Министерство образования и науки Российской Федерации

Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования

Национальный минерально-сырьевой университет «Горный»

Кафедра информатики и компьютерных технологий

ИНФОРМАТИКА

МЕТОДЫ СТАТИСТИЧЕСКОГО АНАЛИЗА ДАННЫХ В СРЕДЕ MATHCAD

Методические указания по выполнению курсовой работы для студентов специальности 022000

САНКТ-ПЕТЕРБУРГ
2015

УДК 681.141.2 (075.83)

МЕТОДЫ СТАТИСТИЧЕСКОГО АНАЛИЗА ДАННЫХ В СРЕДЕ MATHCAD: Методические указания по выполнению курсовой работы / Национальный минерально-сырьевой университет «Горный». Сост. *И.,И.Пивоварова, Прудинский Г.А.*. СПб, 2015, 36 с.

Рассмотрены методы построения различных математических моделей инструментами математического редактора MathCad. Проведен сравнительный анализ возможностей двух вычислительных систем MathCad и MS Excel. Определены преимущества и недостатки применительно к решению конкретной задачи – математического анализа параметров геосистемы.

Методические указания предназначены для студентов специальности «Экология и природопользование»

Табл. 7. Ил. 8. Библиогр.: 6 назв.

Научный редактор доц. *А.Е.Ильин*

ВВЕДЕНИЕ

Специалисты в области защиты окружающей среды изучают сложные экологические системы, состояние и поведение которых зависит от множества условий: географических, метеорологических, промышленных, биологических. В экологических исследованиях при изучении количественных зависимостей различных показателей, значения которых определяются эмпирически, в большинстве случаев, приходится сталкиваться с изменчивостью исследуемых характеристик. Частично она задается неоднородностью самих изучаемых объектов, частично – обусловливается некоторой погрешностью наблюдений и количественной обработки материалов. Поэтому часто для описания поведения экосистемы применяется аппроксимация – приближенное описание корреляционной зависимости переменных подходящим уравнением функциональной зависимости.

Использование аппроксимирующих зависимостей при оценке воздействия на окружающую среду представляет собой процедуру, включающую определение возможных неблагоприятных воздействий на экологическую систему на основании обработки данных, полученных в ходе экологического мониторинга. Так, например, на заводах по выпуску алюминия, применяющих технологию «мокрой» газоочистки в составе отходящих газов присутствуют ПАУ, ПФУ, CO_2, $SO_{2,}$ газообразные и твердые фториды, другие соединения. В связи с этим знания о поведении и распределении сульфатов в окружающей среде приобретают первостепенную важность. Причем если предприятие находится на территории произрастания, например хвойных лесов, данная информация становится особенно актуальной в связи с общеизвестным фактом о негативном влиянии диоксида серы на хвойные породы деревьев. В данном случае берутся пробы на содержание сульфатов в фильтрате снеговой воды и почве. Строятся аппроксимирующие зависимости, а изменение содержания сульфатов описывается с помощью линейных уравнений. Аналогичные примеры можно привести по оценке наличия ионов металла в сточных водах. Так как, такие широко распространенные металлы, как медь, свинец, железо, никель и цинк, попадая в обычные канализационные стоки нарушают работу очистных систем и отравляют водоемы. Использование регрессионных и корреляционных соотношений применяется в некоторых методах расчета максимальных выбросов загрязняющих веществ от автотранспортных потоков на городских автомагистралях и т.д.

В отличие от физических моделей математические модели не требуют огромных вложений на проведение эксперимента, т.к. математическая модель является лишь формальным описанием объекта, а не его физическим аналогом. Кроме того, математическая модель за счет изменения ряда параметров, дает возможность изучить ее поведение в самых разнообразных условиях. Это дает основания называть исследование поведения модели при различных значениях входящих в нее параметров численным экспериментом. Построение математической модели - это необходимый этап при решении научной или технической задачи на компьютере.

Можно выделить некоторые классы моделей, учитывающие различные свойства систем: детерминистские модели и вероятностные модели, статические и динамические модели, линейные и нелинейные модели, оптимизационные модели и т. д. В ряде случаев модель строится в виде уравнений, требующих решения, причем это могут быть уравнения самого разнообразного вида: алгебраические, трансцендентные, системы линейных и нелинейных алгебраических уравнений, обыкновенные дифференциальные уравнения, уравнения с частными производными, интегральные уравнения.

Только наиболее простые математические модели могут сразу представить зависимость между переменными системы в явном виде. В курсовой работе рассматривается подбор именно таких - явных зависимостей, причем заранее не делается никаких предположений о физических свойствах наблюдаемой системы. Известными являются только совместно регистрируемые числовые значения некоторых признаков - величин, описывающих поведение системы. Ставится задача аппроксимации эмпирических табличных данных функцией заданного вида. Рассмотрены примеры построения аппроксимирующих функций одной переменной по таблице совместно наблюдаемых значений двух величин и аппроксимирующих функций двух переменных по таблице совместно наблюдаемых значений трех величин, две из которых считаются независимыми.

Следует отметить, что при наблюдении экологической системы, как правило, требуется учитывать огромное количество параметров. Например, в исследовании экосистемы Великих озер в США американскими учеными проводилась регистрация примерно 300 параметров. При изучении Ладожского озера осуществляется контроль примерно 22-24 параметров. Поэтому рассмотренные примеры носят учебный характер. Тем не менее, они полезны для понимания сути методики расчета и построения аналитических зависимостей, а также способов анализа эмпирических данных в экосистеме.

1. ЗАДАЧА АППРОКСИМАЦИИ ЭМПИРИЧЕСКИХ ДАННЫХ

Пусть состояние некоторой системы описывается рядом переменных - $x_1, x_2, \ldots, x_k, y_1, y_2, \ldots, y_l$. Где, например, $x_1, x_2, \ldots, x_k$ - характеристики производственных мощностей предприятия, $y_1, y_2, \ldots, y_l$- концентрация сульфатов натрия в грунте. Предположим, что величины - $x_1, x_2, \ldots, x_k$ - независимы, а величины $y_1, y_2, \ldots, y_l$ - от них зависят. Эта зависимость может быть выражена с помощью функциональной связи

$$\begin{aligned} y_1 &= f_1(x_1, x_2, \ldots, x_k) \\ y_2 &= f_2(x_1, x_2, \ldots, x_k) \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ y_l &= f(x_1, x_2, \ldots, x_k) \end{aligned}$$

Сделаем допущение, что нам неизвестен вид зависимостей, но мы должны определить его по эмпирическим данным. Пусть получена таблица, в которой даны наблюдаемые значения величин x_i, y_j:

Таблица 1.1

№	x_1	x_2	....	x_k	y_1	y_2	...	y_l
1	x_{11}	x_{12}	...	x_{1k}	y_{11}	y_{12}	...	y_{1l}
n	x_{n1}	x_{n2}	...	x_{nk}	y_{n1}	y_{n2}	...	y_{nl}

В первой строке таблицы указаны обозначения наблюдаемых величин, в следующих строках перечислены наблюдаемые значения.

