохранить в одном файле.

**Предварительная подготовка данных**

**Пример 1**. В предложенном для анализа файле провести очистку данных. Конкретная цель очистки не поставлена и данные из предложенного источника исследуются впервые, поэтому в результате требуется получить три файла: исходный (с именем «1»), с очищенными данными (с именем «2») и с удаленными из «1» данными (с именем «3»). Все операции по очистке занести в ведомость (таблица 1).

Таблица 1. Ведомость очистки

|  |  |  |  |
| --- | --- | --- | --- |
| № | Обнаруженное несоответствие | Строка, столбец | Действие |
| 1 | … | … | … |

Определяем количество данных, количество строк, столбцов, количество данных в каждой строке и столбце. Таким образом определяем наличие пустот. Используем фильтры, находим конкретные места расположения пустот.

Находим минимальные и максимальные значения по столбцам, исходя из физического смысла, определяем наличие ошибок и аномалий. Принимаем решение о заполнении пустот нулями или средними значениями, об удалении строк с аномальными и ошибочными значениями или замене таких значений на подходящие.

Преобразуем данные о времени: создаем несколько столбцов, в каждом из которых данные о дате, месяце, годе, часе, минутах и секундах размещены в числовом формате.

Находим одинаковые строки и столбцы. Принимаем решение, являются ли они дубликатами. Дубликаты удаляем.

Проверяем наличие нерелевантных данных. Видим, что во всех записях одинаковые секунды, месяц и год, следовательно на дальнейший анализ они не повлияют. Принимаем решение об удалении этих столбцов. В одном из столбцов все данные равны 1 также удаляем этот столбец.

В результате очистки и преобразований файл «2» содержит: 9502 записи в 11 столбцах и 864 строках. Файл «3» содержит 6 строк. Ведомость очистки представлена в таблице 2.

Таблица 2. Ведомость очистки файла «Пример 1»

|  |  |  |  |
| --- | --- | --- | --- |
| № | Обнаруженное несоответствие, аномалия, пустоты, ошибки | Строка, столбец | Действие (удален, преобразован, заменен копией) |
| 1 | Пустоты | C36, D548 | копирование строки в файл 3, замена на 0, |
| 2 | Ошибка | H30, D381 | копирование строки в файл 3, замена буквы о на 0 в файле 2 |
| 3 | Аномальное значение | D189 | копирование строки в файл 3, удаление строки из файла 2, |
| 4 | Преобразование даты и времени в числовой формат | А:А | Добавляем 6 столбцов с числовыми данными в файле 2 |
| 5 | Дубликаты | 59 и 259 | копирование строки 259 в файл 3, удаление строки 259 из файла 2 |
| 6 | Нерелевантные данные | I:I | Удаление столбца I и столбцы с информацией о месяце, годе и секундах из файла 2 |

**Задача 1**. Провести очистку данных из файла по qr-коду из приложения. Задание выполнить и оформить в соответствии с примером 1. Результат сохранить в файле с именем, содержащим «task1».

**Пример 2**. Найти противоречия в исходных данных и провести необходимые преобразования. Исходные данные содержат часть информации, полученной детекторами дорожного движения. В таблице 3 приведены обозначения и наименования параметров.

Таблица 3. Условные обозначения для примера 2

|  |  |
| --- | --- |
| Обозначение | Наименование параметра |
| time1 | час |
| time2 | минута |
| counter | счетчик ТС – количество ТС |
| speed | средняя скорость |
| k | плотность |
| occup | занятость дороги |
| Cnt0 | счетчик коротких ТС до 3 м |
| Cnt1 | счетчик легковых ТС от 3 до 5,5 м |
| Cnt2 | счетчик ТС от 5,5 до 7 м |
| Cnt3 | счетчик ТС от 7 до 10 м |
| Cnt4 | счетчик ТС от 10 до 15 м |
| Cnt5 | счетчик ТС свыше 15 м |

Предварительно проводим очистку данных в соответствии с примером 1. Параметр «счетчик ТС» показывает, какое количество транспортных средств идентифицировано на данном участке за последний интервал времени. Параметры , обозначенные Cnt0, Cnt1, Cnt2, Cnt3, Cnt4, Cnt5 показывают количество ТС по разным категориям. При идеальной работе системы counter=Cnt0+Cnt1+Cnt2+Ct3+Cnt4+Cnt5. Если равенство не выполняется, то можно сделать вывод, что часть ТС не отнесена ни к одной из категорий и такая информация может подвергаться специальному анализу.

Создадим столбец CntNC:

CntNC =counter-(Cnt0+Cnt1+Cnt2+Ct3+Cnt4+Cnt5)

С помощью фильтра определяем количество нулевых значений CntNC : 16 строк, что составляет 5,5% от общего количество строк в наборе данных.

**Задача 2**. Найти противоречия в исходных данных и провести преобразования по индивидуальным наборам данных. Задание выполнить и оформить в соответствии с примером 2. Результат сохранить в файле с именем, содержащим «task2».

**Задача 3**. Установить , есть ли противоречия между значениями параметров occup, k и counter в исходных данных к задаче 2 и провести преобразования. Результат сохранить в файле с именем, содержащим «task3».

**Задача 4**. Установить , есть ли противоречия между значениями параметров occup, k и counter в исходных данных к задаче 1 и провести преобразования. Результат сохранить в файле с именем, содержащим «task4».

**Пример 3**. Провести очистку без удаления в файле с текстовыми данными аналогично примеру 1.

Определяем количество данных, количество строк, столбцов, количество данных в каждой строке и столбце. Таким образом определяем наличие пустот. Используем фильтры, находим конкретные места расположения пустот.

