Лабораторная работа 3. Модели статистического моделирования и прогнозирования динамических систем по временному ряду(на основе МНК)

Цель работы

Цель настоящей работы – освоить средства моделирования стохастических временных рядов.

Ход работы

1. Ознакомиться со справочными сведениями.
2. Сформулировать задачу МНК при построении функции регрессии.
3. Разработать программу, моделирующую алгоритм поиска оптимального решения для формализованной задачи, используя математический пакет MatLab или язык программированияPython:
    a. Самостоятельно реализовать МНК для решения задачи поиска коэффициентов модели, заданной в виде полинома второго порядка $f_1(x) = a_2x^2 + a_1x + a_0$.
    b. С использованием встроенной реализации МНК в MatLab илиPython подобрать степень p полиномиальной модели$f_2(x) = \sum_{i=0}^{p} a_i x^i$, наилучшим образом соответствующей исходным данным при визуальной оценке на графике. Для этого построить график с исходными данными (крестики, точки и т.п.) и различными вариантами полиномиальных моделей степени p, где $p \neq 2$.
    c. Аппроксимировать данные функциональной моделью вида $f_3(x) = \sqrt[3]{x+1} + 1$.
    d. Используя скорректированный коэффициент детерминации R^2_{adj}определить наилучшую из трех моделей $f_1(x)$, $f_2(x)$, $f_3(x)$.
4. Сделать прогноз на один шаг. Указать, каким образом можно оценить точность прогноза.
5. Составить и представить преподавателю отчет о работе.
6. Уметь формулировать основные понятия, связанные с МНК, приводить необходимые формулы и их обоснования.

Исходные данные: Варианты задач в Приложении 3 по номеру студента в списке.

Справочные сведения

В рамках лабораторной работы 3 рассматривается метод численного моделирования функции по экспериментальным данным с целью аппроксимации фактических данных (с целью прогнозирования, в том числе).

Для сравнения моделей между собой обычно используют оценку погрешностей аппроксимации или коэффициент детерминации. В данной лабораторной работе предлагается использовать последний.

Коэффициент детерминации модели описывает долю дисперсии зависимой переменной y, объясняемую моделью. В общем случае коэффициент детерминации можно вычислить по формуле:

$$R^2 = 1 - \frac{D[y|x]}{D[y]} = 1 - \frac{\sigma^2}{\sigma_y^2},$$

где $D[y|x] = \sigma^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k-1}$ – условная дисперсия ошибки модели, исправленная нормирующим коэффициентом, $D[y] = \sigma_y^2$ – дисперсия случайной величины y. Здесь $\hat{y}_i = f_j(x_i)$ – результат j-ой модели в точке x_i,n – количество наблюдений за переменными x и y, k – количество параметров j-ой модели. Чем ближе значение коэффициента детерминации к единице, тем лучше данная модель описывает исходные данные.

Коэффициент детерминации обладает существенным недостатком: при увеличении количества параметров k, входящих в модель, его величина растет. Поэтому на практике обычно используют скорректированный коэффициент детерминации, лишенный данного недостатка:

$$R_{adj}^2 = 1 - (1 - R^2)\frac{n-1}{n-k}.$$

Реализация МНК в математических пакетах осуществляется с помощью функции `polyfit` в MatLab или `numpy.polyfit` в Python.

Рекомендуемая литература для лабораторной работы 3.

1. Бокс Дж., Дженкинс Г. Анализ временных рядов. Прогноз и управление. М.: Мир, 1974. Выпуск 1, 2.
2. Воронцов К.В., Егорова Е.В. Динамически адаптируемые композиции алгоритмов прогнозирования // Искусственный Интеллект. – № 10. - 2006. – С. 277–280.
3. Гребенников А.В, Крюков Ю.А, Чернягин Д. В. Моделирование сетевого трафика и прогнозирование с помощью модели ARIMA.
4. Безручко Б.П., Смирнов Д.А. Статистическое моделирование по временным рядам. Учебнометодическое пособие. Саратов: Издательство ГосУНЦ “Колледж”. 2000. 23 с.
5. Афанасьев В.Н., Цыпин А.П. Эконометрика в пакете STATISTICA: учебное пособие по выполнению лабораторных работ. Оренбург: ГОУ ОГУ, 2008. 204 с.
6. Дуброва Т.А. Статистические методы прогнозирования: учеб. пособие для вузов. – М.: ЮНИТИДАНА, 2003. – 206 с.