Вначале мы рассмотрим простейший пример зависимости: величина y зависит от одной переменной x. Теоретические сведения по этому вопросу рассмотрены в разделе 2. Затем в разделе 3 будет рассмотрена зависимость наблюдаемой величины от двух переменных. Этот случай легко обобщается на зависимость от любого числа переменных.

2. АППРОКСИМАЦИЯ ФУНКЦИИ ОДНОЙ ПЕРЕМЕННОЙ

2.1.ТЕОРЕТИЧЕСКИЕ СВЕДЕНИЯ

2.1.1. Понятие аппроксимации

Пусть в результате наблюдений получена таблица совместно наблюдаемых значений x_i, y_i:

Таблица 2.1

x	x_1	x_2	$\dots$	x_i	$\dots$	x_n
y	y_1	y_2	$\dots$	y_i	$\dots$	y_n

Требуется найти некоторую функцию, заданную аналитически и удовлетворительно описывающую зависимость y от x. Приближенное представление исходной функции с помощью некоторой зависимости называется ее аппроксимацией. Выбор вида аппроксимирующей функции остается за исследователем и зависит от ряда соображений. Как правило, предпочтение отдается достаточно простым функциям: линейной, квадратичной, экспоненциальной, логарифмической, обратно пропорциональной.

Зачастую выбору конкретной зависимости помогает анализ графика, построенного по табличным данным, а также физические представления. Выберем аппроксимирующую функцию, зависящую от нескольких параметров:

$$y = f(x; a_1, a_2, \ldots, a_m), \tag{1}$$

Подставив в формулу (1) эмпирическое значение переменной $x = x_i$, получим теоретическое значение величины $y = y_i^{\text{т}}$, вычисленное по формуле

$$y_i^T = f(x_i; a_1, a_2, \ldots, a_m), \tag{2}$$

Разность $y_i^{\text{т}} - y_i$ называется отклонением и представляет ошибку аппроксимации данной табличной функции. Для оценки качества аппроксимации функции в целом требуется оценить суммарную ошибку.

2.1.2. Метод наименьших квадратов

Есть разные способы оценки суммарной ошибки аппроксимации. Чаще всего оценивают суммарную квадратичную ошибку, равную сумме квадратов отклонений эмпирических значений функции от теоретических:

$$S(a_1, a_2, \ldots, a_m) = \sum_{i=1}^{n} \left[y_i - f(x_i; a_1, a_2, \ldots, a_m) \right]^2 \tag{3}$$

Параметры a_1, a_2, ... , a_m должны быть определены из условия минимума суммарной квадратичной ошибки. Запишем необходимое условие экстремума функции многих переменных $S(a_1, a_2, \ldots, a_m)$ - равенство нулю ее частных производных

$$\frac{\partial S}{\partial a_1} = 0, \ \frac{\partial S}{\partial a_2} = 0, \ldots, \frac{\partial S}{\partial a_m} = 0 \tag{4}$$

Формулы (4) представляют собой систему m уравнений с m неизвестными для определения наилучших значений параметров. Если функция (1) линейна относительно параметров a_1, a_2, ... , a_m, то система (4) представляет собой систему линейных уравнений.

Метод определения параметров из условия минимума суммарной квадратичной ошибки называется методом наименьших квадратов.

2.1.3. Определение параметров аппроксимации

Рассмотрим различные случаи аппроксимации эмпирических данных функциями конкретного вида:

1. линейной функцией

$$y=a_1+a_2x \tag{5}$$

Тогда система (4) примет вид:

$$\begin{cases} a_1 n + a_2 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i, \\ a_1 \sum_{i=1}^{n} x_i + a_2 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i\, y_i. \end{cases} \tag{6}$$

2. квадратичной функцией

$$y=a_1+a_2x+a_3x^2. \tag{7}$$

Тогда система (4) будет иметь вид:

$$\begin{cases} a_1 n + a_2 \sum_{i=1}^{n} x_i + a_3 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i, \\ a_1 \sum_{i=1}^{n} x_i + a_2 \sum_{i=1}^{n} x_i^2 + a_3 \sum_{i=1}^{n} x_i^3 = \sum_{i=1}^{n} x_i\, y_i, \\ a_1 \sum_{i=1}^{n} x_i^2 + a_2 \sum_{i=1}^{n} x_i^3 + a_3 \sum_{i=1}^{n} x_i^4 = \sum_{i=1}^{n} x_i^2 y_i. \end{cases} \tag{8}$$

3. экспоненциальной функцией:

$$y = a_1 \cdot e^{a_2 x} \tag{9}$$

В этом случае нужно вначале линеаризовать формулу (9) с помощью логарифмирования, получим:

$$\ln y = \ln a_1 + a_2 x$$

К этому уравнению можно применить формулы (6), но с другими обозначениями. Введем обозначения:

$$z = \ln y,\ c = \ln a_1$$

Тогда уравнение перепишется в виде:

$$z = c + a_2 x$$

и система для определения параметров c, a_2 примет вид:

$$\begin{cases} nc + a_2 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} z_i \\ c \sum_{i=1}^{n} x_i + a_2 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i z_i \end{cases} \tag{10}$$

или, возвращаясь к табличным эмпирическим данным,

$$\begin{cases} nc + a_2 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} ln\, y_i \\ c \sum_{i=1}^{n} x_i + a_2 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i\, ln\, y_i \end{cases} \tag{11}$$

4. логарифмической функцией

$$y = a_1 + a_2 \ln x \tag{12}$$

Этот случай легко сводится к первому, если ввести обозначения $t = \ln x$, а в формулу (6) подставить вместо x_i массив $t_i = \ln x_i$.

