 Авторский материал Макаров Л.М. 2018

ТЕХНОЛОГИИ НАУЧНЫХ ИССЛЕДОВАНИЙ

практикум

ПЕРВАЯ ЧАСТЬ ПРАКТИЧЕСКОГО КУРСА

Постановка задачи.

Имеется тема диссертационной работы. Рассматриваем тему как набор словоформ.
По генеральной гипотезе написание набора словоформ должно сопровождаться понижением энтропии. В таком случае можно говорить о соблюдении основного закона термодинамики для открытой системы.

Термодинамическая система рассматривается как совокупность макроскопических объектов, которые обладают возможностью взаимодействовать. Результатом такого взаимодействия является обмен информационными пакетами, энергией или веществом. В нашем случае в качестве термодинамических объектов рассматриваются словоформы. Возможность создания из словоформ некоторого осмысленного выражения постулируется наличие языка общения и развитой технической системой телекоммуникаций. В самом общем случае система телекоммуникаций, как связанная сеть терминалов, обладает возможностью синтезировать, накапливать и анализировать информационные пакеты. Накопление информационных пакетов можно рассматривать как процесс создания информационного ресурса. Наличие в сети информационных ресурсов позволяет любому пользователю сети установить факт наличия той или иной темы, тематического проекта или суждения, представленного набором словоформ. Обращение к информационным ресурсам сети формируется в виде текстового запроса. Такой формат запроса позволяет установить количественную оценку наличия набора заданных словоформ в множестве информационных ресурсов. Другими словами установить наличие документов (текстов) содержащих словоформы запроса. Такие представления оказываются полезными при исследовании вопроса о уникальности сформированного пользователем набора словоформ, а также о лаконичности построения фразы - текстового транспаранта.

Общее представление о лаконичности построения высказываний, представленных в виде текста, можно найти в работах Аристотеля. В современной трактовке общего правила построения текста, как фразы, единственным и лаконичным образом устанавливающей связь между мысленным образом и реальным набором событий или явлений, пользуемся законом термодинамики. В рамках основного закона термодинамики, устанавливающего возможность извлечения знаний и практического опыта из разрозненных явлений окружающей среды, следует, что набор словоформ (высказывание) должен характеризоваться последовательной серией коэффициентов, характеризующих понижение энтропии.

Для установления реальной ситуации распределения энтропии в наборе словоформ (высказывании) предлагается использовать информационную среду - Интернет. Осуществляя запрос в сети на созданную фразу (набор словоформ), последовательно переходя от первой к последующей словоформе формируется набор коэффициентов. В том случае если набор коэффициентов характеризует нисходящий ряд констатируем лаконичность создания исходной фразы (высказывания). В случае восходящего ряда коэффициентов имеем семантически не ясную конструкцию выражения (фразы). Именно такие простые правила позволяют проанализировать лаконичность высказываний и исключить другие смысловые интерпретации высказывания.

Основные представления о данной процедуре формируются на постулатах:

- в замкнутой системе, где отсутствуют обменные процессы с внешней средой, энтропия постоянна;
- в открытой систем, где имеются потоки обмена с внешней средой, энтропия стремится к возрастанию

Следствием этих постулатов является наличие искусственного механизма понижения энтропии, который реализуется творческой деятельностью, например, при написании осмысленного текста. В качества текста рассматриваем название будущей работы.

Написание текста можно рассматривать как процесс, направленный против факторов внешней среды, изначально позиционирующих бессмысленный набор из словоформ.

Базисным понятием всей теории информации является понятие энтропии. Энтропия – мера неопределенности некоторой ситуации. Можно также назвать ее мерой рассеяния и в этом смысле она подобна дисперсии. Но если дисперсия является адекватной мерой рассеяния лишь для специальных распределений вероятностей случайных величин (а именно – для двухмоментных распределений, в частности, для гауссова распределения), то энтропия не зависит от типа распределения. Это важный вывод, поскольку позволяет работать с источниками бинарной природы.

С другой стороны, энтропия рассматривается нами чтобы осуществить возможность построить оценку лаконичности набора словоформ.

Рассмотрим этапы реализации процедуры формирования оценки лаконичности фразы, построенной из последовательного набора словоформ.