Обнаружены 26 пустых ячейки в категории «отчество», заменяем на текст «null». Аналогично поступаем во всех других категориях.

Определяем наличие ошибок и аномалий. Проводим нормализацию: названия стран , написанные верхним регистром, заменяем на написанные с заглавной буквы; сокращенное наименование пола заменяем на полное.

Дубликатов нет.

Проверяем наличие нерелевантных данных. Видим, что во всех записях одинаковое значение пола, принимаем решение об удалении столбца G.

В результате очистки и преобразований файл содержит: 2152 записи в 8 столбцах и 272 строках. Файл «3» не создавался. Ведомость очистки представлена в таблице 4.

Таблица 4. Ведомость очистки файла «Пример 3»

|  |  |  |  |
| --- | --- | --- | --- |
| № | Обнаруженное несоответствие, аномалия, пустоты, ошибки | Строка, столбец | Действие (удален, преобразован, заменен копией) |
| 1 | Пустоты | 26 ячеек в столбе С, Е177, | замена на null, |
| 2 | Нормализация | 239 ячеек в столбце Е, 18 ячейка в столбце G | замена РОССИЯ на Россия, ИЗРАИЛЬ на Израиль, ГРУЗИЯ на Грузия, АРМЕНИЯ на Армения, Муж на Мужской |
| 3 | Нерелевантные данные | G:G | Удаление столбца G |

**Задача 5**. Провести очистку данных из файла по qr-коду из приложения. Задание выполнить и оформить в соответствии с примером 3. Результат сохранить в файле с именем, содержащим «task5».

**Описательная статистика**

**Пример 4.** Получить характеристики описательной статистики для параметра counter из примера 2. Показать влияние очистки данных на статистические характеристики.

В результате очистки в примере 2 из столбца counter была удалена 271-я строка с аномально большим значением. В таблице 5 приведены результаты работы надстройки «Описательная статистика» для counter до и после очистки.

Таблица 5. Описательная статистика counter

|  |  |  |  |
| --- | --- | --- | --- |
| counter до очистки | | counter после очистки | |
|  |  |  |  |
| Среднее | 454,9 | Среднее | 26,4 |
| Стандартная ошибка | 428,6 | Стандартная ошибка | 0,8 |
| Медиана | 29,0 | Медиана | 29,0 |
| Мода | 37,0 | Мода | 37,0 |
| Стандартное отклонение | 7273,2 | Стандартное отклонение | 13,7 |
| Дисперсия выборки | 52899052,0 | Дисперсия выборки | 188,6 |
| Эксцесс | 288,0 | Эксцесс | -1,0 |
| Асимметричность | 17,0 | Асимметричность | -0,1 |
| Интервал | 123454,0 | Интервал | 58,0 |
| Минимум | 2,0 | Минимум | 2,0 |
| Максимум | 123456,0 | Максимум | 60,0 |
| Счет | 288,0 | Счет | 287,0 |

Из приведенных в таблице 5 статистических характеристик видно, что наличие аномально большого значения повлияло на среднее значение выборки и как следствие еще на несколько характеристик, зависящих от среднего, таких как дисперсия, стандартное отклонение, стандартная ошибка, эксцесс, асимметричность.

**Задача 6.** Получить характеристики описательной статистики для параметра counter из примера 2 для двух вариантов подготовки данных: удаление строки с аномальным значением и замена аномального значения на среднее. Сделать вывод. Результат сохранить в файле с именем, содержащим «task6».

**Задача 7.** Получить характеристики описательной статистики для всех параметров из примера 2 после очистки и преобразований данных. Результат сохранить в файле с именем, содержащим «task7».

**Пример 5**. Провести описательные анализ параметра counter из примера 2.

Описательный анализ проводим после подготовки данных. Аномальное значение 12345 заменено на среднее 26,37. Для описательного анализа используем несколько характеристик:

|  |  |
| --- | --- |
| Среднее | 26,37 |
| Медиана | 29 |
| Мода | 37 |
| Стандартное отклонение | 13,7 |
| Дисперсия выборки | 187,93 |
| Эксцесс | -1,01 |
| Асимметричность | -0,13 |
| Минимум | 2 |
| Максимум | 60 |

Из полученных статистических характеристик видно, что распределение случайной величины counter не является симметричным, так как среднее не равно медиане и не равно моде. Интервал между минимальным и максимальным 58 единиц, эксцесс отрицательный, следовательно пики не сильно выражены, т.е. данные «размазаны» по всему интервалу, что и дает большую дисперсию. Коэффициент асимметрии отрицательный, следовательно форма плотности распределения «скошена» вправо, что дополнительно подтверждается тем, что мода (число с максимальной частотой) больше медианы (числа, находящегося посередине выборки).

**Задача 8.** Провести описательные анализ всех параметров из примера 2 после очистки и преобразований данных аналогично примеру 5. Результат сохранить в файле с именем, содержащим «task8».

**Задача 8.** Провести описательные анализ параметра counter из файла «task4» (очищенные и преобразованные данные детекторов за трое суток) за каждые сутки отдельно. Сделать общий вывод. Результат сохранить в файле с именем, содержащим «task8».