Система (6) примет вид:

$$\begin{cases} a_1 n + a_2 \sum_{i=1}^{n} \ln x_i = \sum_{i=1}^{n} y_i, \\ a_1 \sum_{i=1}^{n} \ln x_i + a_2 \sum_{i=1}^{n} (\ln x_i)^2 = \sum_{i=1}^{n} y_i \ln x_i. \end{cases} \tag{13}$$

Решая системы линейных уравнений (6), (8), (11), определим параметры аппроксимирующих функций. Для решения систем можно использовать различные методы: метод Крамера, метод Гаусса, метод Зейделя, метод обратной матрицы. Матрицы систем (6), (8), (11) являются невырожденными квадратными матрицами, поэтому система уравнений для определения параметров аппроксимирующих функций имеет единственное решение.

Кроме указанных методов можно использовать методы решения систем с симметричной матрицей, т.к. матрицы всех рассмотренных систем являются симметричными.

После вычисления коэффициентов уравнений можно построить таблицы теоретических значений искомых аппроксимирующих функций по формулам:

$$y_i^{лин} = a_1+a_2x_i. \tag{14}$$

$$y_i^{квадр} = a_1+a_2x+a_3x_i^2 \tag{15}$$

$$y_i^{эксп} = a_1 \cdot e^{a_2 x_i} \tag{16}$$

$$y_i^{лог} = a_1+a_2 \ln x_i \tag{17}$$

и вычислить суммарную квадратичную ошибку по формуле (3).

Кроме указанных выше аппроксимирующих функций можно также использовать любые другие аналитические функции , определяемые конечным числом параметров. Например:

$$y = a_1 + a_2 x + a_3 x^2 + a_4 x^3$$

$$y = a_1 + \frac{a_2}{x}$$

$$y = a_1 e^{-a_2(x-a_3)^2}$$

Многочлен третьей степени является линейным относительно параметров – коэффициентов многочлена. Можно рассматривать в качестве аппроксимирующих функций многочлены любой степени. В других приведенных примерах функции легко линеаризовать, если использовать замену переменных и логарифмирование аппроксимирующего уравнения.

2.1.4. Оценка статистических параметров системы наблюдаемых величин

Параметры уравнений (5), (7), (9), (12) связаны некоторыми соотношениями со статистическими оценками эмпирических данных. Особенно это относится к линейному уравнению (5).

Напомним некоторые статистические оценки. Наблюдаемые значения величин x_i, y_i можно рассматривать как выборочные значения двух случайных величин X, Y. По выборочным данным можно найти выборочные средние и выборочные квадратичные отклонения X и Y, а также выборочный коэффициент корреляции, а именно:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \tag{18}$$

$$\sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \sigma_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{19}$$

Для вычисления σ_x, σ_y можно применить и более простые формулы, которые выводятся в курсе теории вероятностей с помощью простых алгебраических преобразований:

$$\sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i - \bar{x}^2}, \quad \sigma_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i - \bar{y}^2} \tag{20}$$

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y}}{\sigma_x \sigma_y} \tag{21}$$

здесь $\overline{x}, \overline{y}$ - выборочные средние величин X, Y; σ_x, σ_y - выборочные квадратичные отклонения величин X, Y; r - выборочный коэффициент корреляции.

Известно, что линейное уравнение (5), называемое в статистике уравнением линейной регрессии, проходит через точку $(\overline{x}, \overline{y})$, а коэффициент a_2, называемый в статистике коэффициентом регрессии, связан с коэффициентом корреляции r. Имеют место следующие соотношения:

$$\overline{y} = a_1 + a_2 \overline{x}, \tag{22}$$

$$a_2 = \frac{\sigma_y}{\sigma_x} r \tag{23}$$

Коэффициент корреляции характеризует меру линейной связи между величинами X,Y и может принимать значения в пределах от -1 до 1. Чем ближе к единице $| r |$, тем теснее линейная связь между X, Y. Если $| r | = 1$, то Y линейно зависит от X, т.е. выполняется соотношение:

$$y_i = a_1 + a_2 x_i ,$$

поэтому ошибка представления эмпирических данных равна 0.

Анализируя формулу (23), легко увидеть, что если линейная функция возрастающая, т.е. $a_2 > 0$, то и коэффициент корреляции положителен. Соответственно, если линейная функция убывает, то коэффициент корреляции отрицателен.

Если коэффициент корреляции равен 0 или близок к 0, то между величинами X,Y нет линейной корреляционной связи, или она является слабой, несущественной. Вообще говоря, этот вопрос требует отдельного рассмотрения – оценки значимости коэффициента корреляции и изучается в разделе статистики «проверка статистических гипотез».

2.1.5. Оценка точности аппроксимации

Квадратичная ошибка (погрешность) аппроксимации функции в соответствии с формулой (3) равна

$$S = \sum_{i=1}^{n} (y_i - y_i^T)^2 .$$

С целью оценки относительной погрешности при аппроксимации функции рассматривают величину суммарной погрешности по отношению к общему разбросу данных. Общий разброс данных складывается из отклонений теоретических значений от среднего и эмпирических значений от теоретических значений. Вводятся обозначения

$$S_{ост} = \sum_{i=1}^{n} (y_i - y_i^T)^2 , \quad (24)$$

$$S_{регр} = \sum_{i=1}^{n} (y_i^T - \bar{y})^2 , \quad (25)$$

$$S_{полн} = S_{ост} + S_{регр} \quad (26)$$

В случае линейной функции получим

$$S_{ост}^{лин} = \sum_{i=1}^{n} (y_i - a_1 - a_2 x_i)^2 \quad (27)$$

В случае квадратичной функции:

$$S_{ост}^{квадр} = \sum_{i=1}^{n} (y_i - a_1 - a_2 x_i - a_3 x_i^2)^2 \quad (28)$$

В случае экспоненциальной функции:

$$S_{ост}^{эксп} = \sum_{i=1}^{n} (y_i - a_1 e^{a_2 x_i})^2 \quad (29)$$

$$S_{ост}^{лог} = \sum_{i=1}^{n} (y_i - a_1 - a_2 \ln x_i)^2 \quad (30)$$

По аналогии легко написать формулы для вычисления регрессионных сумм и ошибки аппроксимации функцией любого вида.

Отметим, что для аппроксимирующей функции, линейной относительно параметров, верно:

$$S_{полн} = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

Относительная ошибка аппроксимации есть отношение

$$\frac{S_{ост}}{S_{полн}}.$$

Величина

$$\eta^2 = 1 - \frac{S_{ост}}{S_{полн}} \qquad (31)$$

называется коэффициентом детерминированности и характеризует меру точности аппроксимации табличных данных функцией любого вида. Если $\eta^2 = 1$, то ошибка аппроксимации равна 0 и теоретические значения функции совпадают с эмпирическими значениями.