Если мы вводим меру неопределенности f , то естественно потребовать, чтобы она была такова, чтобы во-первых, неопределенность росла с ростом числа возможных исходов, а во-вторых, неопределенность составного опыта была равна просто сумме неопределенности отдельных опытов, иначе говоря, мера неопределенности была аддитивной: f(nm)=f(n)+f(m). Именно такая удобная мера неопределенности была введена К. Шенноном:

$$H(x) = -\sum_{i=1}^{N} p(x_i) \lg_2 (x_i)$$

где x – дискретная случайная величина с диапазоном изменчивости N, P(Xi) – вероятность i – го уровня X.

В дальнейшем мы будем рассматривать x как некоторую физическую величину, меняющуюся во времени или пространстве.

При преобразованиях энергии из одного вида в другой ее общее количество, в соответствии с первым началом, сохраняется постоянным, но изменяется ее качество, характеризуемое соотношением между свободной и связанной энергиями. Второе начало термодинамики утверждает, что в закрытых системах процессы преобразования энергии идут в сторону роста связанной энергии, а следовательно, и энтропии. При этом свободная составляющая энергии уменьшается. Можно сказать, что свободная энергия находится в конфликте с энтропией:

чем меньше одна, тем больше другая. В связи с этим свободную энергию часто называют отрицательной энтропией, или негэнтропией, хотя это не совсем корректно, поскольку размерности энтропии и энергии различны.

Напомним основные свойства энтропии.
1. В закрытых системах энтропия всегда неотвратимо растет. Оно выражает суть второго начала термодинамики.
2. Рост энтропии означает ликвидацию различий. Различие – это то, что обеспечивает целенаправленное существование любой сущности. Цель этого существования – уменьшение различий. В термодинамическом понимании системный кризис любой системы означает значительный рост энтропии этой организации, ее деградацию.
3. Чем больше свободы, тем быстрее растет энтропия. Скорость роста энтропии – скорость появления разнообразных способов организации сущностей, а свобода способствует этому появлению, ускоряет рост числа способов организации.

Поэтому чем больше свободы, тем быстрее низкоэнтропийные сущности превращаются в высокоэнтропийные. Энтропия неотвратимо растет только в закрытых системах, не взаимодействующих с другими системами и внешней средой.

Но в открытых системах энтропия может вести себя по-разному: расти, быть постоянной и даже уменьшаться. Причина различного поведения энтропии объясняется тем, что:

1. в открытых системах существуют собственная энтропия, которая, как и в закрытых системах, всегда растет;
2. энтропия, поступающая в открытую систему из внешней среды (импортируемая энтропия);
3. энтропия, удаляемая из открытой системы во внешнюю среду (экспортируемая энтропия).

Комментарии, где рассматриваемые понятия применяются к текстовым массивам, наборам словоформ.
Случай 1. Рост энтропии является естественным процессом «старения» словоформ и ухода их из речи
Случай 2. Постоянное обращение, в данном случае человека, к внешним информационным ресурсам, с целью выражения своих мыслей или индикации будущих действий, технологий. В данном случае поступающая из информационного пространства энтропия необходима для развития самой системы, и конечно человека.
Случай 3. Экспортируемая во внешнюю среду энтропия в виде набора словоформ естественным образом понижает энтропийную характеристику среды, что создает благоприятные условия для жизнедеятельности человека.

Действительно, говоря образно, чем выше информационные возможности среды, в смысле получения различной информации человеком, тем комфортнее условия жизнедеятельности. В таком понимании, в идеале, на любую текущую ситуацию в среде обитания человека всегда можно найти в информационном пространстве прогноз развития событий.

Информационная трактовка второго начала утверждает: В замкнутой системе любое однонаправленное коллективное движение составляющих эту систему элементов не может продолжаться сколь угодно долго и должно перейти в хаотическое движение Однако посколь-

ку сама информация не зависит от времени, то второе начало в общей теории информации связано с материальным свойством нематериальной информации, с носителем информации, с тем свойством, которое мы назвали памятью.

Поэтому более точная информационная трактовка второго начала записывается так: «В природе нет памяти с бесконечным временем существования». Одно из важнейших следствий второго начала термодинамики говорит о том, что не может существовать сколь угодно длительного прямолинейного движения. В качестве примера можно рассматривать линейный процесс приобретения знаний. В этом контексте ясно, что первоначальные установки получения знаний, в той или иной области, со временем будут модифицироваться.