**Статистическое оценивание**

**Пример 6.** Провести статистическое оценивание времени восстановления автомобилеразгрузчика на зерновом терминале.

Случайной величиной является время восстановления оборудования *tв*, ч. По исходным данным можно получить выборку *tвi* (*i*=1, 2,…, *n*) объема *n*=49. Описательная статистика для полученной выборки представлена в таблице 6

Таблица 6. Описательная статистика

|  |  |
| --- | --- |
| *tв, ч* | |
| Среднее | 8,224 |
| Стандартная ошибка | 0,780 |
| Медиана | 8 |
| Мода | 8 |
| Стандартное отклонение | 5,463 |
| Дисперсия выборки | 29,844 |
| Эксцесс | -0,032 |
| Асимметричность | 0,459 |
| Минимум | 0 |
| Максимум | 24 |

Для построения гистограммы требуется задать значения границ интервалов, которые в электронных таблицах называются «карманом». Количество интервалов можно определить по формуле Стерджеса:

*k=*1+[3.322*⋅lg(n)*]*=*1+[3.322*⋅lg*(49)]=1+5=6

На рисунках 1-2 показаны возможные варианты карманов и соответствующие им гистограммы.

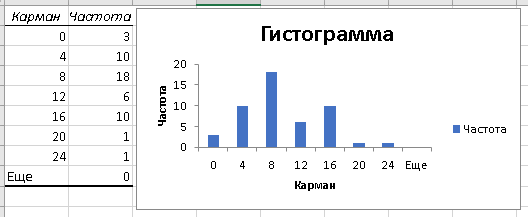


Рисунок 1-Результат построения гистограммы для кармана по формуле Стерджеса

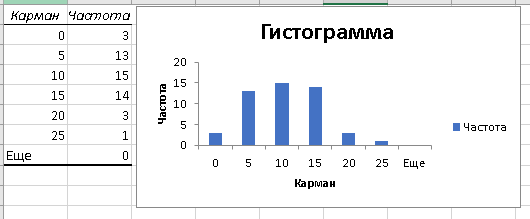


Рисунок 2-Результат построения гистограммы для пяти интервалов

Для построения эмпирической функции распределения (рисунок 3) сначала нужно получить вариационный ряд

*tв1≤ tв2≤ … ≤ tвn*,

затем определить эмпирические вероятности *Fi = i/n* (*i*=1, 2,…, *n*).

Рисунок 3- Эмпирическая функция распределения времени восстановления оборудования *tв*

Описательная статистика и гистограмма на пяти интервалах показывают, что *tв* имеет симметричное распределение и можно выдвинуть гипотезу о нормальном законе распределения с параметрами *a*=8,224 и *σ*=5,463 (рисунок 4).

Рисунок 4- Эмпирическая *Fi* и теоретическая *F(tв)* функции распределения времени восстановления оборудования *tв*

Для проверки гипотезы о нормальном распределении применим два критерия: Пирсона и Колмогорова. Следует задать вероятность *α*=0,05-0,1. Определить степень свободы *m*=6-1. Рассчитать значение критерия *χ2* по формуле:

где *ni* – количество значений выборки, попавших в *i*-ый интервал для гистограммы («частота» на рисунках 5-6), – значение теоретической функции распределения в точке .

Расчетное значение критерия *χ2* сравниваем с критическим значением, полученным с помощью стандартной функции в электронных таблицах для заданных *α* и *m* . Если расчетное значение больше критического, то нулевая гипотеза отклоняется.

Для применения критерия Колмогорова следует найти

и вычислить критическое значение по формуле:

*.*

Если произведение>*Kα*, то нулевая гипотеза отклоняется. Результаты проверки гипотезы о распределении времени восстановления оборудования *tв* приведены в таблице 7.

Таблица 7. Результаты применения критериев согласия

|  |  |
| --- | --- |
| Параметр | Значение |
| Вероятность *α* | 0,05 |
| Критерий Пирсона | |
| Степень свободы *m* | 5 |
| *χ2* расчетный | 1,66 |
| *χ2* критический | 11,07 |
| Вывод: | Нулевая гипотеза о нормальном распределении **не отклоняется** |
| Критерий Колмогорова | |
|  | 1,04 |
|  | 1,36 |
| Вывод: | Нулевая гипотеза о нормальном распределении **не отклоняется** |

Статистическая обработка данных о времени восстановления автомобилеразгрузчика показала, что выборка имеет размах 24 часа, может быть описана с вероятностью 95% нормальным законом с матожиданием 8,224 ч и дисперсией 29,844.

**Задача 9.** Провести статистическое оценивание интервалов прибытия ТС на терминал для разгрузки *Δt*, мин. Исходными данные получены с контрольно-пропускного пункта терминала и представлены в таблице 8.

Таблица 8. Данные о времени прибытия ТС

|  |  |
| --- | --- |
| ID ТС в системе | Время прибытия |
| 195478 | 13:15:00 |
| 195479 | 13:19:00 |
| 141564 | 13:23:00 |
| 195480 | 13:26:00 |
| 195481 | 13:27:00 |
| 321205 | 13:41:00 |
| 195482 | 13:52:00 |
| 595195 | 14:01:00 |
| 195483 | 14:16:00 |
| 195484 | 14:17:00 |

Известно, что *Δt* подчиняется экспоненциальному закону, функция распределения которого имеет вид:

,

где *λ -* интенсивность прибытия ТС и может быть определена по формуле:

,

где – матожидание случайной величины *Δt* .

Результат сохранить в файле с именем, содержащим «task9».

**Задача 10.** Провести статистическое оценивание параметра counter из файла «task4» (очищенные и преобразованные данные детекторов за трое суток) за каждые сутки отдельно. Проверить гипотезы о нормальном, логарифмически-нормальном, экспоненциальном распределениях. Сделать общий вывод. Результат сохранить в файле с именем, содержащим «task10».