2.2. ПРИМЕР ПОСТАНОВКИ И РЕШЕНИЯ ЗАДАЧИ АППРОКСИМАЦИИ

2.2.1 Постановка задачи

Рассмотрим таблицу значений двух наблюдаемых величин x, y:

Таблица 2.1

№п/п	1	2				n
x	x_1	x_2	…	x_i	…	x_n
y	y_1	y_2	…	y_i	…	y_n

Число измерений n может быть произвольным. Предположим, что y является функцией x.
Требуется:

1 построить уравнения линейной, квадратичной и экспоненциальной (или другой заданной) аппроксимирующей функции

2 определить основные статистические параметры данных случайных величин, определить коэффициент детерминированности полученных уравнений

3 построить график табличной функции и в той же системе координат график аппроксимирующей функции (для каждого вида аппроксимации отдельный рисунок)

Решение выполнить с помощью табличного процессора MS Excel и в среде MathCad как непосредственно по формулам, так и с использованием встроенных функций. Сравнить результаты вычислений в MathCad с результатами расчетов в MS Excel и добиться их совпадения в пределах используемой точности.

2.2.2 Расчетные формулы

Перечислим вначале расчетные формулы для решения задачи в MS Excel.

Для определения параметров уравнений регрессии требуется решить системы уравнений (6), (8), (11), (13), элементы которых представляют собой суммарные значения и даны непосредственно в этих формулах. Запишем системы (6), (8), (11), (13) в матричной форме:

SA=Y, где S – матрица системы, A – вектор неизвестных, Y – вектор правой части.

Для системы (6):

$$S = \begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix}, \quad A = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}, \quad Y = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{pmatrix}$$

Для системы (8):

$$S = \begin{pmatrix} n & \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i^3 \\ \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i^3 & \sum_{i=1}^{n} x_i^4 \end{pmatrix}, \quad A = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix}, \quad Y = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} x_{ii}^2 y_i \end{pmatrix}$$

Для системы (11):

$$S = \begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix}, \quad A = \begin{pmatrix} c \\ a_2 \end{pmatrix}, \quad Y = \begin{pmatrix} \sum_{i=1}^{n} \ln y_i \\ \sum_{i=1}^{n} x_i \ln y_i \end{pmatrix}$$

Для системы (13):

$$S = \begin{pmatrix} n & \sum_{i=1}^{n} \ln x_i \\ \sum_{i=1}^{n} \ln x_i & \sum_{i=1}^{n} (\ln x_i)^2 \end{pmatrix}, \quad A = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}, \quad Y = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} (\ln x_i) y_i \end{pmatrix}$$

Для решения систем (6), (8), (11), (13) применим метод обратной матрицы. Решение линейной системы SA=Y имеет вид

$$A = S^{-1} Y \tag{32}$$

- Для вычисления вектора A можно применить матричные функции.
- Вычисление значений аппроксимирующих функций выполним по формулам (14)-(17)

- Остаточные суммы для разных аппроксимаций вычислим по формулам (27)-(30)
- Оценку статистических параметров произведем по формулам (18), (20), (21)
- Ошибки аппроксимации и коэффициенты детерминированности вычислим по формулам (24)-(26) и (31).

2.2.3 Выполнение расчетов в MS Excel по расчетным формулам

1 Составим расчетную таблицу в MS Excel для вычисления сумм, входящих в системы (6), (8), (11) в следующей последовательности:

1) Введем в первую строку таблицы в ячейки A1:J1 заголовки столбцов: номер i, x_i , y_i , x_i^2 , x_i y_i , x_i^3 , x_i^4, x_i^2 y_i , ln y_i , x_i ln y_i
2) Заполним первый столбец номерами наблюдений от 1 до *n* . для этого в ячейку A2 введем 1 и заполним столбец с помощью арифметической прогрессии до значения *n*
3) Заполним следующий столбец значениями x_i
4) Заполним третий столбец значениями y_i
5) В ячейку D2 введем формулу: =B2*B2
6) В ячейку E2 введем формулу: =B2*C2
7) В ячейку F2 введем формулу: =B2*D2
8) В ячейку G2 введем формулу: =D2*D2
9) В ячейку H2 введем формулу: = D2*C2
10) В ячейку I2 введем формулу: = LN(C2)
11) В ячейку J2 введем формулу: = B2*I2
12) Формулы в ячейках D2:J2 скопируем в нижележащие ячейки для всех номеров *i*
13) В строке с номером *n+2=22* вычислим суммы столбцов с именами от B до J. Выполним Автосуммирование во втором столбце таблицы, при этом в ячейке B22 получим формулу: =СУММ(B2:B21), а затем скопируем полученную формулу в ячейки этой строки на всю ширину таблицы.

Результаты вычислений представлены в табл.2.2

Таблица 2.2

Образец вычислений

	A	B	C	D	E	F	G	H	I	J
1	i	x_i	y_i	x_i^2	$x_i y_i$	x_i^3	x_i^4	$x_i^2 y_i$	$\ln y_i$	$x_i \ln y_i$
2	1	0,86	1,89	0,74	1,63	0,64	0,55	1,40	0,64	0,55
3	2	1,54	6,76	2,37	10,41	3,65	5,62	16,03	1,91	2,94
4	3	2,78	9,79	7,73	27,22	21,48	59,73	75,66	2,28	6,34
5	4	3,99	15,98	15,92	63,76	63,52	253,45	254,40	2,77	11,06
6	5	5,34	54,67	28,52	291,94	152,27	813,14	1558,95	4,00	21,37
7	6	5,78	76,98	33,41	444,94	193,10	1116,12	2571,78	4,34	25,11
8	7	6,45	122,32	41,60	788,96	268,34	1730,77	5088,82	4,81	31,00
9	8	7,23	167,89	52,27	1213,84	377,93	2732,46	8776,10	5,12	37,04
10	9	7,65	188,65	58,52	1443,17	447,70	3424,88	11040,27	5,24	40,09
11	10	7,91	219,87	62,57	1739,17	494,91	3914,77	13756,85	5,39	42,66
12	11	8,34	256,78	69,56	2141,55	580,09	4837,98	17860,49	5,55	46,27
13	12	8,78	277,71	77,09	2438,29	676,84	5942,62	21408,22	5,63	49,40
14	13	8,88	300,45	78,85	2668,00	700,23	6218,02	23691,80	5,71	50,66
15	14	9,15	342,56	83,72	3134,42	766,06	7009,46	28679,98	5,84	53,40
16	15	9,25	378,45	85,56	3500,66	791,45	7320,94	32381,13	5,94	54,91
17	16	9,54	399,89	91,01	3814,95	868,25	8283,11	36394,63	5,99	57,16
18	17	9,75	425,98	95,06	4153,31	926,86	9036,88	40494,72	6,05	59,03
19	18	10,32	466,89	106,50	4818,30	1099,10	11342,76	49724,91	6,15	63,43
20	19	10,67	488,98	113,85	5217,42	1214,77	12961,57	55669,84	6,19	66,07
21	20	10,78	500,12	116,21	5391,29	1252,73	13504,39	58118,15	6,21	67,00
22	n	144,99	4702,61	1221,07	43303,24	10899,93	100509,22	407564,11	95,76	785,48
23		Σx_i	Σy_i	Σx_i^2	$\Sigma x_i y_i$	Σx_i^3	Σx_i^4	$\Sigma x_i^2 y_i$	$\Sigma \ln y_i$	$\Sigma x_i \ln y_i$

2. Составим и решим систему (6) для вычисления параметров линейной аппроксимации.