Второе начало термодинамики - всеобщий закон природы, который распространяется на любую физическую систему, в том числе и на стационарные формы существования материи. Ведь стационарная форма существования материи - результат информационного взаимодействия. Направленное движение материальной точки, единичного объекта - это простейший вид существования информации, но он является основой возникновения любой другой формы материального мира. Еще в XVIII веке П. Мопертюи сформулировал принцип, который называется сегодня принципом наименьшего действия Мопертюи-Лагранжа. П. Мопертюи сформулировал его так: «Природа, производя действия, всегда пользуется наиболее простыми средствами», «количество действия всегда является наименьшим». В термодинамике сформулирован другой принцип – принцип наименьшего рассеяния энергии, Он обоснован в теореме Онсагера - одной из основных теорем термодинамики неравновесных процессов, установленной американским физиком в Л. Онсагером. На основании теоремы Оснагера Пригожиным в 1947 доказана еще одна теорема термодинамики неравновесных процессов, согласно которой при установленных внешних условиях, препятствующих достижению системой равновесного состояния, стационарному (неизменному по времени) состоянию системы соответствует минимум производства

Если таких препятствий нет, то производство энтропии достигает своего абсолютного минимума - нуля. К этому ряду принципов можно отнести также принцип наименьшего принуждения Гаусса, принцип наименьшей кривизны Герца и рад других принципов физики.

Следуя этим представлениям рассмотрим пример. Используем «ЯНДЕКС браузер»

Положим имеется фраза, отражающая название научной работы:

ФРАЗА	**Исследование**	**канала**	**связи**	**компьютеров**		
					Вариации слова	
Количество источников	75*10^6	6,04*10^6	5,0*10^6	2,0*10^6		
					биокомпьютеров	29,80*10^6
					биосенсоров	0,15*10^6
					серверов	0,15*10^6
					терминалов	1,00*10^6

Основная фраза: **Исследование канала связи компьютеров,** позиционируется набором информационных источников, обнаружение которых характеризуется уровнями вероятности:

$$R = 75 + 6.04 + 5.0 + 2.0 = 88.04$$

$$p_1 = \frac{75}{88.04} = 0.8519 \qquad p_2 = \frac{6.04}{88.04} = 0.0686$$

$$p_3 = \frac{5}{88.04} = 0.0568 \qquad p_4 = \frac{2.0}{88.04} = 0.0227$$

Имеем:

Номер словоформы	1.	2.	3.	4.
Вероятность обнаружения	0,8519	0,0686	0,0568	0,0227

Аналогичные вычисления можно провести для альтернативных вариантов построения фразы, используя словоформы из правого столбца таблицы.

Очевидно, требование понижения вероятности встречаемости слова в многочисленных информационных ресурсах (Интернет ресурс), создает условия для написания лаконичной фразы, отражающей семантическую направленность научной работы. Понижение уровня вероятности обнаружения словоформы свидетельствует о уникальности, отличии этой части высказывания от других. Постепенное понижение уровня вероятности от перового слова до последнего свидетельствует о совершаемой работе, направленной на понижение энтропии, характеризующей семантику сообщения. Иллюстрация этого факта представлена на рисунке ниже.

В заключении рассмотрим доверительный интервал. Для этого проведем вычисление среднего значения мости слов в выше представленном наборе словоформ.

Имеем:

$$R_{среднее} = \frac{88,04}{4} = 29,35$$

Дисперсия $D = \frac{\sum (x - x_{ср})^2}{(n-1)} = 1250.91$

Полагаем, что просмотрено более 100 источников (n>100).

Воспользуемся сервисом, для вычисления доверительного интервала, по адресу

http://math.semestr.ru/group/interval.php

Доверительный интервал для генерального среднего.

$$\left(\overline{x} - t_{kp}\frac{s}{\sqrt{n}}; \overline{x} + t_{kp}\frac{s}{\sqrt{n}}\right)$$

Поскольку n>30, то определяем значение t_{kp} по таблицам функции Лапласа.

В этом случае $2\Phi(t_{kp}) = \gamma$
$\Phi(t_{kp}) = \gamma/2 = 0.95/2 = 0.475$
По таблице функции Лапласа найдем, при каком t_{kp} значение $\Phi(t_{kp}) = 0.475$
$t_{kp}(\gamma) = (0.475) = 1.96$

$$\epsilon = t_{kp}\frac{s}{\sqrt{n}} = 1.96\frac{35.37}{\sqrt{100}} = 6.932$$

(29.35 - 6.932;29.35 + 6.932) = (22.42; 36.28)

С вероятностью 0.95 можно утверждать, что среднее значение при выборке большего объема не выйдет за пределы найденного интервала.