**Дисперсионный анализ**

**Пример 7.** Провести однофакторный дисперсионный анализ средних скоростей движения ТС .

В исходных данных представлено 3 выборки средних скоростей, по 288 значений в каждой. Предварительная проверка показала, что очистка не требуется. Проводим однофакторный анализ для трех выборок объемом по 288 значений. N=288\*3=864. J=3. Исходными положениями дисперсионного анализа являются: случайный и независимый характер выборки ; нормальное распределение значений изучаемого признака в генеральной совокупности; равенство дисперсий в сравниваемых генеральных совокупностях. Результаты анализа трех выборок при уровне значимости α=0,05представлены на рисунке 5.



Рисунок 5 – Результат работы надстройки «однофакторный дисперсионный анализ»

Нулевая гипотеза *H0*состоит в равенстве средних значений выборки. Так как *F*>*Fкритическое*, нулевая гипотеза *H0*не принимается. Попарное сравнение выборок дает результаты приведенные в таблице 9.

Таблица 9. Результаты попарного дисперсионного анализа.

|  |  |  |  |
| --- | --- | --- | --- |
| Группы | *F* | *F критическое* | Вывод |
| speed1 | 6,7856 | 3,85 | *H0* не принимается |
| speed2 |
| speed2 | 10,56254 | 3,85 | *H0* не принимается |
| speed3 |
| speed3 | 35,73352 | 3,85 | *H0* не принимается |
| speed1 |

Проведем анализ средних скоростей за час наблюдений. По исходным данным получим средние скорости движения за час как среднее арифметическое.. Полученные выборки представлены в таблице 10.

Таблица 10. Средние скорости ТС

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Час | speed1 | speed2 | speed3 | Час | speed1 | speed2 | speed3 |
| 14 | 22 | 16 | 13 | 2 | 32 | 28 | 28 |
| 15 | 22 | 20 | 19 | 3 | 33 | 29 | 25 |
| 16 | 20 | 20 | 18 | 4 | 28 | 36 | 33 |
| 17 | 22 | 24 | 15 | 5 | 43 | 39 | 42 |
| 18 | 30 | 26 | 11 | 6 | 37 | 36 | 38 |
| 19 | 29 | 28 | 16 | 7 | 28 | 22 | 20 |
| 20 | 30 | 29 | 27 | 8 | 25 | 19 | 19 |
| 21 | 33 | 33 | 31 | 9 | 25 | 19 | 17 |
| 22 | 33 | 35 | 30 | 10 | 22 | 20 | 18 |
| 23 | 30 | 32 | 28 | 11 | 21 | 18 | 15 |
| 0 | 27 | 33 | 26 | 12 | 20 | 13 | 14 |
| 1 | 29 | 25 | 25 | 13 | 21 | 17 | 14 |

Результаты однофакторного дисперсионного анализа часовых средних скоростей ТС представлены в таблице 11.

Таблица 11. Результаты анализа для часовых средних скоростей ТС

|  |  |  |  |
| --- | --- | --- | --- |
| Группы | *F* | *F критическое* | Вывод |
| speed1 | 1,107 | 4,05 | *H0* принимается |
| speed2 |
| speed2 | 1,64 | 4,05 | *H0* принимается |
| speed3 |
| speed3 | 5,65 | 4,05 | *H0* не принимается |
| speed1 |
| speed1 | 2,80 | 3,13 | *H0* принимается |
| speed2 |
| speed3 |

Применение критерия Фишера при α=0,05для предложенных выборок дало противоречивые результаты, которые требуют дополнительных проверок апостериорных предположений (в частности, о нормальном распределении значений средних скоростей) и применения других критериев проверки гипотезы о равенстве средних значений выборки.

**Задача 11**. Провести однофакторный дисперсионный анализ для всех параметров из файла «task4» аналогично примеру 7 при уровнях значимости 0,05 и 0,1. Результат сохранить в файле с именем, содержащим «task11».

**Корреляционный анализ**

**Пример 8.** Провести корреляционный анализ параметров counter, speed, k из файла данных о параметрах транспортного потока.

В файле содержится 4 столбца по 288 чисел с данными о времени наблюдений и значениями параметров counter, speed, k. Проверка показала, что очистка данных не требуется.

Результат работы надстройки «корреляция» для выборок объема 288 и средних за час наблюдений объема 24 показан в таблице 12.

Таблица 12 . Парные коэффициенты корреляции *RXY*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *n*=288 | | | | *n*=24 | | | |
|  | counter | speed | k |  | counter | speed | k |
| counter | 1 |  |  | counter | 1 |  |  |
| speed | -0,432 | 1 |  | speed | -0,633 | 1 |  |
| k | 0,895 | -0,657 | 1 | k | 0,943 | -0,804 | 1 |

Проверка значимости коэффициентов корреляции проводится по критерию Стьюдента для уровня значимости α=0,05. Нулевая гипотеза *H0*состоит в том, что парный коэффициент корреляции *RXY* не значим. Критическое значение коэффициента Стьюдента *tкр* для числа степеней свободы *n*-2и α=0,05 , расчетное значение коэффициента Стьюдента и вывод о значимости коэффициента корреляции для пар X и Y представлены в таблице 13.

