МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное
образовательное учреждение высшего образования
«Юго-Западный государственный университет»
(ЮЗГУ)

Кафедра биомедицинской инженерии

Лабораторный практикум по дисциплине
«Биоинформатика»

для студентов направления подготовки
12.03.04 «Биотехнические системы и технологии»

Курск 2023

УДК 50+51-7+57

Составители: М.В. Артеменко, Н.М. Калугина

Рецензент
Доктор биологических наук, профессор Привалова И.Л.

Лабораторный практикум по дисциплине «Биоинформатика» / Юго-Зап. гос. ун-т; сост. М.В. Артеменко, 2023. −91 с.: Приложений - 6

Лабораторный практикум содержит краткие теоретические сведения, порядок выполнения и содержание отчета по лабораторным работам по дисциплине «Биоинформатика» и соответствуют требованиям Федерального государственных образовательных стандартов высшего образования направлений подготовки 12.03.04 «Биотехнические системы и технологии». Рассматриваются разделы: представления информации о поведении биологических объектов инструментальными средствами офисных программ, анализ и разработка алгоритмов сравнения генетических последовательностей, синтеза решающих диагностических (классификационных) правил, расчет критериев качества диагностического процесса, анализ экологической ситуации и прогноз заболеваемости в регионе, выделение ритмических составляющих в биомедицинских сигналах, применение корреляционного анализа в биомедицинских исследованиях. Экспериментальная часть работ основывается на применении компьютерных технологий для обработки результатов исследований.

Предназначено для студентов для студентов направления подготовки 12.03.04 «Биотехнические системы и технологии»

Текст печатается в авторской редакции

Подписано в печать Формат 60х84х 1/16.
Усл.печ.л. . Уч.-изд.л. . Тираж 100 экз. Заказ. Бесплатно.
Юго-Западный государственный университет.
305040, г.Курск, ул. 50 лет Октября,

СОДЕРЖАНИЕ

ЛАБОРАТОРНАЯ РАБОТА №1. СРАВНИТЕЛЬНЫЙ АНАЛИЗ СТРУКТУР ДНК

Цель работы: овладение навыками разработки и анализа алгоритмов генетических последовательностей с помощью современных информационных средств и технологий.

Краткие теоретические сведения.

Нуклеотидная последовательность, генетическая последовательность - это порядок следования нуклеотидных остатков в нуклеиновых кислотах. Определяется при помощи секвенирования. Для записи нуклеотидных последовательностей ДНК по рекомендации IUPAC используются символы латинского алфавита: A =аденин; C=цитозин; G =гуанин; T =тимин. Для записи последовательностей РНК обычно достаточно символов A, C, G, U (уридин).

Все последовательности записываются без пробелов. Сравнительный анализ нуклеотидных последовательностей позволяет судить о степени родства сравниваемых организмов. Это обстоятельство широко применяется на практике (в частности, для установления отцовства).

Изучая семьи с известной генеалогией, генетики оценивают скорость накопления различий в ДНК. В частности, большую помощь оказало исследование ДНК населения Исландии - уникальной страны, где каждый житель знает всех своих предков вплоть то первых колонистов, прибывших в Исландию из Норвегии в IX веке (причем из останков нескольких первопоселенцев тоже удалось извлечь ДНК для анализа). Теми же методами можно реконструировать историю целых народов или, к примеру, находить среди современных азиатов потомков Чингисхана.

Результаты генетического анализа при этом хорошо согласуются с сохранившимися историческими сведениями.

В ходе многочисленных исследований такого рода, где можно было непосредственно сравнить генетические данные с историческими, генетики раз за разом убеждались в достоверности

оценок родства на основе сравнения ДНК, а используемые методы развивались и совершенствовались.

Генетическое родство человека и шимпанзе доказывается даже не столько сходством последовательностей, сколько характером различий между ними. Легко заметить, что характер этих различий полностью соответствует предсказаниям эволюционной теории.

Выравнивание аминокислотных или нуклеотидных последовательностей – это процесс сопоставления сравниваемых последовательностей для такого их взаиморасположения, при котором наблюдается максимальное количество совпадений аминокислотных остатков или нуклеотидов. Различают 2 вида выравнивания: парное (выравнивание двух последовательностей ДНК, РНК или белков) и множественное (выравнивание трех и более последовательностей).

Наиболее популярной серией программ для множественного выравнивания последовательностей является Clustal. Первая программа серии Clustal была создана Д.Хиггинсом в 1988 году. Затем она была усовершенствована Д. Фенгом, Р. Дулиттл и В. Тейлором путем добавления прогрессивного выравнивания, то есть созданием множественного выравнивания в результате серий попарных выравниваний, следуя ветвлению направляющего дерева, построенного методом UPGMA.

В 1992 году появилась второе поколение программ Clustal. Программа, названная Clustal V, отличалась способностью проводить сопоставления существующих выравниваний и построением направляющего дерева методом NJ. Третье поколение программ, появившееся в 1994 году и названное Clustal W, стало значительно проще в работе благодаря усовершенствованному алгоритму. Кроме этого появилась возможность выбирать матрицы сравнения аминокислот и нуклеотидов, а также устанавливать штрафы за внесение пробелов. Следует отметить, что высокая совместимость программ этого поколения с другими пакетами программ обусловлена за счет предоставления результатов выравнивания в виде формата FASTA.

Последним представителем серии является программа Clustal X, для которой характерен более удобный интерфейс и более

легкая оценка результатов выравниваний. В настоящее время именно последние программы серии Clustal этого поколения (версия 1.83) позволяют создавать наиболее биологически корректные множественные выравнивания дивергировавших последовательностей.

Программы третьего поколения серии Clustal доступны на многих серверах (http://npsa-pbil.ibcp.fr, http://www.ebi.ac.uk) в двух вариантах – интерактивном и почтовом. Интерактивный вариант предполагает ожидание пользователем получения результатов выравнивания (целесообразно применять при небольшом (<100) количестве последовательностей), а почтовый – по электронной почте (применяется при большом числе последовательностей).

Принципы работы CLUSTAL. Первоначально необходимо ввести на одном из серверов изучаемые аминокислотные или нуклеотидные последовательности в одном из 7 возможных форматов (NBRF/PIR, EMBL/SWISSPROT, Pearson (Fasta), Clustal (*.aln), GCG/MSF (Pileup), GCG9/RSF, GDE). Наиболее часто используется формат FASTA, сущность которого заключатся во введении знака «>» перед названием каждой последовательности, а затем (с новой строки) однобуквенном обозначении аминокислот и нуклеотидов. Суммарная длина вводимых последовательностей не должна превышать 40000 для WWW и 60000 для e-mail серверов.