Сформируем матрицу системы (6) в ячейках C27:D28 по формулам:

1) В ячейку C27 введем формулу =A21 (n)
2) В ячейку D27 введем формулу =B22 (сумма x)
3) В ячейку C28 введем формулу =D27 (сумма x)
4) В ячейку D28 введем формулу =D22 (сумма x^2)

Сформируем правую часть системы (6) в ячейках E27:E28 по формулам:

1) В ячейку E27 введем формулу =C22 (сумма y)

2) В ячейку E28 введем формулу =E22 (сумма xy)

Решим систему по формуле (32) с использованием матричных функций. Для этого выделим область результата, например ячейки H27:H28 и введем формулу: {=МУМНОЖ(МОБР(C27:D28); E27:E28)}. В ячейках E27:E28 получены параметры линейной аппроксимирующей функции.

Расчеты показаны в табл. 2.3, скопированной из MS Excel:

Таблица 2.3

	C	D	E	F	G	H
25	Расчет параметров линейной аппроксимации					
26	Матрица системы		Правая часть		Решение	
27	20	144.99	4702.61		a_1	-157.781
28	144.99	1221.07	43303.2		a_2	54.19843

Уравнение линейной аппроксимации имеет вид:

$$y = -157.781 + 54.19843\,x \qquad (40)$$

3. Составим и решим систему (8) для вычисления параметров квадратичной аппроксимации. Расчеты показаны в табл. 2.4:

Таблица 2.4

	C	D	E	F	G	H
31	Расчет параметров квадратичной аппроксимации					
32	Матрица системы			Правая часть	Решение	
33	20	144.99	1221.07	4702.61	a_1	35.79013
34	144.99	1221.07	10899.9	43303.24	a_2	-34.5043
35	1221.1	10899.9	100509	407564.1	a_3	7.362073

Уравнение квадратичной аппроксимации имеет вид:

y = 35.79013 - 34.5043 x+ 7.362073 x^2 (41)

4. Составим и решим систему (11) для вычисления параметров экспоненциальной аппроксимации.

Таблица2.5

	C	D	E	F	G	H
37	Расчет параметров экспоненциальной аппроксимации					
38	Матрица системы		Правая часть		Решение	
39	20	144.99	95.7595		c	0.894814
40	144.99	1221.07	785.482		a_2	0.537025
41					a_1	2.446881

Найдем коэффициент a_1 по формуле $a_1 = e^c$, для этого в ячейку H41 введем формулу =EXP(H39)

Уравнение экспоненциальной аппроксимации имеет вид:

y = 2.446881e $^{0.537025}$ (42)

5. Вычислим теоретические значения функции по формулам (14)-(16):

введем в ячейку K2 формулу =$H$27+$H$28*B2, в ячейку L2 - формулу =$H$33+$H$34*B2+$H$35*B2^2, в ячейку M2 - формулу =$H$41*EXP($H$40*B2)

6. Вычислим остаточные суммы по формулам (27)-(29) или (24) введем в ячейку O2 формулу =(K2-C2)^2, в ячейку P2 - формулу =(L2-C2)^2, в ячейку Q2 - формулу =(M2-C3)^2.

Эти формулы скопируем в нижележащие ячейки для всех значений x.

7. Вычислим статистические оценки случайных величин по формулам (18), (20), (21):

в ячейку D50 введем формулу =B22/A21, в ячейку D51 - формулу =C22/A21 для вычисления средних;

перед тем, как вычислять среднее квадратичное отклонение по формуле (21), вычислим сумму квадратов значений Y:

в ячейку N2 введем формулу =C2^2 и скопируем формулу для всех значений i;
в ячейке N22 получим сумму квадратов по формуле =СУММ(N2:N21).
В ячейку D53 введем формулу =(D22/A21-D50^2)^(1/2);
в ячейку D54 введем формулу =(N22/A21-D51)^(1/2);
вычислим коэффициент корреляции по формуле (21):
в ячейку D55 введем формулу ==(E22/A21-D50*D51)/(D53*D54).

8. Для оценки регрессионных сумм вычислим квадраты разностей между теоретическими значениями наблюдаемой функции и средним значением по формуле (25). Для этого в ячейку R2 введем формулу =(K2-$D$51)^2, в ячейку S2 формулу =(L2-$D$51)^2, в ячейку T2 формулу =(M2-$D$51)^2.
9. Для вычисления остаточных и регрессионных сумм введем в ячейку O22 формулу =СУММ(O2:O21) и скопируем ее в ячейки P22:T22.

2.2.4 Построение графиков и вывод уравнений регрессии с помощью встроенных средств MS Excel

Представим графическую интерпретацию полученных уравнений, сравнив их с эмпирическими данными.

1. Построение прямой линии тренда

   Строим график исходной эмпирической функции. Для этого выделяем диапазон B2:C21 и строим точечную диаграмму:

   Выделяем на диаграмме ряд полученных точек и правой кнопкой мыши вызываем контекстное меню, выбираем команду - Добавить линию тренда. В диалоговом окне команды выбираем тип: «Линейная» и параметры: «Показывать уравнение на диаграмме» и «Поместить на диаграмме значение достоверности аппроксимации»;
2. Построение квадратичной линии тренда.

   Выполняется так же как построение прямой линии, но в диалоговом окне Линия тренда выбираем тип Полиномиальная и степень аппроксимирующего полинома равной 2. Параметры установить такие же, как в пункте 1

3. Построение экспоненциальной линии тренда.
   Выполняется аналогично пункту 1, но на вкладке Тип выбираем тип Экспоненциальная.