Таблица 13. Результаты проверки гипотезы о значимости *RXY*

|  |  |  |  |
| --- | --- | --- | --- |
| *X* | *Y* | *t* | Вывод |
| *n*=288 | | | *tкр*(α, *n*-2) = 1,968 |
| speed | counter | 8,105 | *H0* не принимается, *RXY*  - значим |
| k | counter | 33,83 | *H0* не принимается, *RXY*  - значим |
| k | speed | 14,75 | *H0* не принимается, *RXY*  - значим |
| *n*=24 | | | *tкр*(α, *n*-2) = 2,074 |
| speed | counter | 3,84 | *H0* не принимается, *RXY*  - значим |
| k | counter | 13,32 | *H0* не принимается, *RXY*  - значим |
| k | speed | 6,35 | *H0* не принимается, *RXY*  - значим |

Можно сделать вывод, что между параметрами counter и k сильная статистическая связь, между speed и k сильная или средняя связь, а counter и speed имеют среднюю и слабую статистическую связь для предложенных выборок и средних за час наблюдений, соответственно.

**Задача 12**. Провести корреляционный анализ для всех параметров из файла «task4» аналогично примеру 8 при α=0,05 и α=0,1. Результат сохранить в файле с именем, содержащим «task12».

**Регрессионный анализ**

**Пример 9**. Получить линейную многофакторную регрессионную модель по данным о транспортном потоке и провести ее анализ.

В файле с исходными данными, предложенными для анализа, содержится 4 столбца по 24 числа. Наименования столбцов: час, counter, speed, k. Очистка данных не требуется.

Найдем регрессионное уравнение вида Y=b0+b1x1+b2x2+b3x3, полагая x1 — это «час», x2 - counter, x3 - speed, Y – k. С помощью надстройки «регрессия» получены необходимые статистические показатели, основные из них приведены в таблице 14.

Таблица 14. Коэффициенты регрессии и расчетные значения *t*

|  |  |  |
| --- | --- | --- |
| Коэффициенты регрессии | | t-статистика |
| b0 | 614,7358071 | 1,961305177 |
| b1 | 2,065860849 | 0,193830681 |
| b2 | 0,563556305 | 1,973901532 |
| b3 | -6,955962264 | -2,036992602 |

Там же определены коэффициент корреляции *R*=0,8066 и коэффициент Фишера *F*=12,415.

Для проверки значимости коэффициента корреляции (аналогично примеру 8) используем распределение Стьюдента: критическое значение коэффициента Стьюдента *tкр*(0,05,22) = 2,0738; расчетное *t*=6,4. Так как *t*> *tкр* , то коэффициент корреляции значим *R*.

Аналогично проверим значимость коэффициентов регрессии. Нулевая гипотеза *H0*состоит в том, что bj =0 (*j*=0, 1, 2, 3). Критическое значение коэффициента Стьюдента для числа степеней свободы *n-(m-1)* (*m –* число коэффициентов регрессии)и α=0,05  *tкр*(0,05,20) = 2,085. Так как все расчетные значения t по модулю меньше критического, то нулевая гипотеза принимается. В полученном уравнении нет значимых коэффициентов.

Проверка адекватность модели проводится по критерию Фишера. Нулевая гипотеза *H0*состоит в том, регрессионная модель не адекватна экспериментальным данным. Если расчетное значение коэффициента Фишера больше критического то, гипотеза отвергается. Критическое значение для α=0,05 и двух степеней свободы *n-m* и (*m*-1): *Fкр*( 0,05;20;3)=8,66. Так как *F*>*Fкр,* то гипотеза *H0* отвергается, модель адекватна.

Получены противоречивые результаты: с одной стороны исследуемые параметры имеют достаточно сильную статистическую связь *R*=0,8066 и этот коэффициент корреляции значим с вероятностью 0,95 и модель (линейная многофакторная) адекватна, с другой стороны все коэффициенты регрессии не значимы.

Проверим значимость коэффициентов регрессии при α=0,075. Критическое значение коэффициента Стьюдента  *tкр*(0,075,20) = 1,878.  При таком уровне α коэффициенты b0, b2, b3 значимы.

Далее отбрасываем фактор x1 и проводим анализ для трех выборок. Получаем уравнение вида Y=b0+b1x1+b2x2, полагая x1 — это counter, x2 - speed , Y – k. С помощью надстройки «регрессия» получены необходимые статистические показатели, основные из них приведены в таблице 15.

Таблица 15. Коэффициенты регрессии и расчетные значения *t*

|  |  |  |
| --- | --- | --- |
| Коэффициенты регрессии | | t-статистика |
| b0 | 581,0991785 | 2,279211741 |
| b1 | 0,610276234 | 4,082909696 |
| b2 | -6,571785876 | -2,419402355 |

Там же определены коэффициент корреляции *R*=0,8066 и коэффициент Фишера *F*=19,49.

Для проверки значимости коэффициента используем распределение Стьюдента: *tкр*(0,05,22) = 2,0738; расчетное *t*=6,4. Так как *t*> *tкр* , то коэффициент корреляции значим *R*. Критическое значение коэффициента Фишера для α=0,05: *Fкр*(0,05;21;2)=19,45. Так как *F*>*Fкр,* то гипотеза *H0* отвергается, модель адекватна. Проверим значимость коэффициентов регрессии: критическое значение *tкр*(0,05,22) = 2,074. Так как все расчетные значения t по модулю больше критического, то нулевая гипотеза не принимается. В полученном уравнении все коэффициенты регрессии значимы.

Таким образом, предложенные для анализа данные можно описать многофакторным линейным уравнением :

k = 581,0991785+0,610276234\* counter - 6,571785876\* speed

Все параметры имеют статистическую связь с коэффициентом корреляции *R*=0,8066, значимость которого подтверждается с вероятностью 0,95. Полученная модель адекватна с вероятностью 0,95.