При использовании данной программы выравнивание состоит из трех этапов: парных выравниваний, построения направляющего дерева и множественного выравнивания.

В ходе парных выравниваний предварительно сравниваются все возможные пары изучаемых последовательностей. На основании проведенных сравнений вычисляются показатели сходства в соответствии с выбранными матрицами.

Существуют 2 разновидности парного выравнивания: медленное (slow) и быстрое (fast). Медленное выравнивание является более точным, но его не рекомендуется применять в случае большого количества (более 20) последовательностей значительной длины (более 1000 остатков). Медленное выравнивание характеризуется 4 параметрами:

- штрафом на внесение делеции (gap open penalty). Уменьшение этого параметра способствует внесению разрывов в выравнивание,

что ухудшает качество. Увеличение – приводит к тому, что выравнивание будет представлять собой длинные участки последовательностей почти без вставок или делеций.

- штраф на продолжение делеции (gap extension penalty). Этот параметр контролирует возможность внесения длинных вставок или делеций.
- матрица сравнений нуклеотидов (DNA weight matrix, Clustal W 1.6). В наиболее широко используемой матрице DNA identity совпадение нуклеотидов оценивается в 1 балл, а несовпадение – -10000 баллов. Такой высокий штраф за несоответствие облегчает внесение пробелов.
- матрица сравнения аминокислот (protein weight matrix) – PAM, Blosum и Gonnet.

Выбор матрицы оказывает большое влияние на получаемые результаты, так как каждая матрица представляет отражение отдельных эволюционных гипотез. Известно, что все замены аминокислот не являются равновероятными и в ходе эволюции чаще происходят замены на сходные по физико-химическим свойствам аминокислоты. Так в ходе эволюции гидрофобный изолейцин достаточно часто заменяется на гидрофобный валин и редко на гидрофильный цистеин.

Исследования эволюционных изменений различных белковых семейств позволили установить частоты фиксированных мутаций аминокислот и нуклеотидов и обобщить полученную информацию в виде матриц. В настоящее время используются серии белковых матриц Blosum, PAM и Gonnet. Матрицы серии Blosum преимущественно используются при проведении локальных выравниваний (поиск сходных последовательностей по базам данных).

Матрицы серии PAM, предложенные М. Дэйхофф, широко используются с 70-х годов. Основными отличиями матриц PAM и Blosum являются:

1) использование матрицами PAM простой эволюционной модели (подсчет замен на ветвях филогенетического древа);
2) матрицы PAM основаны на учете мутаций по принципу глобального выравнивания (в высококонсервативных и

высокомутабельных участках), а матрицы Blosum – локального (только высококонсервативных участков);
3) для матриц PAM замены в группах последовательностей подсчитываются сходным образом.

Матрицы этих двух серий сопоставимы следующим образом PAM 100 – Blosum 90, PAM 120 – Blosum 80, PAM 160 – Blosum 60, PAM 200 – Blosum 52, PAM 250 – Blosum 45. Наиболее часто используются матрицы Blosum 62 и PAM 160 (при среднем сходстве последовательностей). При выравнивании близко родственных последовательностей следует использовать матрицы Blosum с большим порядковым номером и матрицы PAM с меньшим номером.

Матрицы Gonnet (рис. 4) представляют собой усовершенствованный вариант матриц Дэйхофф, основанный на большей базе данных. Использование этой матрицы наиболее целесообразно для инициальных сравнений.

Геном - совокупность информации, передаваемой живыми существами по наследству. Геномика – это наука, изучающая геном. В её задачи входят, среди прочего, секвенирование геномов и определение механизма связи между генами и признаками. Мутация - это самопроизвольное или спровоцированное изменение генома, передающееся по наследству. Мутации могут быть точечными, а могут затрагивать большие участки ДНК. Считается, что мутации являются необходимым фактором эволюции и образования новых видов.

В настоящее время существует несколько способов электронного представления генетических данных. Среди них наиболее популярные — форматы FASTA и FASTQ. Оба эти формата текстовые и кодируют каждый нуклеотид одной буквой.
Существующие алгоритмы позволяют выравнивать биологические последовательности. Это позволяет оценить их степень сходства, но без учёта не точечных мутаций.

На молекулярном уровне гены представляют собой участки молекул ДНК или РНК. Большая часть живых существ хранят свои гены в нескольких длинных молекулах ДНК, находящихся в клеточном ядре. При экспрессии генов соответствующий участок ДНК транскрибируется, в результате чего создаётся молекула

транспортной РНК (тРНК), которая в дальнейшем может используется для синтеза молекулы белка.

Генетический код - способ кодирования структуры белков при помощи последовательности нуклеотидов. Молекула белка представляет собой цепочку из аминокислот. Хотя общее число различных аминокислот достаточно велико, лишь небольшое число аминокислот, так называемые стандартные аминокислоты, могут становиться частью молекул белка. Всего известно 20 стандартных аминокислот.

Поскольку число возможных триплетов больше, чем число стандартных аминокислот, некоторые аминокислоты кодируются несколькими возможными триплета- ми (до шести триплетов для одной аминокислоты). Это придаёт коду некоторую степень помехоустойчивости - некоторые изменения последовательности нуклеотидов не приводят к изменению последовательности аминокислот, которую она кодирует.

Представление генетической информации в электронном виде. Задача электронного представления генетической информации встала перед исследователями, когда появились устройства, способные считывать эту информацию. Поскольку различных нуклеотидов и стандартных аминокислот немного, логично, что их стали кодировать одним символом. Обычно для кодирования нуклеотида ис- пользуется первая буква в его названии. В то же время, названия многих аминокислот начинаются с одинаковых букв, поэтому коды многих аминокислот не совпадают с из первыми буквами — используются те буквы, которые не заняты.

Формат FASTA. FASTA — один из наиболее популярных форматов для представления последовательностей нуклеотидов или аминокислот. Файл в формате FASTA - простой текстовый файл. Первая строка должна начинаться с символа «>» или «;». Она содержит имя последовательности и некоторую дополнительную информацию, предназначенную для идентификации. Другие строки, начинающиеся с «;», являются комментариями и игнорируются.

После первой строки начинается, собственно, описание последовательности. При кодировании последовательности

нуклеотидов, буквы A, C, G, T и U кодируют, соответственно, аденин, цитозин, гуанин, тимин и урацил. Также некоторые буквы кодируют позиции, в которых находится один нуклеотид из некоторого множества (это используется, если неизвестно, какой именно нуклеотид там находится). Символ - (дефис) кодирует неизвестную последовательность произвольной длины.