В процессе построения линий тренда вычисляются параметры аппроксимирующих функций и коэффициент детерминированности с помощью встроенных статистических функций Excel.

3. РЕШЕНИЕ ЗАДАЧИ В СИСТЕМЕ МАТЕМАТИЧЕСКИХ ВЫЧИСЛЕНИЙ MATHCAD

Методы построения различных математических моделей, отражающих существенные черты данного явления или процесса, проверка качества модели и границ ее применимости, применение модели для проведения конкретных расчетов и предсказания поведения системы составляют предмет математического моделирования в данной предметной области. При описании различных природных и технологических процессов, происходящих в некоторой системе, часто ставится задача построения математической модели поведения данной системы. Для решения такого рода математических задач существует множество пакетов, как узкоспециализированных, так и для решения широкого класса проблем. Большинство пакетов являются узкопрофильными: анализ статистических данных, решение определенного типа уравнений. Только немногие способны решать широкий спектр задач. К таким относятся Mathematica, Mathcad, Maple, MathLab, Derive, Statistica. Эти системы содержат большой набор готовых к употреблению алгоритмов и программ, позволяющих решать задачи математического анализа, линейной алгебры, геометрии, дифференциальных уравнений. Программа MathCad — это редактор математических текстов с широкими возможностями символьных вычислений и достаточно простым интерфейсом. Все выражения записываются в естественной математической форме. Вычисления происходят в автоматическом режиме, по мере пролистывания документа (принцип живых страниц). Текст, формулы и графики можно свободно сочетать, передвигая их как выделенные штриховой рамкой объекты, и помещать их в произвольной точке экрана; при изменении хотя бы в одном из объектов последовательно пересчитываются все остальные данные. Все содержимое рабочей области можно экспортировать в MS Word в формате RTF. Пакет располагает широким набором панелей интерфейса, содержащих практически все известные в высшей математике операторы и нотационные элементы формул.

Среда MathCad представляет достаточно удобное исполнение, поставленной в курсовом проекте, задачи. Так как MathCad – это что-то среднее между средой программирования, вычислительным процессором и графическим редактором, в ней легко и быстро создаются таблицы и матрицы, представляемые как массивы данных. Обрабатывать их можно с помощью стандартных математических функций, результат выводить – как в обычной численной форме, так и с помощью графиков. Наличие большого выбора всевозможных графиков является еще одним несомненным преимуществом MathCad перед другими математическими программами.

3.1. ВЫПОЛНЕНИЕ РАСЧЕТОВ В MATHCAD

3.1.1. Расчет статистических характеристик

1. Введем исходные данные x_i и y_i в виде матрицы с определенным количеством строк и столбцов с помощью инструмента «Матрица или вектор» на панели инструментов «Матрица» (рис.1).

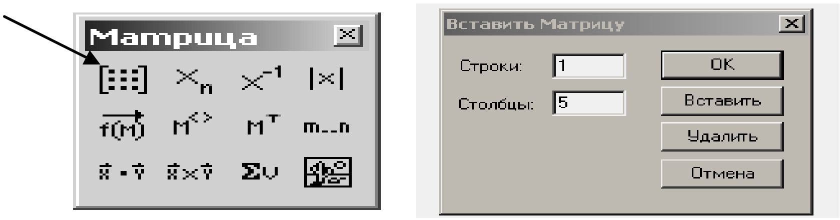

Рисунок 1. Панель инструментов «Матрица»

Значения векторов x_i и y_i можно выводить в виде столбцов либо в виде строк:

$$x := (16.67\ \ 14.71\ \ 12.99\ \ 11.76\ \ 10.20)$$

$$y := (1.79\ \ 1.67\ \ 1.58\ \ 1.50\ \ 1.42)$$

2.Вычислим произведения векторов, используя инструмент «Векторизация» на панели инструментов Матрица (рис.2).

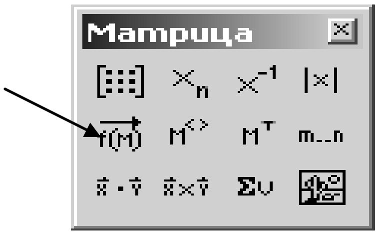

Рисунок 2. Инструмент «Векторизация»

Получим произведение x_i и y_i :

$$\left(\overrightarrow{(x \cdot y)}\right) = (29.839\ 24.566\ 20.524\ 17.64\ 14.484)$$

3. Найдем значения x^2 ,x^3, x^4 . Для этого нужно использовать инструмент возведения в степень на панели инструментов «Калькулятор» (рис.3).

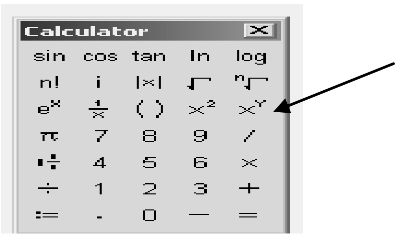

Рисунок3. Инструмент «Калькулятор»

Для того чтобы математическая операция была произведена для каждого элемента вектора x_i, необходим инструмент «Векторизация» на панели инструментов Матрица, Пример записи представлен ниже.

$$\overrightarrow{x^2} = (277.889\ 216.384\ 168.74\ 138.298\ 104.04)$$

$$\overrightarrow{x^3} = \left(4.632\times 10^3\ \ 3.183\times 10^3\ \ 2.192\times 10^3\ \ 1.626\times 10^3\ \ 1.061\times 10^3\right)$$

$$\overrightarrow{x^4} = \left(7.722\times 10^4\ \ 4.682\times 10^4\ \ 2.847\times 10^4\ \ 1.913\times 10^4\ \ 1.082\times 10^4\right)$$

4. Найдем натуральный логарифм вектора y_i пользуясь теми же инструментами:

$$\overrightarrow{\ln(y)} = (0.582\ 0.513\ 0.457\ 0.405\ 0.351)$$

5. Произведения $x_i^2\ y_i$ и x_i,$\ln(y_i)$ руководствуясь рассмотренными правилами:

$$\overrightarrow{(x\cdot\ln(y))} = (9.706\ 7.544\ 5.942\ 4.768\ 3.577)$$

$$\overrightarrow{\left(x^2\cdot y\right)} = (497.421\ 361.361\ 266.609\ 207.446\ 147.737)$$

6. Итогом этапа подготовки к составлению аппроксимирующих зависимостей будет являться вычисление сумм рассчитанных значений. Для этого используем инструмент «Суммирование вектора» на панели инструментов «Матрица» (рис.4).