ПРИЛОЖЕНИЕ 4. Пример оформления лабораторной работы 4 «Компьютерное моделирование временного ряда по методу МНК»

Лабораторная работа № XXX.

Компьютерное моделирование временного ряда по методу МНК.
Выбор наилучшей модели.

Вариант 1. Выполнил: Иванов И.И., гр. 4536

Исходные данные: экспериментальные данные о значениях показателей *X*и *Y*

X,Y	*i*=1	*i*=2	*i*=3	*i*=4	*i*=5
x_i	0	1	2	4	5
y_i	2,1	2,4	2,6	2,8	3,0

Задание: В результате предварительной аппроксимации данных получена зависимость (тренд) – функциональная модель $g(x)=\sqrt[3]{x+1}+1$. Используя метод наименьших квадратов, аппроксимировать эти данные линейной моделью $y=ax+b$ (найти параметры *a* и *b*). Выяснить, какая из двух моделей лучше (адекватность в смысле метода наименьших квадратов) моделирует экспериментальные данные. Сделать чертеж. Сделать прогноз в момент *i*=6 по «лучшей» модели.

Ход выполнения задания

1. Согласно методу наименьших квадратов (МНК) задача заключается в нахождении коэффициентов линейной зависимости, при которых функция двух переменных*a* и *b* $F(a,b)=\sum_{i=1}^{n}\left(y_i-\left(ax_i+b\right)\right)^2 \xrightarrow[a,b]{} \min$ (принимает наименьшее значение).

Решение примера сводится к нахождению экстремума функции двух переменных.

2. Вывод формул для нахождения коэффициентов *a* и *b*.

Составляется и решается система из двух уравнений с двумя неизвестными. Находим частные производные функции $F(a,b)$ по переменным *a* и *b*, приравниваем эти производные к нулю:

$$\frac{\partial F}{\partial a}=-2\sum_{i=1}^{n}\left(y_i-\left(ax_i+b\right)\right)x_i=0,$$

$$\frac{\partial F}{\partial b}=-2\sum_{i=1}^{n}\left(y_i-\left(ax_i+b\right)\right)=0.$$

Получаем соотношения для *a* и *b* в виде:

$$a=\frac{n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2},\ b=\frac{\sum_{i=1}^{n}y_i-a\sum_{i=1}^{n}x_i}{n}.$$

Убедимся, что в найденной стационарной точке (a,b) функция $F(a,b)$

принимает минимум. Дифференциал второго порядка должен быть положительно определенным, или матрица квадратичной формы дифференциала второго порядка для функции $F(a,b)$, в точке (a,b) она должны быть положительно определенной (по критерию Сильвестра).

Дифференциал второго порядка имеет вид:

$$d^2F = \frac{\partial^2 F}{\partial a^2}d^2a + 2\frac{\partial^2 F}{\partial a \partial b}dadb + \frac{\partial^2 F}{\partial b^2}d^2b.$$

Находим соответствующие величины:

$$\frac{\partial^2 F}{\partial a^2} = 2\sum_{i=1}^{n} x_i^2,\ \frac{\partial^2 F}{\partial b^2} = 2n,\ \frac{\partial^2 F}{\partial a \partial b} = 2\sum_{i=1}^{n} x_i$$

Тогда $d^2F = 2\sum_{i=1}^{n} x_i^2 d^2a + 4\sum_{i=1}^{n} x_i dadb + 2nd^2b$, а матрица квадратичной формы имеет вид

$$D = \begin{bmatrix} \frac{\partial^2 F}{\partial a^2} & \frac{\partial^2 F}{\partial a \partial b} \\ \frac{\partial^2 F}{\partial a \partial b} & \frac{\partial^2 F}{\partial b^2} \end{bmatrix} = \begin{bmatrix} 2\sum_{i=1}^{n} x_i^2 & 2\sum_{i=1}^{n} x_i \\ 2\sum_{i=1}^{n} x_i & 2n \end{bmatrix}.$$

Значения элементов не зависят от a и b, главные миноры положительны (докажите методом математической индукции), следовательно, характер экстремума определен по критерию Сильвестра.