Формат FASTQ. FASTQ - формат представления биологической последовательности совместно с данными о качестве. Этот формат используется для представления данных секвенирования, так как позволяет представить как саму последовательность, так и вероятность, что каждый из элементов последовательности указан правильно. Для этого, кроме символов, кодирующих элементы последовательности, используется символы, кодирующие уровень качества.

Уровень качества - целое число в некотором диапозоне. Известно два различных способа выражать уровень качества через вероятность ошибки. Чаще всего используется следующая формула: $Q = -10\log(p)$ Здесь Q - уровень качества, p - вероятность, что этот элемент последовательности - ошибочный.

Сам же файл содержит четыре строки для каждой последовательности. Первая строка начинается с символа «@», после которого идёт описание последовательности, как и в формате FASTA. Следующая строка содержит последовательность символов, кодирующих саму последовательность, аналогично формату FASTA. За ней идёт строка, начинающаяся с символа «+», после которого может идти описание последовательности (третья строка будет отличаться от первой только тем, что первый символ заменён на «+»), а может ничего не идти.

Последняя строка содержит уровни качества. Её длина равна длине второй строки, а каждый символ кодирует информацию о качестве элемента последовательности, закодированного соответствующим символом второй строки. Сами же уровни кодируются таким образом: ASCII-код символа равен уровню качества плюс неко- торая константа. Константа обычно имеет значение 33 или 64. В любом случае, код символа не должен превышать 127.

Формат GenBank. Формат GenBank позволяет представить больше дополнительной информации о последовательности. Файл в формате GenBank состоит из нескольких записей, каждая из которых может занимать несколько строк. Все строки в записи, кроме первой, начинаются с пробела, это позволяет легко находить границы записей. Каждая запись начинается с имени, за которым идёт значение (через один или несколько пробелов). Некоторые записи могут содержать подзаписи. Они форматируются аналогично записям.

После этого в начало каждой строки записывается несколько пробелов. Таким образом, значения подзаписей выравнены правее значений записей, а те выравнены правее имён записей. Имена записей имеют предопределённое значение. Например, запись DESCRIPTION хранит описание последовательности, запись SOURCE идентифицирует особь, с кото- рой считана последовательность, а запись REFERENCE (их может быть несколько) используется для ссылок на публикации. Запись ORIGIN содержит саму последовательность.

Каждая строка, кроме последней, содержит 60 элементов, разбитых пробелами на группы по 10. Для представления элементов используются строчные буквы. В начало каждой строки добавляется текужая позиция, начиная с единицы, и пробел.

Существующие методы поиска гомологий в биологических последовательностях.

Гомология - структурное сходство. В генетике под гомологиями понимаются участки белков или ДНК, имеющую сходную последовательность аминокислот или нуклеотидов. Существующие методы поиска гомологий умеют находить участки (подстроки) двух последовательностей, которые отличаются не очень сильно.

Алгоритм Нидлмана-Вунша. Алгоритм Нидлмана-Вунша впервые был опубликован в 1970 году и позволяет определять степень сходства последовательностей, а также находить глобальное выравнивание, - находить, какой именно символ из одной последовательности соответствует некоторому символу из другой последовательности.

Для своей работы алгоритм использует матрицу сходства, которая указывает, насколько схожими считать разные нуклеотиды. Для различных нуклеотидов используются отрицательные элементы. Поэтому последовательности должны содержать некоторую долю совпадающих нуклеотидов, для того чтобы быть признанными гомологичными.

Использование матрицы позволяет придавать разный вес разным заменам нуклеотидов. Например, поскольку транзиции более вероятны, чем трансверсии, логично считать последовательности, отличающиеся заменой пурина на пурин или пиримидина на пиримидин, более схожими, чем те, которые отличаются заменой пурина на пиримидин или наоборот.

Вообще, матрица позволяет приписать любой вес любым заменам. Обычно используется симметричная матрица, однако применение несимметричной матрицы позволяет различать замены в одну и в другую сторону.

Пример матрицы сходства:

	А	Г	Т	Ц
А	10	−1	−4	−3
Г	−1	7	−3	−5
Т	−4	−3	8	0
Ц	−3	−5	0	9

Здесь А, Г, Т и Ц обозначают, соответственно, аденин, гуанин, тимин и цитозин, а числа в матрице указывают степень сходства между двумя нуклеотидами.

Алгоритм Нидлмана-Вунша способен сопоставлять символы двух последовательностей так, что сумма значений сходства для соответствующих символов максимально. Кроме того, алгоритм может учитывать вставки и удаления. При этом считается, что символ в одной строке, которому не соответствует никакой символ из другой строки, имеет некоторый уровень сходства, который является параметром алгоритма (например, −5).

Алгоритм Смита-Вотермана. Алгоритм Смита-Вотермана аналогичен алгоритму Нидлмана-Вунша, но при этом решается задача локального выравнивания: находит подстроки первой и второй строк, обладающие максимальным сходством, а также выравнивает их. Алгоритм был опубликован в 1981 году. Как и

алгоритм Нидлмана-Вунша, алгоритм Смита-Вотермана использует матрицу сходства. Это позволяет учитывать различные замены с различным весом. Также он позволяет учитывать разные добавления и удаления по-разному в зависимости от того, какой именно нуклеотид был добавлен или удалён.

Как и алгоритм Нидлмана-Вунша, алгоритм Смита-Вотермана всегда находит оптимальное решение. Однако время работы и занимаемая память делают эти алгоритмы неприемлемыми для работы с большим количеством генетического материала.

Критерий сходства биологических последовательностей.

Чтобы оценивать сходство двух последовательностей, необходимо придумать некоторую меру сходства. Предложенная в данной работе мера сходства - число от нуля до единицы, где единица соответствует максимальному сходству, а ноль — минимальному. Кроме того, мера удовлетворяет следующим свойствам:

- Сходство последовательности с самой собой равно единице.
- Сходство несхожих последовательностей обычно невелико.
- Сходство не изменяется значительно, если одну из последовательностей подвергнуть мутации. Чем менее вероятна мутация, тем сильнее изменяется мера сходства.
- Существует способ эффективно вычислять меру сходства для практически встречающихся последовательностей.
- Сложность разработки подобной меры заключается в том, что нелокальные мутации могут приводить к значительным перестановкам частей последовательности.

Рассмотрим типовой способ решения проблемы. Назовём одну из последовательностей исходной, а другую — целевой. Будем рассматривать все возможные разбиения целевой последовательности на фрагменты (подстроки) и для каждого из них подсчитаем степень различия — число, тем большее, чем хуже, с точки зрения данного разбиения, целевая последовательность аппроксимирует исходную. Затем найдём минимум степени различия по всем раз- биениям.