**Задача 13**. Получить линейную многофакторную регрессионную модель по данным из файла «task4» и провести ее анализ. Результат сохранить в файле с именем, содержащим «task13».

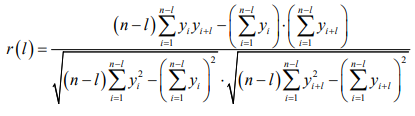
**Задача 14.** Применить методы регрессионного анализа для построения основной диаграммы транспортного потока для исходных данных из файла «task4». Результат сохранить в файле с именем, содержащим «task14».

**Задача 15.** Подобрать однофакторную нелинейную регрессионную модель с максимальным коэффициентом корреляции для исходных данных из файла «task15».

**Анализ временных рядов**

**Пример 10.** Построить коррелограмму и провести ее анализ. В файле с исходными данными, предложенными для анализа, содержится 2 столбца по 72 числа. Наименования столбцов: час, counter, speed, k. Очистка данных не требуется.

Найдем корреляционную функцию *r(l)* с помощью встроенного макроса cor\_gr() по формуле:

,

где *n* – число наблюдений, *l* – сдвиг (временной лаг), *yi* – значение случайной величины в *i*-ый момент времени. Для реализации макроса требуется разместить *yi* в столбце С, *n* – в ячейке E1, *l* – в ячейке E2. Рекомендуется выбирать *l*<*n*/4.

Результаты вычислений приведены в таблице 16, соответствующая коррелограмма – на рисунке 6.

Таблица 16 Значения корреляционной функции *r(l)* , расчетное *t* и выводы

|  |  |  |  |
| --- | --- | --- | --- |
| *l* | *r(l)* | *t* | Вывод |
| 1 | 0,902743 | 17,55753 | значим |
| 2 | 0,756799 | 9,686909 | значим |
| 3 | 0,602583 | 6,317313 | значим |
| 4 | 0,408158 | 3,740659 | значим |
| 5 | 0,186742 | 1,590376 | не значим |
| 6 | -0,04336 | 0,363102 | не значим |
| 7 | -0,23074 | 1,984021 | не значим |
| 8 | -0,39185 | 3,563426 | значим |
| 9 | -0,51298 | 4,999913 | значим |
| 10 | -0,58432 | 6,024267 | значим |
| 11 | -0,62976 | 6,783029 | значим |
| 12 | -0,62669 | 6,728378 | значим |
| 13 | -0,60333 | 6,329666 | значим |
| 14 | -0,57753 | 5,918871 | значим |
| 15 | -0,50429 | 4,885939 | значим |
| 16 | -0,37635 | 3,398686 | значим |
| 17 | -0,18657 | 1,588859 | не значим |

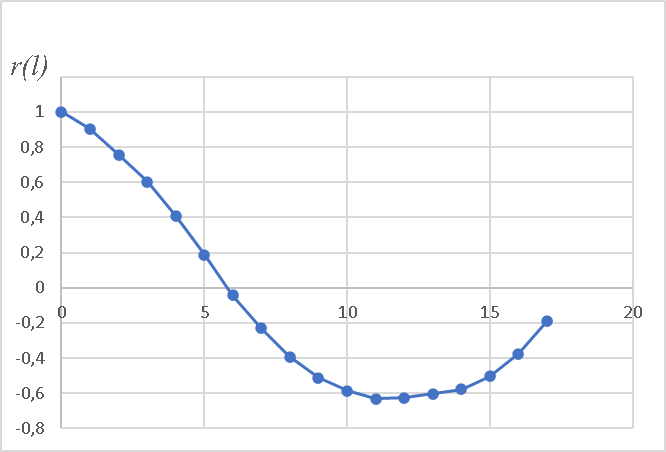


Рисунок 6 – Коррелограмма параметра counter для *l*=17

Значимость *r(l)* проверяется по критерию Стьюдента аналогично примеру 8. Критическое значение коэффициента Стьюдента *tкр* для числа степеней свободы *n*-2и α=0,05 равно 1,994. Расчетные коэффициенты Стьюдента *t* и выводы о значимости - в таблице 16.

Так как наиболее высоким оказался коэффициент автокорреляции *r(1)=*0.927, то исследуемый ряд содержит только трендовую составляющую.

Наличие не значимых *r(l)* может свидетельствовать о том, что временной ряд содержит нелинейный тренд, для выявления которого необходимо провести дополнительный анализ.

Увеличение лага до 70 приводит к появлению нового максимального значения *r(*68*)*=0,93 , что потенциально свидетельствует о существовании периодической составляющей (рисунок 7). Для более точного вывода необходимо большее количество наблюдений.

Подтвердить наличие тренда можно, проверив гипотезу о наличии неслучайной составляющей, которая, по существу, состоит в проверке гипотезы о постоянстве среднего значения временного ряда. Нулевая гипотеза *H0*состоит в равенстве средних значений. Разделим ряд на 2 части, c с равным количеством наблюдений *nI = nII*=36, найдем дисперсии *sI2 sII2* и применим критерий Фишера.

Расчетное *F* = max{*sI2*, *sII2* }/ min{*sI2*, *sII2* } = 37913/34245 = 1,107.

Критическое *Fкр*( 0,05;  *nI*-1 ;*nII*-1) = 1,757.

Так как *F*<*Fкр*, то нулевая гипотеза *H0*принимается и можно сделать вывод о наличии неслучайной составляющей. Это подтверждает ранее сделанное заключение о наличии тренда.