3. Заполняем таблицу для удобства нахождения коэффициентов a и b.

X,Y	i=1	i=2	i=3	i=4	i=5	$\sum_{i=1}^{5}(\ )$
x_i	0	1	2	4	5	$\sum_{i=1}^{n} x_i = 12$
y_i	2,1	2,4	2,6	2,8	3,0	$\sum_{i=1}^{n} y_i = 12{,}9$
$x_i y_i$	0	2,4	5,2	11,2	15	$\sum_{i=1}^{n} x_i y_i = 33{,}8$
x_i^2	0	1	4	16	25	$\sum_{i=1}^{n} x_i^2 = 46$

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \approx 0{,}165,\ b = \frac{\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i}{n} \approx 2{,}184.$$

4. Записываем итоговое выражение для линейной модели

$$y(x) = ax + b = 0{,}165x + 2{,}184.$$

5. Для ответа на вопрос: какая из кривых

$y(x) = ax + b = 0{,}165x + 2{,}184$ или $g(x) = \sqrt[3]{x+1} + 1$

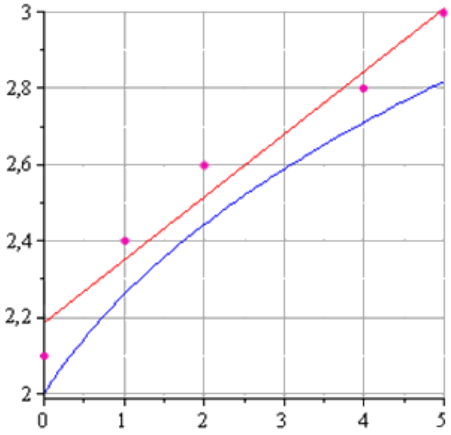
лучше аппроксимирует исходные данные, следует оценить погрешности аппроксимаций по формулам:

$$\sigma_1 = \sum_{i=1}^{n} \left(y_i - (ax_i + b)\right)^2 \approx 0{,}019,\ \ \sigma_2 = \sum_{i=1}^{n} \left(y_i - g(x_i)\right)^2 \approx 0{,}096.$$

Сравнивая $\sigma_1 < \sigma_2$, делаем **вывод**: прямая $y = 0{,}165x + 2{,}184$ является лучшим

приближением исходных данных по сравнению с кривой $g(x)=\sqrt[3]{x+1}+1$.

6. Графическая интерпретация степени близости кривых к исходным данным: кривая $y(x)=0{,}165x+2{,}184$ (красное начертание), $g(x)=\sqrt[3]{x+1}+1$ (синее начертание), исходные данные – набор точек $(x_i, y_i), i=\overline{1,5}$ (сиреневый цвет начертания).



ПРИЛОЖЕНИЕ 3. Варианты для моделирования временного ряда к ЛР-3

Вариант 1

Изучается динамика потребления молока в регионе. Для этого собраны данные об объемах среднедушевого потребления мяса (кг) $Y(t)$ за 7 месяцев. Обосновать и построить тренд данного ряда. Оценить достоверность уточненной по МНК модели.

t	1	2	3	4	5	6	7
$Y(t)$	8,16	8,25	8,41	8,76	9,2	9,78	10,1

Вариант 2

Банк изучает динамику изменения величины депозитов физических лиц за несколько лет (млн.$ в сопоставимых ценах). Обосновать и построить тренд данного ряда. Оценить достоверность уточненной по МНК модели.