В каждом разбиении каждый фрагмент можно считать свободным или связанным. От этого зависит вклад этого фрагмента в степень различия. Вклад свободного фрагмента в

степень различия пропорционален его длине с константой, являющейся параметром метода, и не зависит от его содержимого. Вклад же связанного фрагмента зависит от его содержимого: каждому связанному фрагменту сопоставляется подстрока исходной последовательности или комплементарной к ней, и вклад фрагмента в степень различия равен редакционному расстоянию между содержимым фрагмента и этой подстрокой. Здесь редакционное расстояние рассматривается в обобщённом смысле: можно приписывать различные веса разным заменам, вставкам и удалениям.

Как и разбиение на фрагменты, назначение типов фрагментов и соответствующих подстрок происходит таким образом, чтобы минимизировать суммарную степень различия. Кроме того, чтобы сделать большое число коротких фрагментов менее оптимальным, к степени различие добавляется ещё одно слагаемое, пропорциональное общему числу фрагментов в разбиении.

Мера сходства вычисляется по степени различия путём применения линейного преобразования, переводящего нулевую степень различия в единицу, а максимально возможную степень различия в ноль. Поскольку одним из разбиений является разбиение на один свободный фрагмент, степень различия не превосходит такую, которая соответствует этому разбиению, поэтому удобно принимать это значение соответствующим нулевой степени сходства.

Если редакционное расстояние между связанным фрагментом и соответствующей подстрокой достаточно велико, то такое разбиение заведомо не оптимально: можно сделать этот фрагмент свободным и таким образом уменьшить суммарную степень различия.

Более того, даже если фрагмент содержит немного изменений относительно его длины, но все эти изменения расположены недалеко друг от друга, можно разбить этот фрагмент на три фрагмента, вырезав из него кусок, содержащий большую часть изменений, и заменив его свободным фрагментом. В результате, во многих случаях можно не считать (или не продолжать считать) редакционное расстояние между фрагментами, поскольку и так понятно, что оптимальное разбиение не содержит такой фрагмент.

Кроме того, использование редакционного расстояния не даёт локальным мутациям существенно изменять степень различия: если мутация попала в свободный фрагмент, то степень различия может лишь слегка измениться за счёт изменения длины фрагмента, а если мутация попала в связанный фрагмент, то, опять же, соответствующее изменение редакционного расстояния не будет большим.

В отличии от локальных мутаций, которые сказываются на редакционном расстоянии, нелокальные мутации непосредственно сказываются на самом разбиении. Например, если подвергнуть последовательность дупликации (вставить в неё копию её подстроки), то в соответствующее разбиение можно тоже вставить копию его подстроки (при этом крайние фрагменты могут оказаться обрезанными). При этом вклад в степень различия от добавленных фрагментов не превосходит вклад от исходных фрагментов. Кроме того, некоторый вклад в степень различия происходит из-за того, что тот фрагмент, в котором находится место вставки, оказывается разделён на два.

Другие нелокальные мутации учитываются аналогично. При учёте инверсий используется то, что подстроки, соответствующие связанным фрагментам, можно брать как из исходной последовательности, так и из комплементарной к ней.

Представление последовательности нуклеотидов в памяти компьютера. Существует несколько форматов файлов, предназначенных для хранения нуклеотидных последовательностей. Однако, эти форматы создавались для удобного восприятия человеком, а также для совместимости с программами и протоколами, рассчитанными на работу с текстовыми данными.

В то же время, для работы с нуклеотидной последовательностью желательно такое представление, с которым можно быстро выполнять следующие операции:

- Индексация - определение нуклеотида по порядковому номеру.
- Взятие подстроки - получение последовательности, содержащей те из нуклеотидов исходной последовательности, порядковые номера которых лежат в заданном диапазоне.

- Сравнение - определение, совпадают ли две заданные последовательности.
- Хеширование - операция, получающая по заданной последовательности число таким образом, чтобы одинаковым последовательностям соответствовали одинаковые числа, а разным, по возможности, разные.

Форматы типа FASTA допускают наличие в файле произвольного числа символов перевода строки между символами, кодирующими нуклеотиды, что препятствует эффективной индексации и сравнению, поэтому этот формат не подходит для выполнения операций, при которых необходима возможность произвольного доступа к последовательности.

Вместо этого, используется двоичное представление последовательности. Поскольку в одной последовательности возможны четыре разновидности нуклеотидов, каждый нуклеотид кодируется двумя битами. Таким образом, каждый байт кодирует сразу четыре нуклеотида. Поскольку современные процессоры обрабатывают информацию блоками сразу по четыре или даже по восемь байт, подобное кодирование позволяет в несколько раз быстрее выполнять операции, которые не требуют обращения к индивидуальным нуклеотидам, например, сравнение, которое можно делать поблочно.

Кроме того, для представления подстрок используется механизм ссылок: вместо копирования данных при взятии подстроки создаётся объект, ссылающийся на исходные данные, с указанием смещения и длины подстроки.

Алгоритм сравнения генетических последовательностей. Сравнение нуклеотидных последовательностей происходит в два этапа.

На первом этапе создаётся словарь, позволяющий быстро находить участки исходной последовательности, совпадающие с заданными короткими последовательностями (в качестве которых используются подстроки целевой последовательности).

На втором этапе ищется оптимальное разбиение целевой последовательности на фрагменты. При этом словарь используется, для того чтобы быстро находить подстроки исходной последовательности, соответствующие связанным

фрагментам. Поскольку в оптимальном покрытии редакционное расстояние между связанным фрагментом и соответствующей ему подстрокой невелико, в них найдётся подстрока, совпадающая полностью. Каждое такое совпадение расширяется до тех пор, пока ошибок не станет настолько много, что дальнейшее расширение не имеет смысла.

Для подсчёта степени различия используется динамический алгоритм: подсчитывается степень различия для каждого префикса целевой последовательности, а фрагменты генерируются от начала целевой последовательности к её концу, что позволяет эффективно пересчитывать степени различия. Результатом является степень различия для всей целевой последовательности.

Создание словаря.

Для быстрого поиска схожих участков в данной работе используется словарь - структура, позволяющая по нуклеотидной последовательности определённой длины (ключу) быстро найти все её вхождения в исходной подпоследовательности, а также в последовательности, комплементарной к ней.

В качестве словаря используется хеш-таблица, в которой ключами служат последовательности, а значениями — списки вхождений. Для каждого вхождения указывается, в какой из строк находится это вхождение (исходной или комплементарной), а также смещение этого вхождения относительно начала последовательности.