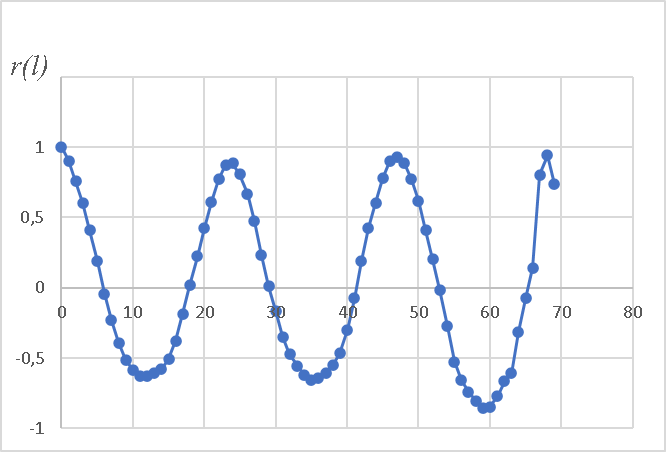


Рисунок 7 – Коррелограмма параметра counter для *l*=70

**Задача 16**. Построить коррелограммы и провести из анализ аналогично примеру 10 для исходных данных из файла «task4». Результат сохранить в файле с именем, содержащим «task16».

**Пример 11.** Применить гармонический анализ к исходным данным о транспортном потоке.

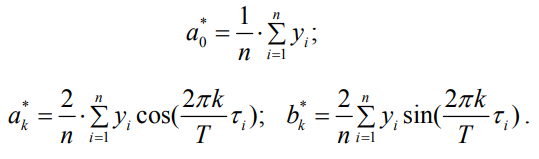
В исходном файле 4 столбца по 24 числа – данные о часовой интенсивности движения за 3 дня. Столбцы имеют метки hour, data1, data2, data3. Очистка данных не требуется.

Графики предложенных для анализа временных рядов показывают колебательный характер изменения значений (рисунок 8), поэтому возможно применение методов гармонического анализа.

Периодическую составляющую временного ряда найдем в виде ряда Фурье:



где *k*- номер гармоники ( *k*= 0, 1,…), *Т* – период, τ – время, *ak* и *bk* – коэффициенты, оценки которых *ak\** и *bk\** можно найти по приближенным формулам для наблюдений *yi*:



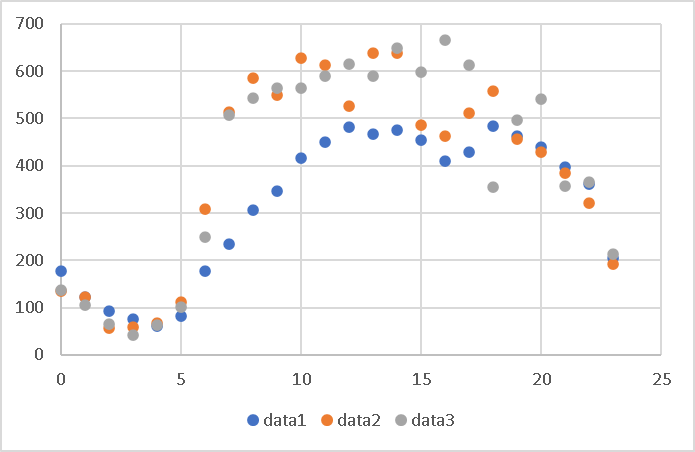


Рисунок 8 – Зависимость интенсивности от времени в разные дни

Значения временного ряда заданы в дискретные моменты времени τi и представляют собой арифметическую прогрессию с шагом Δ :

τ =τнач + (*i*-1)Δ, i = 1,2,…,*n*.

В качестве периода принимается величина *Т*= *n*Δ = 24, τнач=0, *n*=24. Наблюдения *yi* – средние за три дня значения интенсивности. Промежуточные результаты вычислений представлены в таблице приложения 2. Оценки коэффициентов Фурье: *a0*=368,9;*a1*=-207,3;*b1*=124,3;*a2*=-5,8;*b2*=-82,0.

Известно, что величины Sk = *ak2 + bk2* характеризуют спектральный состав (или спектр) функции *φ(*τ*).* Сравнительно не большие величины S*k* указывают на слагаемые, которыми можно пренебречь. S*0* =136120, S*1* =58410, S*2* =6756, S*3* =233. Четвертой гармоникой можно пренебречь, т.к. S*3* значительно меньше S*0*. Построим периодическую составляющую по двум и по трем гармоникам и сравним результаты на графике. На рисунке 9 представлены средние по трем дням интенсивности движения *y*, периодические составляющие временного ряда по двум гармоникам *φ1* и по трем гармоникам *φ2*.