Формально можно констатировать, что «центр фразы» характеризуется интервалом протяженностью от 22,42 до 36,28 единиц. Или иначе: избранная фраза характеризуется набором словоформ, которые можно обнаружить как минимум на 22,42 *10^6 источниках или по максимуму на 36,28 *10 ^6 источниках.

Другими словами, просмотрев минимальное количество, указанное выше, источников в сети, можно с вероятностью 0,95, обнаружить близкие по семантике фразы.

ЗАДАНИЕ 1.

Используя теоретический материал по оценке энтропии в наборе словоформ, продемонстрировать оптимальный выбор названия научной работы.

Расчеты провести в среде EXCEL а комментарии представить в виде текстового материала с иллюстрациями

ЗАДАНИЕ 2.

Фрактал

Феномен фракталов и самоподобия хорошо известен начиная с последней четверти XX века. Фрактал – это множество дробной размерности. Природа устроена так, что многие объекты и события в ней – фрактальны, то есть их можно описать как фрактал с определенными параметрами. Фракталы обладают свойством самоподобия, то есть каждая их часть в каком-то смысле подобна более мелким частям, из которых эта большая часть состоит.

Рассмотрим фракталы, как *самоподобные случайные процессы*.

Случайный процесс (вероятностный процесс, случайная функция, стохастический процесс) в теории вероятностей — семейство случайных величин, индексированных некоторым параметром, чаще всего играющим роль времени или координаты.

Функция X(t) называется случайной, если ее значение при любом аргументе t является случайной величиной. Случайные функции времени называют случайными процессами. Реализацией случайной функции X(t) (выборочной функцией) называется конкретный вид, который она принимает в результате опыта. Реализация случайного процесса может рассматриваться как элемент множества возможных физических реализаций случайного процесса. Совокупность реализаций случайного процесса называется ансамблем реализаций, например, набором амплитудных значений случайного сигнала. Совокупность значений реализаций в фиксированный момент времени (выборка случайных значений) называется сечением случайного процесса.

По современным научным представлениям любая задача прогнозирования предполагает построение математической модели. Зная модель и имея первоначальный набор наблюдений, можно вычислить параметры этой модели и построить прогноз. Одним из наиболее ярких проявлений самоподобных процессов являются процессы телетрафика, то есть загрузки коммуникационных сетей. В данном направлении создается много моделей позволяющих найти оптимальные режимы эксплуатации оборудования.

Для полной характеристики случайного процесса недостаточно наличие оценки математического ожидания и дисперсии. Для зависимых наблюдений вводится понятие «связанного ряда»: вероятность возникновения на определенном месте тех или иных конкретных значений зависит от того, какие значения случайная величина уже получила раньше или будет получать позже. Иными словами, существует поле рассеяния пар значений x(t), x(t+k) временного ряда, где k - постоянный интервал или задержка, характеризующее взаимозависимость последующих реализаций процесса от предыдущих. Теснота этой взаимосвязи оценивается коэффициентом автокорреляции

$r(k) = E[(x(t) - m)(x(t + k) - m)] / D$,

где m и D - математическое ожидание и дисперсия случайного процесса. Для расчета автокорреляции реальных процессов необходима информация о совместном распределении вероятностей уровней ряда $p(x(t_1))$, $p(x(t_2))$.

Рассмотрим пример

Известны амплитудные значения двух случайных процессов:

$$A_1(t) = D_1 + S(t_i) + S(t_{i+1})$$

$$A_2(t) = D_2 + S(t_i) + S(t_{i+1})$$

Где D – постоянный коэффициент. D_1= 10; D_2= 12 ; S(t) – амплитудные значения случайного процесса (0, 1);
t – интервал наблюдения событий [1, 10]

В среде EXCEL создадим вычислительную процедуру. Используя введенные условия можно получить две цепи значений случайных процессов. Например:

Таблица 2.1

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
A1	11,42386	11,31552	10,76609	10,9063	11,19209	10,73874	10,48874	11,23388	11,34262	11,25246
A2	11,48963	12,61871	11,93073	11,92906	11,78515	12,66821	11,58179	11,67307	12,21819	11,87196