Для создания словаря перебираются все подстроки исходной последовательности некоторой длины, а также подстроки последовательности, комплементарной к ней. Длина ключа является параметром алгоритма и подбирается так, чтобы обеспечить оптимальное потребление ресурсов: если ключи будут слишком короткими, то каждой подстроке будет соответствовать слишком длинный список позиций, поэтому увеличится время работы.

Если же ключи будут слишком длинными, то алгоритм будет потреблять слишком много памяти. Кроме того, из-за слишком длинных ключей алгоритм может начать пропускать фрагменты, в которых нет полностью совпадающих участков достаточной

длины. На практике оптимальная длина ключа получается около $\log_4(ls)$-1, где *ls* - длина исходной последовательности.

Для каждой подстроки соответствующей длины проверяется, есть ли она в словаре. Если есть, то в соответствующий список добавляется новое вхождение. Если нет, то в словарь добавляется новая запись, в котором ключом служит подстрока, а значением — новый список, содержащий единственное вхождение.

Поиск оптимального покрытия.

Поиск оптимального покрытия производится методом динамического программирования: для каждого префикса в целевой строке хранится минимальная степень различия и пересчитывается на основе информации о возможных фрагментах. Кроме минимальной степени различия хранится минимальная степень различия для случаев, когда в конце префикса находится свободный фрагмент. Это включает случаи, когда этот фрагмент имеет нулевую длину. Это требуется для того, чтобы легче было учитывать свободные фрагменты.

Пусть g_i — минимальная степень различия для префикса длины i, а h_i — минимальная степень различия для префикса длины i при условии, что последний фрагмент - свободный.

Тогда выполняется неравенство: $g_i \le h_i \le g_i + Df$.

Здесь Df — коэффициент к добавке к степени различия, пропорциональной числу фрагментов.

Во всех точках, кроме нуля, эти значения инициализируются бесконечностями. В нуле используются значения $g_0 = -Df$ (ноль фрагментов, поэтому отрицатель- ное число) и $h_0 = 0$ (здесь один пустой свободный фрагмент). Заметим, что g_0 — единственный из всех g_i и h_i может быть отрицательным, поскольку остальные соответствуют покрытиям, содержащим хотя бы один фрагмент.

Первое неравенство выполняется, потому что g_i - минимум на множестве покрытий префикса длины i, а h_i - минимум на подмножестве этого множества. Второе неравенство выполнено, так как к любому покрытию можно добавить свободный фрагмент длины ноль, что увеличит степень различия на Df .

В процессе работы алгоритма поддерживается инвариант, что для всех префиксов длины меньше некоторого i числа g_i и h_i уже вычислены и в дальнейшем не будут изменяться, а для остальных

префиксов значения ещё не вычислены, поэтому не будут использоваться. В процессе работы алгоритма i может только увеличиваться. Для того чтобы выполнить пересчёт для соответствующего фрагмента, необходимо выполнить операцию $g_e \leftarrow \min(g_e, g_b+d+Df)$, где b - индекс начала фрагмента, e - индекс его окончания, d - его вклад в степень различия. Существенно, что g_b уже вычислено, а g_e ещё нет.

При таком методе требуется выполнять пересчёт для каждого фрагмента в тот момент, когда позиция с индексом i лежит внутри него. Это условие выполняется автоматически, если начинать поиск фрагмента с ключа, начинающегося с позиции i. После того, как массивы g и h обновлены с учётом найденных связанных фрагментов, необходимо увеличить i на один. Перед этим следует пересчитать g_{i+1} и h_{i+1} для учёта свободных фрагментов. При этом g_{i+1} вычисляется по формуле: $g_{i+1} \leftarrow \min(g_{i+1}, h_i + Dl)$ (можно взять значение, полученное при пересчёте или продолжить свободную последовательность). Здесь Dl — вклад единицы длины свободного фрагмента в степень различия. Число h_{i+1} вычисляется по формуле: $h_{i+1} \leftarrow \min(h_i+Dl, g_{i+1}+Df)$ (можно продолжить свободную последовательность или начать новую). Заметим, что g_{i+1} может присваиваться до этого пересчёта, поэтому использование этого значения до присваивания осмысленно.

Алгоритм заканчивает работу, когда $i=l_t+1$- когда пройдена вся целевая строка. Тогда $D = gl_{t+1}$ по индукции.

Сканирование целевой строки

Чтобы быстро учесть все связанные фрагменты, получающиеся из вхождения ключа, используется такой метод: пусть ключ входит, начиная с позиции i целевой строки и j исходной. Поскольку сам ключ совпадает полностью, редакционное расстояние между фрагментом и соответствующей подстрокой складывается из двух слагаемых: редакционное расстояние между участками до ключа и после ключа.

Соответственно, если s и e — индексы начала и конца фрагмента в целевой строке, то выполняется неравенство:

$$g_e \leq g_s + d_{s,i} + d_{i,e}.$$

Здесь $d_{s,i}$ и $d_{i,e}$ — вклад в редакционное расстояние участков от s до i и от i до e. Чтобы обеспечить это неравенство, можно было

бы перебирать все возможные значения s и e в некотором диапазоне, но можно сделать это намного быстрее: вначале найти $\min_s(g_s+d_{s,i})$ перебором s, а затем перебирать только значения e.

Для определения границ перебора s и e используется изложенный выше критерий оптимальности: если для какого-то s выгодно заменить подстроку от s до i свободным фрагментом, то дальнейшее уменьшение s не имеет смысла. Аналогично для e.

Оценка сложности алгоритма. Поскольку представленный алгоритм использует эвристики, его сложность существенно зависит от входных данных. В частности, чем более схожи исходная и целевая последовательности, тем медленнее будет работать алгоритм.

Порядок выполнения работы:

1. Изучите теоретический материал.
2. Составьте блок-схему алгоритма сравнения двух генетических текстов по 4 «буквам» с подсчетом доли совпадений.
3. В информационных источниках найти две любых генетических последовательности и выполнить трассировку разработанного алгоритма. Результаты трассировки оформить в виде таблицы.
4. Средствами Excel выполните сравнение генетических текстов и сравните результаты, полученные в п.3. Сделайте выводы.
5. Оформите отчет, включающий результаты выполнения п.3 и п.4. и краткие ответы на контрольные вопросы (не менее 4).
6. Составьте аннотацию не менее 3 информационных источников, найденных в сети интернет по тематике лабораторной работы (объемом 150-350 слов).