Время, t	1	2	3	4	5	6	7
Размер депозитов, $Y(t)$	2	6	7	3	10	12	13

Вариант 3

Изучается динамика рождаемости в России. Собраны данные о числе рожденных (млн) $Y(t)$ за 7 лет (2009-2015). Обосновать и построить тренд данного ряда. Оценить достоверность уточненной по МНК модели.

t	2009	2010	2011	2012	2013	2014	2015
$Y(t)$	1,767	1,788	1,796	1,902	1,895	1,947	1,944

Вариант 4

Изучается динамика потребления сахара в России. Для этого собраны данные об объемах среднедушевого потребления сахара (г/сутки) $Y(t)$ за 7 десятилетий. Обосновать и построить тренд данного ряда. Оценить достоверность уточненной по МНК модели.

t	1(1950)	2(1960)	3(1970)	4(1980)	5(1990)	6(2000)	7(2015)
$Y(t)$	32	85	115	130	130	96	107

Вариант 5
Изучается динамика потребления мяса птицы в Европе. Для этого собраны данные об объемах среднедушевого потребления мяса (кг/чел/год) $Y(t)$ за 10 лет (2000-2009). Обосновать и построить тренд данного ряда. Оценить достоверность уточненной по МНК модели.

t	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
$Y(t)$	16,0	17,9	18,6	18,3	19,0	19,3	19,2	20,3	21,1	21,9

Вариант 6
Изучается динамика потребления мяса птицы в Азии. Для этого собраны данные об объемах среднедушевого потребления мяса (кг/чел/год) $Y(t)$ за 10 лет (2000-2009). Обосновать и построить тренд данного ряда. Оценить достоверность уточненной по МНК модели.

t	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
$Y(t)$	6,7	6,6	6,8	7,0	7,0	7,5	7,7	8,2	8,6	8,8

Вариант 7
Изучается динамика потребления мяса птицы в Африке. Для этого собраны данные об объемах среднедушевого потребления мяса (кг/чел/год) $Y(t)$ за 10 лет (2000-2009). Обосновать и построить тренд данного ряда. Оценить достоверность уточненной по МНК модели.

t	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
$Y(t)$	4,2	4,3	4,5	4,7	4,6	4,7	4,8	5,2	5,4	5,5

Вариант 8
Изучается динамика объема депозитов и прочих средств,размещенных в банках в России. Для этого собраны данные об объемах указанных средств (млн.р.) $Y(t)$ за 12 месяцев 2017г. Обосновать и построить тренд данного ряда. Оценить достоверность уточненной по МНК модели.

t	1	2	3	4	5	6	7	8	9	10	11	12
$Y(t)$	10,64	10,61	10,64	10,73	10,84	10,92	11,04	11,19	11,38	11,54	11,69	11,88

Вариант 9

Изучается динамика объема кредитов, выданных в России. Для этого собраны данные об объемах указанных средств в иностранной валюте (усл.ед.) $Y(t)$ за 12 месяцев 2017г. Обосновать и построить тренд данного ряда. Оценить достоверность уточненной по МНК модели.

t	1	2	3	4	5	6	7	8	9	10	11	12
$Y(t)$	160330	152249	144849	134451	132446	129460	137186	135531	129578	122239	115773	118181

Вариант 10

Изучается динамика производства стали в мире. Для этого собраны данные об объемах ее производства (млн.т.) $Y(t)$ за первые 7 месяцев 2018 года. Обосновать и построить тренд данного ряда. Оценить достоверность уточненной по МНК модели.

t	1	2	3	4	5	6	7
$Y(t)$	145	132	149	149	155	152	155

Вариант 11

Изучается динамика производства чугуна в стране. Для этого собраны данные об объемах его производства (млн.т.)$Y(t)$ за 7 лет (1991-1996). Обосновать и построить тренд данного ряда. Оценить достоверность уточненной по МНК модели.

t	1	2	3	4	5	6	7
$Y(t)$	107	108	107	110	111	110	112

Вариант 12

Исследуется динамика цен на недвижимость. Для этого собраны данные о средней стоимости 1 м^2 на вторичном рынке жилья Санкт-Петербурга (руб.)$Y(t)$ за период с августа по октябрь 2019 года (шаг измерений – 14 дней). Обосновать и построить тренд данного ряда. Оценить достоверность уточненной по МНК модели.

t	1	2	3	4	5	6	7
$Y(t)$	143427,5	139720,1	137696,4	137833,2	136591,1	135856,2	135597,8