Рассмотрим суммарный процесс

Таблица 2.2

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
	22,91349	23,93423	22,69681	22,83536	22,97724	23,40695	22,07052	22,90695	23,56082	23,12443
G - Дисперсия	0,261									
μ - Средне значение	23,00									

Воспользуемся сервисом для вычисления автокорреляционной функции https://math.semestr.ru/corel/autocorrelation.php

Для выявления структуры ряда (т. е. состава компонент) создаем автокорреляционную функцию.
Автокорреляция уровней ряда – корреляционная между последовательными уровнями одного и того же ряда динамики (сдвинутыми на определенный промежуток времени *L* – лаг). То есть связь между рядом: X_1, X_2, ... X_{n-L} и ряд ом X_{1+L}, X_{2+L}, ... X_n, где *L* – положительное целое число. Автокорреляция может быть измерена коэффициентом автокорреляции.
Лаг (сдвиг во времени) определяет порядок коэффициента автокорреляции. Если L = 1, то имеем коэффициент автокорреляции 1-го порядка $r_{t,t-1}$. Если L = 2, то коэффициент автокорреляции 2-го порядка $r_{t,t-2}$ и т.д.
Следует учитывать, что с увеличением лага на единицу число пар значений, по которым рассчитывается коэффициент автокорреляции, уменьшается на 1. Поэтому обычно рекомендуют максимальный порядок коэффициента автокорреляции, равный n/4.
Рассчитав несколько коэффициентов автокорреляции, можно определить лаг (I), при котором автокорреляция ($r_{t,t-L}$) наиболее высокая, выявив тем самым **структуру временного ряда.**

Чтобы найти коэффициент корреляции 1-го порядка, нужно найти корреляцию между рядами (расчет производится не по 10, а по 9 парам наблюдений):

Два важных свойства коэффициента автокорреляции:

1) Ряд строится по аналогии с линейным коэффициентом корреляции и таким образом характеризует тесноту только линейной связи текущего и предыдущего уровней ряда. Поэтому по коэффициенту автокорреляции можно судить о наличии линейной (или близкой к линейной) тенденции. Для некоторых временных рядов, имеющих сильную нелинейную тенденцию (например, параболу второго порядка или экспоненту), коэффициент автокорреляции уровней исходного ряда может приближаться к нулю.

2) По знаку коэффициента автокорреляции нельзя делать вывод о возрастающей или убывающей тенденции в уровнях ряда. Большинство временных рядов экспериментальных данных содержит положительную автокорреляцию уровней, однако при этом могут иметь убывающую тенденцию.

Работаем с материалом примера. Сдвигаем исходный ряд на 1 уровень. Получаем следующую таблицу:

Таблица 2.3

	y_t	y_{t-1}
1.	22.91	23.94
2.	23.94	22.69
3.	22.69	22.83
4.	22.83	22.97
5.	22.97	23.4
6.	23.4	22.07
7.	22.07	22.9
8.	22.9	23.56
9.	23.56	23.12

Расчет коэффициента автокорреляции 1-го порядка. Параметры уравнения авторегрессии.
Выборочные средние.

$$\overline{x} = \frac{\sum x_i}{n} = \frac{207.27}{9} = 23.03$$
$$\overline{y} = \frac{\sum y_i}{n} = \frac{207.48}{9} = 23.05$$
$$\overline{xy} = \frac{\sum x_i y_i}{n} = \frac{4777.65}{9} = 530.85$$

Выборочные дисперсии:

$$S^2(x) = \frac{\sum x_i^2}{n} - \overline{x}^2 = \frac{4775.79}{9} - 23.03^2 = 0.26$$
$$S^2(y) = \frac{\sum y_i^2}{n} - \overline{y}^2 = \frac{4785.45}{9} - 23.05^2 = 0.26$$

Среднеквадратическое отклонение.

$$S(x) = \sqrt{S^2(x)} = \sqrt{0.26} = 0.51$$
$$S(y) = \sqrt{S^2(y)} = \sqrt{0.26} = 0.51$$

Коэффициент автокорреляции.