Примечание. Сравнение генетических текстов в Excel можно осуществить следующим образом: преобразовать последовательность символов в генетических текстах, найти алгебраическую разницу между полученными кодами, подсчитайте количество подряд стоящих «0», если таковых более трех, оцените процент совпадений.

Контрольные вопросы.

1. Что характеризует нуклеотидная генетическая последовательность?

2. Как осуществляется выравнивание нуклеотидных последовательностей? Какие компьютерные программы для этого используются в настоящее время?
3. В чем заключаются принципы работы программного инсрументария CLUSTAL?
4. Как наука изучает информацию заключенную в геноме? Характеристики объектов и методологии исследования.
5. Общее и отличия форматов представления генетических данных FASTA, FASTQ и GenBank.
6. Каким образом осуществляется представление генетической информации в электронном виде.
7. Охарактеризуйте основные существующие методы поиска гомологий в биологических последовательностях?
8. Опишите математический аппарат обработки биоинформации в алгоритме Нидлмана-Вунша.
9. Опишите математический обработки биоинформации в алгоритме Смита-Вотермана.
10. По какому критерию оценивается сходство сходство биологических последовательностей.
11. Охарактеризуйте этапы алгоритма сравнения генетических последовательностей?
12. Каким образом создается словарь для сравнения генетических последовательностей?
13. Каким образом осуществляется поиск оптимального покрытия?
14. Как осуществляется сканирование целевой строки?

Библиография.

1. Бородовский М., Екишева С. «Задачи и решения по анализу биологических последовательностей». — М.-Ижевск: НИЦ «Регулярная и хаотичная динамика», 2018. — 420 с.
3. Компьютерный анализ генетических текстов Александров А.А., Александров Н.Н., Бородовский М.Ю. и др.- М.:Наука; 2000, - 267с.
4. Леск, Артур. Введение в биоинформатику [Текст]: пер. с англ. / под ред. А. А. Миронова, В. К. Швядаса. - М.: БИНОМ. Лаборатория знаний, 2019. - 318 с.

ЛАБОРАТОРНАЯ РАБОТА № 2. ГРАФИЧЕСКОЕ ПРЕДСТАВЛЕНИЕ ИНФОРМАЦИИ БИОМЕДИЦИНСКОГО ХАРАКТЕРА

Цель работы: изучение возможностей стандартных инструментариев ПЭВМ для отображения биомедицинской информации в различных статических и динамических формах.

Краткие теоретические сведения.

В процессе проведения биомедицинских и экологических исследований (включая системы мониторинга) осуществляется регистрация значений различных информационных показателей, характеризующих поведение исследуемого биообъекта как во времени так и пространстве. Для принятия решения о дальнейших наблюдениях и управлениях о биообъекте лицо принимающее решение (человек или некоторая автоматизированная система, обладающая свойствами и возможностями искусственного интеллекта) должно создать некоторый образ (модель) биообъекта. Простейший такой образ создается с помощью графических и аудио средств в динамике или статике, поскольку данные органы чувств у человека, как наиболее типичного лица принимающего решения в биомедицинских исследованиях, наиболее развиты.

Для искусственного интеллекта, применяемого в автоматизированных системах поддержки принятия решений, данные образы представляются многомерными кортежами данных или лингвистическими переменными.

Мало исследованными направлениями в данной области генерации новых знаний являются возможности создания и использования образов нацеленных на другие органы чувств (различных анализаторов человеческого мозга) человека (различных анализаторов головного и-или спинного мозгов) или использования «человеческих» образов как вторичной интегральной информации для компьютерных систем.

Визуальное представление биомедицинской информации в различных компьютеризированных системах осуществляется в следующих направлениях:

- дружественный интерфейс (в том числе графический и аудио);
- оптические изображения физических биообъектов, регистрируемые специализированной аппаратурой с помощью различных методов;
- статические изображения: графики, рисунки, «смайлики», диаграммы;
- изображения, отражающие результаты имитационного моделирования;
- таблицы с определенными образом выделенными информационными составляющими (структурами);
- динамические объекты изображений и другие мультимедийные средства.

Основной задачей визуализации данных является задача получения образа, однозначно соответствующего полученному сигналу. Обычно под *визуализацией биомедицинских данных* понимается построение графиков функций и поверхностей. В медицинских исследованиях в качестве индикатора общего состояния организма определена сердечно-сосудистая система, наиболее информативным показателем которой является ритм сердечных сокращений. Математический анализ сердечного ритма позволяет получить информацию, характеризующую состояние регуляторных механизмов, так как в ряде кардиосигналов содержится информация не только о деятельности регуляторных механизмов сердечно-сосудистой системы, но и о многочисленных функциях целостного организма. Наиболее перспективным диагностическим направлением является метод электрокардиографии.

Для отображения графической информации возможно использование готовых компонентов, содержащихся в используемой среде разработки . Такие компоненты, как правило, содержат множество свойств, методов и событий.

При этом все функции построения графиков выполняются средствами центрального процессора без использования специализированных технологий, поддерживаемых видеоадаптером. Этот недостаток приводит к необходимости разработки собственных компонент эффективного отображения

графической информации использующих возможности графических адаптеров.

Анализ информационных источников показывает, что визуализация значительного объема данных невозможна без применения специализированных графических технологий, поддерживаемых видеоадаптером . Среди существующих графических компонентов, отличающихся быстродействием можно выделить «MultipleLayer AFM Surface3D PRO» , «3D-Splot» , HOOPS Visualize .

Приведенные компоненты являются коммерческими продуктами, имеющими высокую стоимость. В этой связи обычно принимается решение о разработке собственного компонента на основе специализированных технологий, например, доступных в операционных системах Microsoft-Windows – DirectX3D или OpenGL.

Обобщенный алгоритм типовой визуализации представлен на рисунке 1.

Рис. 1. Алгоритм визуализации.

В соответствии с данным алгоритмом, программное представление результата вычислений, визуализацию которого должен осуществлять компонент, представляет собой двумерный массив вещественных значений. Таким образом, каждый элемент набора данных, подлежащий исследованию, характеризуется тремя величинами: индексом строки массива, индексом столбца массива и величиной, хранимой в ячейке массива. Естественным представлением множества точек, характеризуемых тремя величинами, является поверхность. Таким образом ,визуализация заключается в построении изображения проекции трехмерного объекта. Дискретный характер определения зависимости параметров элементов набора данных не позволяет представить их в виде непрерывной поверхности. В качестве удовлетворительного приближения к идеальному визуальному представлению может быть использовано построение поверхности в виде множества плоских многоугольников, вершины которых расположены в опорных точках, соответствующих представлению элементов набора данных. Возможность идентификации параметров элементов набора данных предоставляется за счет совмещения объекта поверхности с системой линий-отметок и цифровых и буквенных обозначений.