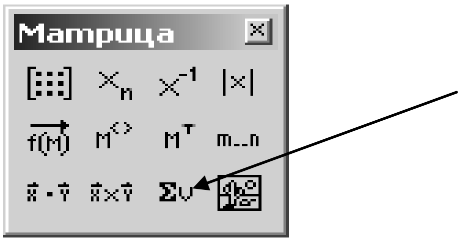

Рисунок 4. Инструмент «Суммирование вектора»

Получим следующие суммарные значения:

$\sum x = 66.33$

$\sum \overrightarrow{x^2} = 905.351$

$\sum \overrightarrow{x^3} = 1.269 \times 10^4$

$\sum \overrightarrow{x^4} = 1.825 \times 10^5$

$\sum \overrightarrow{(x \cdot y)} = 107.053$

$\sum y = 7.96$

$\sum \overrightarrow{\left(x^2 \cdot y\right)} = 1.481 \times 10^3$

$\sum \overrightarrow{\ln(y)} = 2.309$

$\sum \overrightarrow{(x \cdot \ln(y))} = 31.536$

3.1.2. Построение аппроксимирующих зависимостей

1.Составим и решим систему (6)

$$\begin{cases} a_1 n + a_2 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \quad , \\ a_1 \sum_{i=1}^{n} x_i + a_2 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i \; y_i \quad . \end{cases}$$

для вычисления параметров линейной аппроксимации. Для этого сформируем две матрицы и условно назовем их - А и В. Параметры матрицы А:

n- число членов ряда (в нашем случае n=5),

$\sum x = 66.33$

$$\sum \overrightarrow{x^2} = 905.351$$

Матрица А сформирована с помощью того же инструмента «Матрица или вектор» на панели инструментов «Матрица» и имеет вид:

$$A := \begin{pmatrix} 5 & 66.33 \\ 66.33 & 905.351 \end{pmatrix}$$

Параметры матрицы В:

$$\sum y = 7.96$$

$$\sum \overrightarrow{(x \cdot y)} = 107.053$$

Матрица В выглядит следующим образом:

$$B := \begin{pmatrix} 7.96 \\ 107.053 \end{pmatrix}$$

Решение предлагается выполнить методом Крамера и проверить правильность расчетов с помощью встроенной функции «lsolve».

2. Суть решения системы линейных алгебраических уравнений методом Крамера заключается в поочередной замене каждого столбца матрицы А столбцом матрицы В и дальнейшем нахождении определителя каждой матрицы. Однако применение метода Крамера возможно, если определитель, составленный из коэффициентов при переменных (его называют главным определителем системы), не равен нулю, в противном случае система имеет бесконечно много решений или несовместна. Поэтому первым шагом решения будет являться нахождение определителя матрицы А с помощью инструмента «Определитель» на панели инструментов «Матрица» (рис.5)

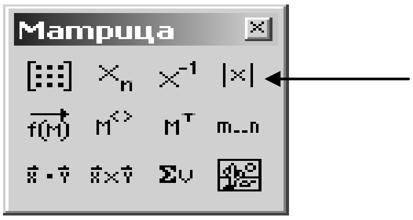

Рисунок 5. Инструмент «Определитель»

Определитель матрицы A равен:

$|A| = 127.086$

Далее воспользуемся инструментом «Столбец матрицы», предварительно создав матрицу A1 и присвоив ей значения матрицы A (рис.6).

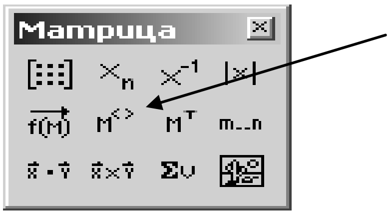

Рисунок 5. Инструмент «Столбец матрицы»

Соответствующая запись в Mathcad выглядит следующим образом:

$A1 := A$

$A1^{\langle 0 \rangle} := B$

$$A1 = \begin{pmatrix} 7.96 & 66.33 \\ 107.053 & 905.351 \end{pmatrix}$$

Определитель матрицы A1:

$|A1| = 105.768$

Аналогично составляем матрицу А2 и находим ее определитель:

$A2 := A$

$A2^{\langle 1 \rangle} := B$

$$A2 = \begin{pmatrix} 5 & 7.96 \\ 66.33 & 107.053 \end{pmatrix}$$

Определитель матрицы А2:

$|A2| = 7.278$

В итоге параметры линейной аппроксимации равны:

$a1 := \dfrac{|A1|}{|A|}$ $\qquad$ $a2 := \dfrac{|A2|}{|A|}$

$a1 = 0.832$ $\qquad$ $a2 = 0.057$

А уравнение линейной аппроксимации имеет вид:

$$y = 0.832 + 0.057x$$

Проверяем полученные результаты с помощью встроенной функции «lsolve».

$$\mathrm{lsolve}(A, B) = \begin{pmatrix} 0.832 \\ 0.057 \end{pmatrix}$$

Делаем вывод – параметры линейной аппроксимации определены правильно, соответственно уравнение составлено верно.

3. Используя системы (8) и (11) аналогичным способом можно определить параметры для квадратичной и экспоненциальной аппроксимации.

3.1.3. Оценка погрешности аппроксимации

1. Степень связи характеристик xi и yi предлагается оценить с помощью коэффициента корреляции:

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad (43)$$

Расчет всех входящих в формулу параметров можно вычислить используя те же инструменты, что и ранее: «Суммирование вектора» и Векторизация на панели инструментов «Матрица» и «Корень» и «Возведение в степень» плюс обычные арифметические операторы на панели инструментов «Калькулятор». Среднее значение находится с помощью записи «mean(x)», например:

$$xsr := mean(x)$$

$$ysr := mean(y)$$

Итоговая запись в MathCad:

$$p := \frac{\sum \overrightarrow{(xrazn yrazn)}}{\sqrt{\sum \overrightarrow{xrazn^2}} \cdot \sqrt{\sum \overrightarrow{yrazn^2}}}$$

$$p = 0.999$$

Проверить свои вычисления можно с помощью встроенной функции corr(x,y). Запись выглядит следующим образом:

$\mathrm{corr}(x, y) = 0.999$

2. Расчет коэффициентов детерминации необходимо выполнить для трех вариантов аппроксимации, чтобы посмотреть какая из моделей наиболее близка к исходной выборке. Опять же коэффициенты можно определить как вручную с помощью формул, так и с помощью встроенной функции. Общая формула для коэффициента детерминации:

$$r^2 = 1 - \frac{S_{ост}}{S_{полн}} \qquad (44)$$

где остаточные и регрессионные суммы вычисляются по формулам:

$$S_{ост} = \sum_{i=1}^{n} (y_i - y_i^m)^2, \qquad (45)$$

$$S_{полн} = S_{ост} + S_{регр} \qquad (46)$$

$$S_{регр} = \sum_{i=1}^{n} (y_i^m - \overline{y})^2 \qquad (47)$$

В пакете MathCad расчет всех составляющих, приведенных уравнений, выполняется с помощью стандартных инструментов, рассмотренных выше. При использовании встроенной функции, вначале записывается само уравнение регрессии, затем непосредственно коэффициент детерминации, например, для линейной зависимости это выглядит так:

$ylin(x) := a1 + a2 \cdot x$

$r2lin := corr(y, ylin(x))^2$

$r2lin = 0.9989$

Сравнение результатов вычислений, полученных в MathCad с результатами, полученными по расчетным формулам и на графиках в MS Excel , позволяет проверить правильность расчетов. Все результаты, кроме коэффициентов детерминированности экспоненциальной аппроксимации, должны совпасть.

3.1.4. Построение графиков в MathCad

Для построения графиков в среде MathCad воспользуемся инструментом «X-У Зависимость» на панели инструментов «Графики» (рис.6)

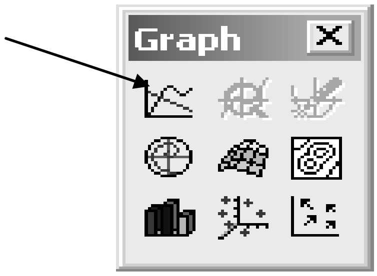

Рисунок 6. Панель инструментов «Графики»

В одной системе координат нам необходимо построить график по точкам *xi* и *yi* , а затем добавить график, исходя из определенного уравнения аппроксимации. Для этого сформируем массивы *xi* и *yi* в виде двух столбцов (панель инструментов «Матрица») и запишем уравнение аппроксимации. Затем введем график, по оси абсцисс запишем **x** по оси ординат запишем **y** и здесь же, через запятую, название аппроксимирующей функции (рис.7).

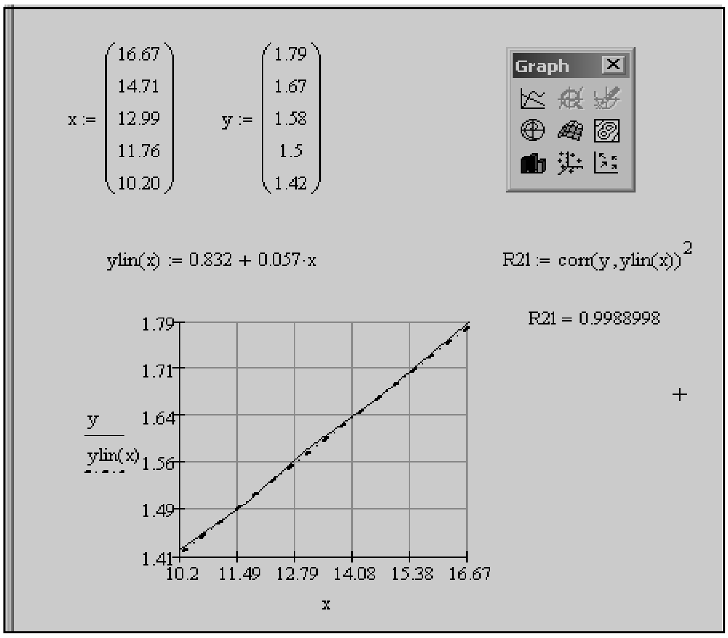

Рисунок 7. Страница MathCad с массивами, функцией, графиком и коэффициентом детерминации

Так как в данном примере высокий коэффициент детерминации, кривые практически совпадают, поэтому важно грамотно отформатировать график. Для этого воспользуемся панелью инструментов «Форматирование выбранного графика X-Y», где можно, например, переформатировать оси графика, добавить линии сетки, изменить границы осей, цвет графиков и т.д. (рис.8)

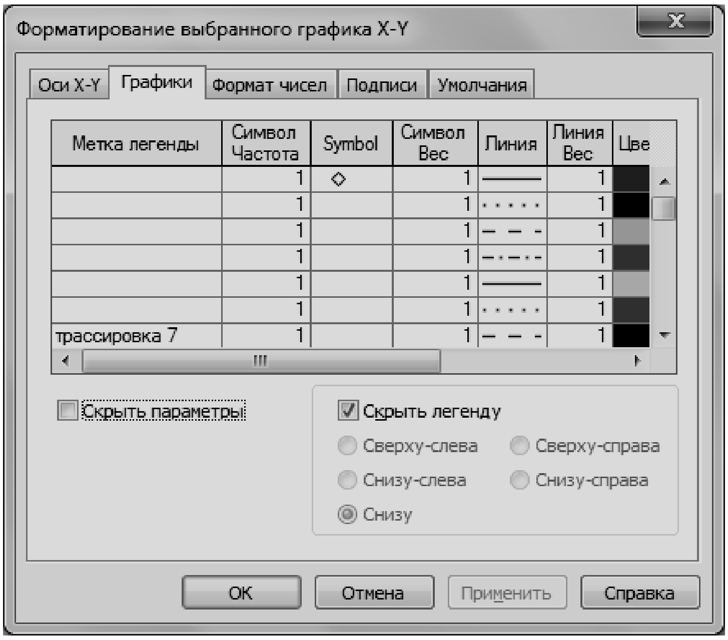

Рисунок 8. Панель «Форматирование графика»

Таким образом должны быть построены графики для всех видов аппроксимирующих зависимостей, проведен сравнительный анализ возможностей двух вычислительных систем MathCad и MS Excel. Определены преимущества и недостатки применительно к решению конкретной задачи – математического анализа параметров экосистемы.

4. ВАРИАНТЫ ЗАДАНИЙ

Варианты заданий даны в отдельном файле Варианты.xls в каталоге группы.

РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

1. *Пискунов Н.С.* Дифференциальное и интегральное исчисление Том1, М: Интеграл-Пресс, 2009-412 с
2. *Фурасов В. Д* Задачи гарантированной идентификации Издательство: Бином. Лаборатория знаний, 2005 – 356 с.
3. *Мешалкин В. П, Бутусов О. Б, Гнаук А. Г* Основы информатизации и математического моделирования экологических систем М: Инфа-М, 2001-358с.
4. *Шерри Виллард Кинкоф* Microsoft Excel 2003, М: НТ Пресс, 2004 -416 с.

ОГЛАВЛЕНИЕ