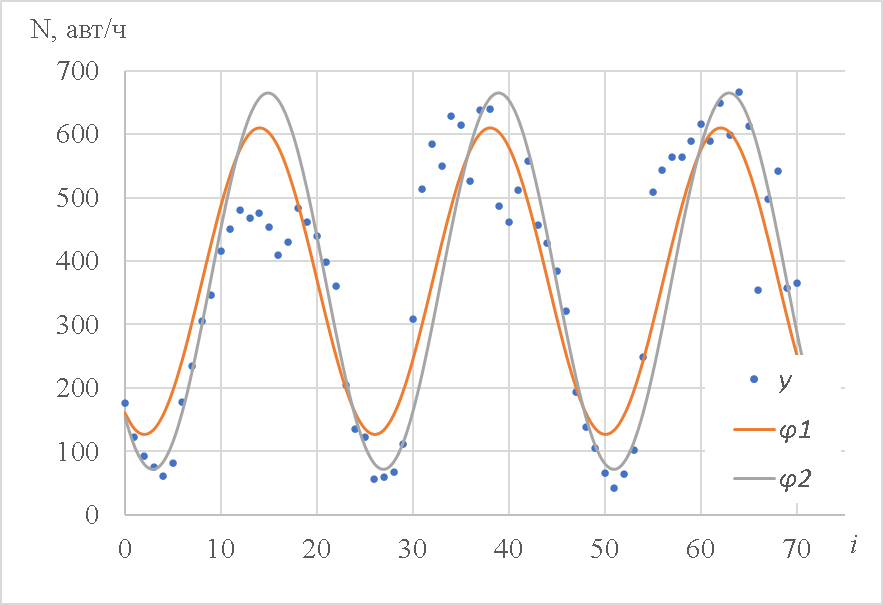


Рисунок 9. Результат применения гармонического анализа

Для проверки адекватности полученной гармонической модели применяется распределение Стьюдента. Нулевая гипотеза *H0*состоит в равенстве нулю математического ожидания разницы между *yi* и *φ(*τ*i).* Расчетное значение коэффициента Стьюдента . Для предоставленных исходных данных *t*=0. Критическое значение коэффициента Стьюдента *tкр*(0,05,22) = 2,0738 . Так как *t<tкр*, то гипотеза принимается и модель можно считать адекватной.

**Задача 17**. Применить гармонический анализ к исходным данным о транспортном потоке из файла «task4» аналогично примеру 11. Результат сохранить в файле с именем, содержащим «task17».

**Задача 18**. Сравнить модели из задач 15 и 17. Результат сохранить в файле с именем, содержащим «task18».

**Приложение 1**

qr-код для доступа к исходным данным

Ссылка для доступа к исходным данным <https://do.skif.donstu.ru/course/view.php?id=4610>

**Приложение 2**

Таблица. Результаты промежуточных вычислений для оценок коэффициентов Фурье

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| τi | *yi* | τi(2*π/T*) | *yi*cos(τi(2*π/T*)) | *yisin*(τi(2*π/T*)) | *yi*cos(τi(4*π/T*)) | *yisin*(τi(4*π/T*)) | *yi*cos(τi(6*π/T*)) | *yisin*(τi(6*π/T*)) |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 0 | 150,3 | 0,0 | 150,3 | 0,0 | 150,3 | 0,0 | 150,3 | 0,0 |
| 1 | 116,7 | 0,3 | 112,7 | 30,2 | 101,0 | 58,3 | 82,5 | 82,5 |
| 2 | 72,0 | 0,5 | 62,4 | 36,0 | 36,0 | 62,4 | 0,0 | 72,0 |
| 3 | 59,7 | 0,8 | 42,2 | 42,2 | 0,0 | 59,7 | -42,2 | 42,2 |
| 4 | 64,7 | 1,0 | 32,3 | 56,0 | -32,3 | 56,0 | -64,7 | 0,0 |
| 5 | 98,7 | 1,3 | 25,5 | 95,3 | -85,4 | 49,3 | -69,8 | -69,8 |
| 6 | 245,3 | 1,6 | 0,0 | 245,3 | -245,3 | 0,0 | 0,0 | -245,3 |
| 7 | 418,7 | 1,8 | -108,4 | 404,4 | -362,6 | -209,3 | 296,0 | -296,0 |
| 8 | 478,3 | 2,1 | -239,2 | 414,2 | -239,2 | -414,2 | 478,3 | 0,0 |
| 9 | 487,0 | 2,4 | -344,4 | 344,4 | 0,0 | -487,0 | 344,4 | 344,4 |
| 10 | 536,0 | 2,6 | -464,2 | 268,0 | 268,0 | -464,2 | 0,0 | 536,0 |
| 11 | 551,3 | 2,9 | -532,5 | 142,7 | 477,5 | -275,7 | -389,9 | 389,9 |
| 12 | 541,0 | 3,1 | -541,0 | 0,0 | 541,0 | 0,0 | -541,0 | 0,0 |
| 13 | 565,0 | 3,4 | -545,7 | -146,2 | 489,3 | 282,5 | -399,5 | -399,5 |
| 14 | 587,7 | 3,7 | -508,9 | -293,8 | 293,8 | 508,9 | 0,0 | -587,7 |
| 15 | 513,3 | 3,9 | -363,0 | -363,0 | 0,0 | 513,3 | 363,0 | -363,0 |
| 16 | 512,3 | 4,2 | -256,2 | -443,7 | -256,2 | 443,7 | 512,3 | 0,0 |
| 17 | 518,0 | 4,5 | -134,1 | -500,3 | -448,6 | 259,0 | 366,3 | 366,3 |
| 18 | 464,7 | 4,7 | 0,0 | -464,7 | -464,7 | 0,0 | 0,0 | 464,7 |
| 19 | 471,7 | 5,0 | 122,1 | -455,6 | -408,5 | -235,8 | -333,5 | 333,5 |
| 20 | 469,7 | 5,2 | 234,8 | -406,7 | -234,8 | -406,7 | -469,7 | 0,0 |
| 21 | 379,7 | 5,5 | 268,5 | -268,5 | 0,0 | -379,7 | -268,5 | -268,5 |
| 22 | 349,3 | 5,8 | 302,5 | -174,7 | 174,7 | -302,5 | 0,0 | -349,3 |
| 23 | 203,7 | 6,0 | 196,7 | -52,7 | 176,4 | -101,8 | 144,0 | -144,0 |