Для построения модели поверхности на основе двумерного массива значений массив интерпретируется как горизонтальная регулярная карта высот. Номера строк и номера столбцов массива отображаются в координаты на горизонтальной плоскости, а величина, хранящаяся в соответствующей ячейке массива, интерпретируется как высота точки над такой плоскостью. Интерпретация карты высот проиллюстрирована на рис. 2.

Строки массива входных данных обрабатываются попарно для формирования последовательности смежных треугольников, образующих элемент поверхности.

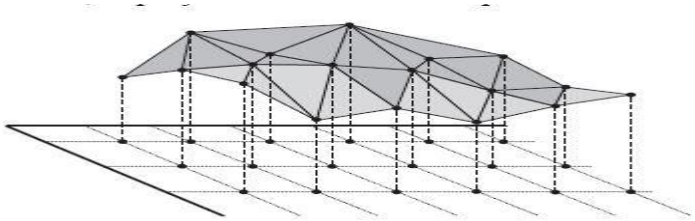


Рис. 2. Интерпретация исходных данных

Особенностью разработанного алгоритма и дополнительным способом визуализации одного из параметров элементов набора является присвоение точкам поверхности цвета в зависимости от величины одной из координат точки. Отображение величин в цветовые значения осуществляется по цветовым шкалам. Цветовые шкалы представляют собой упорядоченный набор цветовых значений, соответствующих пространству кодируемой величины. Пространство допустимых значений кодируемой величины разделено на смежные интервалы равной величины, и цвета шкалы соответствуют границам таких интервалов.

В большинстве случаев, в настоящее время в биомедицинских исследованиях используются пакеты статистической обработки данных со встроенными компонентами визуализации. Продолжительное время анализ медицинских данных был уделом специалистов, так как это требовало серьезной предварительной подготовки. С появлением и совершенствованием современных программ обработки данных статистическая обработка поднялась на новый уровень. Теперь исследователь-медик может и не иметь математической подготовки. Достаточно оперировать статистическими понятиями и, самое главное, правильно выбрать метод анализа. Все осуществимо благодаря компьютеру и новейшим программам.

Все программы статистической обработки данных можно разделить на профессиональные, полупрофессиональные (популярные) и специализированные. Статистические программы относятся к наукоемкому программному обеспечению, цена их часто недоступна индивидуальному пользователю.

Профессиональные пакеты имеют большое количество методов анализа, популярные пакеты - количество функций, достаточное для универсального применения.

Специализированные же пакеты ориентированы на какую-либо узкую область анализа данных. Создатели программных статистических пакетов заявляют, что их продукт превосходит аналоги. Отсутствие у большинства исследователей времени для освоения нескольких программ, делает непростым ее выбор. В данной статье приведена базовая информация о присутствующих на рынке основных полупрофессиональных программных пакетах пригодных для статистической обработки биомедицинских данных.

MS Excel. Самой часто упоминаемой (и используемой) в отечественных статьях является приложение MS Excel из пакета офисных программ компании Microsoft ? MS Office. Причины этого кроются в широком распространении этого программного обеспечения, наличии русскоязычной версии, тесной интеграцией с MS Word и PowerPoint. Однако, MS Excel - это электронная таблица с достаточно мощными математическими возможностями, где некоторые статистические функции являются просто дополнительными встроенными формулами. Расчеты сделанные при ее помощи не признаются авторитетными биомедицинскими журналами. Также в MS Excel невозможно построить качественные научные графики. Безусловно, MS Excel хорошо подходит для накопления данных, промежуточного преобразования, предварительных статистических прикидок, для построения некоторых видов диаграмм. Однако окончательный статистический анализ необходимо делать в программах, которые специально созданы для этих целей. Существует макрос-дополнение XLSTAT-Pro http://www.xlstat.com для MS Excel который, включает в себя более 50 статистических функций, включая анализ выживаемости, которых в основных случаях достаточно для обычного применения. Пробную версию макроса можно взять на сайте производителя.

STADIA. Программа отечественной разработки с 16-и летней историей. Включает в себя все необходимые статистические функции. Она прекрасно справляется со своей задачей - статистическим анализом. Но. Программа внешне фактически не изменяется с 1996 года. Графики и диаграммы, построенные при помощи STADIA, выглядят в современных презентациях архаично.

Цветовая гамма программы (красный шрифт на зеленом) очень утомляет в работе. К положительным качествам программы можно отнести русскоязычный интерфейс и наличие книг описывающих работу. Самый часто используемый пакет статистической обработки данных с более чем 30-и летней историей http://www.spss.com Отличается гибкостью, мощностью применим для всех видов статистических расчетов применяемых в биомедицине. Существует русскоязычное представительство компании http://www.spss.ru, которое предлагает полностью русифицированную версию SPSS 12.0.2 для Windows.

STATA. Профессиональный статистический программный пакет с data-management system, который может применятся для биомедицинских целей. Один из самых популярных в образовательных и научных учреждениях США наряду с SPSS. Официальный сайт http://www.stata.com. Программа хорошо документирована, издается специальный журнал для пользователей системы. Однако возможности предварительного ознакомления с демо-версией нет.

STATISTICA. Производителем программы является фирма StatSoft Inc. (США) http://www.statsoft.com которая выпускает статистические приложения, начиная с 1985 года. STATISTICA включает большое количество методов статистического анализа (более 250 встроенных функций) объединенных следующими специализированными статистическими модулями: Основные статистики и таблицы, Непараметрическая статистика, Дисперсионный анализ, Множественная регрессия, Нелинейное оценивание, Анализ временных рядов и прогнозирование, Кластерный анализ, Факторный анализ, Дискриминантный функциональный анализ, Анализ длительностей жизни, Каноническая корреляция, Многомерное шкалирование, Моделирование структурными уравнениями и др. Несложный в освоении этот статистический пакет может быть рекомендован для биомедицинских исследований любой сложности.

JMR. Один из мировых лидеров в анализе данных. Развивает этот статистический пакет SAS Institute http://www.jmp.com который выкупил в конце 2002 года известную статистическую программу StatView. Однако особых

преимуществ для медико-биологической статистики этот программный продукт не имеет.

SYSTAT Статистическая система для персональных компьютеров http://systat.com Последняя 11 версия обладает неплохим интуитивно понятным интерфейсом. Компания Systat Software также разрабатывает популярные у отечественных исследователей SigmaStat и SigmaPlot, которые являются соответственно, программой статистической обработки и программой построения диаграмм. При совместной работе становятся единым пакетом для статистической обработки и визуализации данных.