Вариант 13

Исследуется динамика производства стали. Для этого собраны данные об объемах ее производства (млн.т.)$Y(t)$ за первые 7месяцев 2018 года. Обосновать и построить тренд данного ряда. Оценить достоверность уточненной по МНК модели.

t	1	2	3	4	5	6	7
$Y(t)$	145	132	149	149	155	152	155

Вариант 14

Исследуется динамика цен на недвижимость. Для этого собраны данные о средней стоимости 1 м^2 в новостройках Санкт-Петербурга (руб.)$Y(t)$ за лето 2019 года (шаг измерений – 14 дней). Обосновать и построить тренд данного ряда. Оценить достоверность уточненной по МНК модели.

t	1	2	3	4	5	6	7
$Y(t)$	115113,8	116620,5	117377,2	116770,5	118621,8	118173,4	118447

Вариант 15

Исследуется динамика производства стали. Для этого собраны данные об объемах ее производства (млн.т.)$Y(t)$ за первые 7месяцев 2017 года. Обосновать и построить тренд данного ряда. Оценить достоверность уточненной по МНК модели.

t	1	2	3	4	5	6	7
$Y(t)$	138	127	143	142	145	143	146

Вариант 16

Исследуется динамика цен на недвижимость. Для этого собраны поквартальные данные о средней стоимости 1 м^2 на вторичном рынке жилья Москвы (руб.)$Y(t)$ за 2018-2019 года. Обосновать и построить тренд данного ряда. Оценить достоверность уточненной по МНК модели.

t	2018 (1)	2018 (2)	2018 (3)	2018 (4)	2019 (1)	2019 (2)	2019 (3)
$Y(t)$	168453,79	170645,34	171764,67	171177,91	177859,12	176030,45	188992,3

Вариант 17

Исследуется бедность населения. Для этого собрана информация о численности населения РФ с денежными доходами ниже величины прожиточного минимума(млн. чел., Федеральная служба государственной статистики)$Y(t)$ за 2010-2017 года. Обосновать и построить тренд данного ряда. Оценить достоверность уточненной по МНК модели.

t	2010	2011	2012	2013	2014	2015	2016	2017
$Y(t)$	17,7	17,9	15,4	15,5	16,1	19,5	19,5	19,4

Вариант 18

Собрана информация о дефиците денежного дохода населения Российской Федерации, имеющегодоход ниже величины прожиточного минимума (млрд. руб., Федеральная служба государственной статистики)$Y(t)$ за 2010-2017 года. Обосновать и построить тренд данного ряда. Оценить достоверность уточненной по МНК модели.

t	2010	2011	2012	2013	2014	2015	2016	2017
$Y(t)$	375,0	424,1	370,5	417,9	478,6	700,8	706,8	719,1

Вариант 19

Исследуется демографическая ситуация. Для этого собрана информация о численности населения Российской Федерации (млн. чел., Федеральная служба государственной статистики)$Y(t)$ за 2012-2019 года. Обосновать и построить тренд данного ряда. Оценить достоверность уточненной по МНК модели.

t	2012	2013	2014	2015	2016	2017	2018	2019
$Y(t)$	143,0	143,3	143,7	146,3	146,5	146,8	146,9	146,8

Вариант 20

Собрана информация о числе женщин на 1000 мужчин для возрастной группы 20-24 лет населения Российской Федерации (чел., Федеральная служба государственной статистики) $Y(t)$ за 2012-2019 года. Обосновать и построить тренд данного ряда. Оценить достоверность уточненной по МНК модели.

t	2012	2013	2014	2015	2016	2017	2018	2019
$Y(t)$	967	966	962	959	960	960	959	961

Вариант 21

Собрана информация о количестве разводов на 1000 человек населения Российской Федерации (шт., Федеральная служба государственной статистики)$Y(t)$ за 2011-2018 года. Обосновать и построить тренд данного ряда. Оценить достоверность уточненной по МНК модели.

t	2011	2012	2013	2014	2015	2016	2017	2018
$Y(t)$	4,7	4,5	4,7	4,7	4,2	4,1	4,2	4,0

Вариант 22

Изучается динамика потребления стали в мире. Для этого собраны данные об объемах ее потребления (млн.т.) $Y(t)$ с 2010 года. Обосновать и построить тренд данного ряда. Оценить достоверность уточненной по МНК модели.