Линейный коэффициент автокорреляции $r_{t,t-1}$:

$$r_{t,t-1} = \frac{\overline{x_t \cdot x_{t-1}} - \overline{x_t} \cdot \overline{x_{t-1}}}{S(x_t) \cdot S(x_{t-1})} = \frac{530.85 - 23.03 \cdot 23.05}{0.51 \cdot 0.51} = -0.26$$

Линейный коэффициент корреляции принимает значения от –1 до +1.
Связи между признаками могут быть слабыми и сильными (тесными). Их критерии оцениваются по шкале Чеддока:
$0.1 < r_{t,t-1} < 0.3$: слабая;
$0.3 < r_{t,t-1} < 0.5$: умеренная;
$0.5 < r_{t,t-1} < 0.7$: заметная

ВЫВОД. Результат вычисления r = -0.26 . Это значение коэффициента указывает на слабую связь исследуемых процессов.

Перейдем к построению фрактального образа случайного процесса.

Создаем фрактальный образ аддитивной композиции двух сигналов. Рассмотрим уравнение эпициклоиды.

ЭПИЦИКЛОИДА (от греч. epi - на, над, при, после и kukloz - окружность, круг) - плоская кривая, траектория точки производящей окружности радиуса r, катящейся без скольжения по другой неподвижной окружности радиуса R, Параметрические уравнения:

$$X(t) = (R+r)COS(t) - rCOS(\frac{R+r}{r}t)$$ где t – (0, 2π)

$$Y(t) = (R+r)SIN(t) - rRSIN(\frac{R+r}{r}t)$$

Проведем построение фрактального образа: R=23; r= 2,3 . Отобразим полученную фигуру их 10 звеньев, связанных между собой без разрыва. Непрерывный цельный фрактальный образ создан при условии кратности R и r. В таком образе присутствуют самоподобные звенья, зрительно отражающие подобие элементов структуры временного ряда.