t	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
$Y(t)$	1401	1492	1552	1613	1627	1591	1611	1693	1739	1761

Вариант 23

Исследуется информация о браках и разводах. Для этого собрана информация о количестве браков на 1000 человек населения Российской Федерации (шт., Федеральная служба государственной статистики)$Y(t)$ за 2011-2018 года. Обосновать и построить тренд данного ряда. Оценить достоверность уточненной по МНК модели.

t	2011	2012	2013	2014	2015	2016	2017	2018
$Y(t)$	9,2	8,5	8,5	8,4	7,9	6,7	7,1	6,1

Вариант 24

Собрана информация о числе женщин на 1000 мужчин для возрастной группы 15-19 лет населения Российской Федерации (чел., Федеральная служба государственной статистики) $Y(t)$ за 2012-2019 года. Обосновать и построить тренд данного ряда. Оценить достоверность уточненной по МНК модели.

t	2012	2013	2014	2015	2016	2017	2018	2019
$Y(t)$	959	956	954	953	955	957	956	956

Вариант 25

Изучается динамика производства стали в мире. Для этого собраны данные об объемах ее производства (млн.т.) $Y(t)$ за первые 7 месяцев 2017 года. Обосновать и построить тренд данного ряда. Оценить достоверность уточненной по МНК модели.

t	1	2	3	4	5	6	7
$Y(t)$	138	127	143	142	145	143	146

Вариант 26

Исследуется демографическая ситуация. Для этого собрана информация о числе женщин на 1000 мужчин для возрастной группы 30-34 лет населения Российской Федерации (чел., Федеральная служба государственной статистики)$Y(t)$ за 2012-2019 года. Обосновать и построить тренд данного ряда. Оценить достоверность уточненной по МНК модели.

t	2012	2013	2014	2015	2016	2017	2018	2019
$Y(t)$	1014	1009	1004	1001	1002	1000	994	992

Вариант 27

Исследуется социально-экономическое положение семей и тенденции их жизнедеятельности. Для этого собрана информация о числе многодетных семей, состоящих на учете в качестве нуждающихся в жилых помещениях на территории Российской Федерации(единиц, Федеральная служба государственной статистики)$Y(t)$ за 2012-2018 года. Обосновать и построить тренд данного ряда. Оценить достоверность уточненной по МНК модели.

t	2012	2013	2014	2015	2016	2017	2018
$Y(t)$	125198	124971	127194	131213	129207	131585	132323

Вариант 28

Исследуется социально-экономическое положение семей и тенденции их жизнедеятельности. Для этого собрана информация о числе многодетных семей, получивших жилые помещения и улучшивших жилищные условия на территории Российской Федерациив отчетном году(единиц, Федеральная служба государственной статистики)$Y(t)$ за 2012-2018 года. Обосновать и построить тренд данного ряда. Оценить достоверность уточненной по МНК модели.

t	2012	2013	2014	2015	2016	2017	2018
$Y(t)$	5268	6292	6751	5538	6013	4963	4367

Вариант 29

Исследуется дифференциация оплаты труда работников. Для этого собрана информация о медианной заработной плате работников в сфере образованияна территории Российской Федерации(руб., Федеральная служба государственной статистики)$Y(t)$ за 2005-2019 года. Обосновать и построить тренд данного ряда. Оценить достоверность уточненной по МНК модели.

t	2005	2007	2009	2011	2013	2015	2017	2019
$Y(t)$	3815	5818	9040	10186	15785	18925	20 363	53592

Вариант 30

Исследуется социальное обеспечение. Для этого собрана информация о среднем размере назначенных пенсий по старости (возрасту) в Российской Федерации на 1 января соответствующего года (руб., Федеральная служба государственной статистики)$Y(t)$ за 2008-2019 года. Обосновать и построить тренд данного ряда. Оценить достоверность уточненной по МНК модели.

t	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
$Y(t)$	4909,8	6630,1	8165,8	8876,1	9790,1	10716,4	11569,1	12830,4	13172,5	14151,6	14986,2