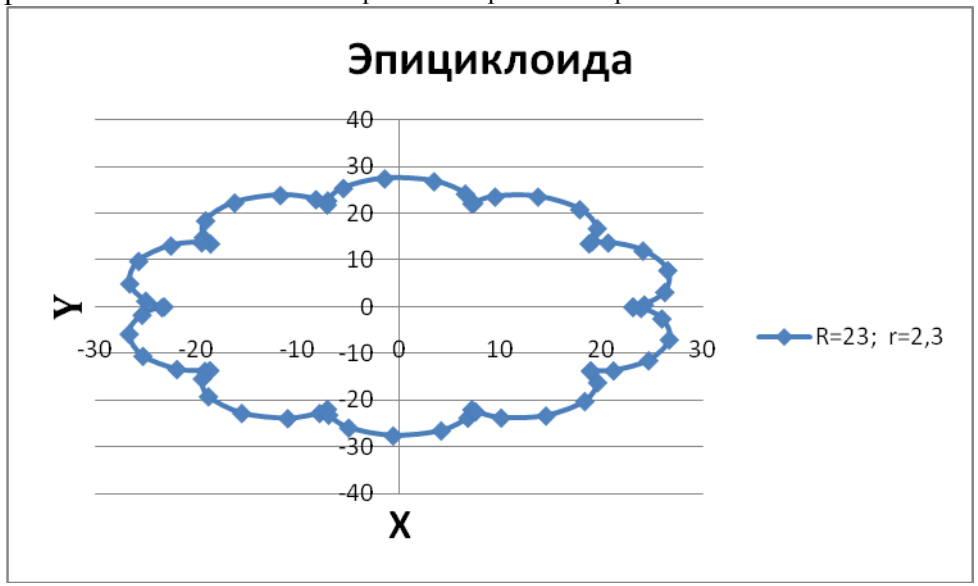


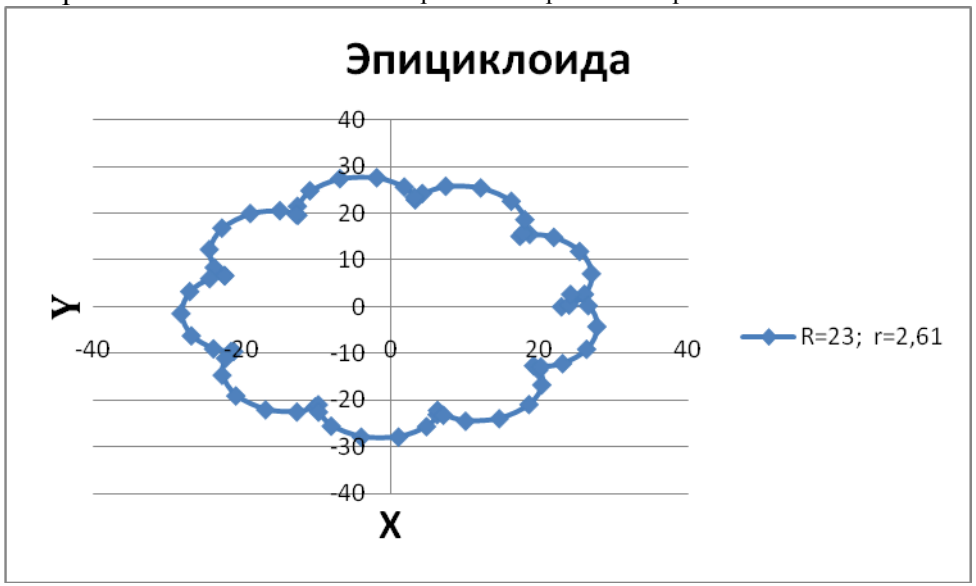
Рисунок 2.1 Эпициклоида – фрактальный образ

Модифицируем исходное уравнение для эпициклоиды:

$$X(t) = (G + 10\mu)COS(t) - 10\mu COS(\frac{G + 10\mu}{10\mu} t)$$

$$Y(t) = (G + 10\mu)SIN(t) - 10\mu RSIN(\frac{G + 10\mu}{10\mu} t)$$

Используем данные по оценке параметров двух сигналов: G = 23; μ = 2.61 . Отобразим фрактальный образ.



Используем данные по расчету коэффициента автокорреляции: G = 23; μ = 2,6. Полагаем, что μ = r. Отобразим фрактальный образ.

Рисунок 2.2 Модифицированный фрактальный образ

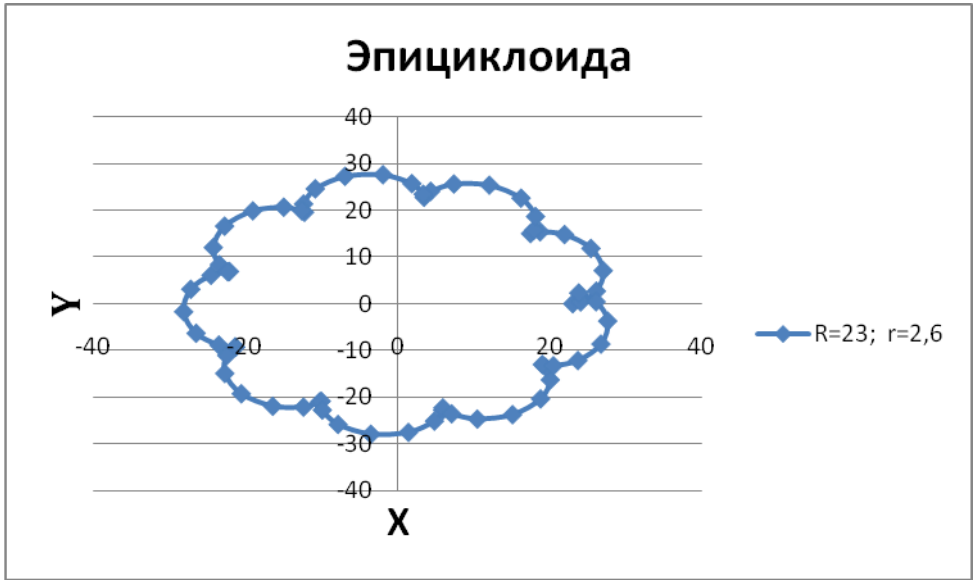


Рисунок 2.3 Фрактальный образ, построенный на наборе экспериментальных данных

Вывод

Модельный фрактальный образ, созданный по выражению для типичной эпициклоиды, подобен фрактальному образу, созданному на основе экспериментального материала. Наличие возможности масштабировать образы подчеркивает особенности фрактальной геометрии, стремящейся воспроизвести хорошо запоминающиеся зрительные образы.

Задание

Провести построение фрактального образа по экспериментальным данным с использованием уравнения эпициклоиды.

Исходные данные

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
	64,59349	65,61423	64,37681	64,51536	64,65724	65,08695	63,75052	64,58695	65,24082	64,80443
G - Дисперсия	0,261088									
μ - Средне значение	25,88907									

Указание

Результаты представить в форме отчета в редакторе Word.

Сделать выводы.