NCSS. Программа развивается с 1981 года и рассчитана на непрофессионалов в области статистической обработки. Интерфейс системы многооконный и как следствие этого явления - немного непривычный в использовании. Все действия пользователя сопровождаются подсказками. Сейчас доступна версия 2004 г. С сайта http://www.ncss.com можно переписать полнофункциональную пробную версию, работающую 30 дней.

MINITAB 14. Статистический пакет MINITAB в настоящее время выпускается в версии 14. С сайта производителяhttp://www.minitab.com можно взять полнофункциональный пробный вариант программы, которая работает 30 дней. Это достаточно удобный в работе программный пакет, имеющий хороший интерфейс пользователя, хорошие возможности по визуализации результатов работы. Имеет подробную справку.

STATGRAPHICS PLUS. Довольно мощная статистическая программа. Содержит более 250 статистических функций, генерирует понятные, настраиваемые отчеты. Последняя доступная версия - 5.1. Ее можно получить на сайте http://www.statgraphics.com. Есть возможность скачать демо-версию. Следует отметить, что ранние версии этой программы были весьма популярны у отечественных исследователей.

PRISM. Эта программа создавалась специально для биомедицинских целей. Интуитивно понятный интерфейс позволяет в считанные минуты проанализировать данные и построить качественные графики. Программа содержит основные

часто применяемые статистические функции, которых в большинстве исследований будет достаточно. Однако, как отмечают сами разработчики, программа не может полностью заменить серьезных статистических пакетов. На сайте http://www.graphpad.com помимо возможности ознакомления с демо-версией Prism можно получить справочник в формате PDF по биомедицинской статистике.

К рекомендациям выбора программ можно отнести:

- Если нужен мощный, общепризнанный пакет с простым и понятным даже начинающим пользователям интерфейсом, то лучше воспользоваться SPSS.
- Для начинающих и профессионалов, которым нужна подсказка и развитая документация на русском языке, можно рекомендовать STATISTICA. Это мощное приложение с профессиональными возможностями.
- Для непритязательных пользователей, которые ограничиваются в своих исследованиях стандартными статистическими методами можно рекомендовать англоязычную программу Prism.

Виды графиков.

Существует множество видов графических изображений (рис. 3 и 4). Классификация графиков основвывается на ряде признаков:

1. способ построения графического образа;
2. геометрические знаки, изображающие статистические показатели;
3. задачи, решаемые с помощью графического изображения.

Рис. 3. Классификация статистических графиков по форме графического образа.

По способу построения статистические графики делятся надиаграммы и статистические карты.

Диаграммы - наиболее распространенный способ графических изображений. Это графики количественных отношений. Виды и способы их построения разнообразны. Диаграммы применяются для наглядного сопоставления в различных аспектах (пространственном, временном и др.) независимых друг от друга величин: территорий, населения и т. д. При этом сравнение исследуемых совокупностей производится по какому-либо существенному варьирующему признаку.

Рис. 4. Классификация статистических графиков по способу построения и задачам изображения

Статистические карты - графики количественного распределения по поверхности. По своей основной цели они близко примыкают к диаграммам и специфичны лишь в том отношении, что представляют собой условные изображения статистических данных на контурной географической карте, т. е. показывают пространственное размещение или пространственную распространенность статистических данных.

Геометрические знаки, как было сказано выше, - это либо точки, либо линии или плоскости, либо геометрические фигуры. В соответствии с этим различают графики точечные, линейные, плоскостные и пространственные (объемные).

При построении точечных диаграмм в качестве графических образов применяются совокупности точек; при построении линейных - линии. Основной принцип построения всех плоскостных диаграмм сводится к тому, что статистические величины изображаются в виде геометрических фигур и, в свою

очередь, подразделяются на: столбиковые, полосовые, круговые, квадратные, фигурные.

Статистические карты по графическому образу делятся на картограммы и картодиаграммы.

В зависимости от круга решаемых задач выделяют

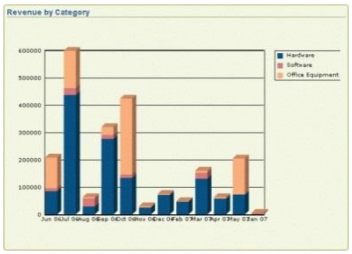
- диаграммы сравнения,
- структурные диаграммы
- диаграммы динамики.

Особым видом графиков являются диаграммы распределения величин, представленных вариационным рядом. Это гистограмма, полигон, огива, кумулята.

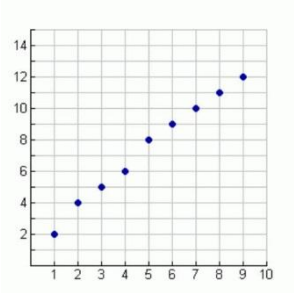
ВИДЫ ДИАГРАММ

Прежде чем составить какой либо график, необходимо определиться с вопросом о том, какие виды диаграмм вас именно интересуют. Рассмотрим основные из них.

Гистограмма. Само название этого вида позаимствовано из греческого языка. Дословный перевод – писать столбом. Это своеобразный столбчатый график. Диаграммы в Excel такого вида могут быть объемные, плоские, отображать вклады (прямоугольник в прямоугольнике) и т.д.

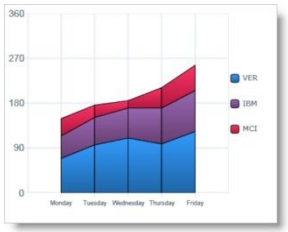


Точечная диаграмма показывает взаимную связь между числовыми данными в некотором количестве рядов и представляет собой пару групп цифр или чисел в виде единственного ряда точек в координатах. Виды диаграмм такого типа отображают кластеры данных, используются для научных целей. При предварительной подготовке к построению точечной диаграммы все данные, которые вы хотите расположить по оси ординат, следует расположить в одной строке/столбце, а значения по оси «абсцисс» - в смежной строке/столбце.

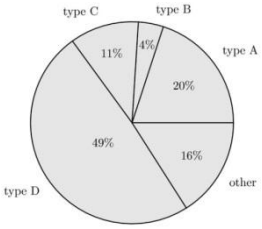


Линейчатая диаграмма и график. Диаграмма линейчатая описывает некое соотношение отдельных данных. На такой диаграмме значения располагаются по вертикальной оси, категории же – по горизонтальной. Из этого следует, что большее внимание такая диаграмма уделяет сопоставлению данных, нежели изменениям, происходящим с течением времени. Данный вид диаграмм существует с параметром «накопление», что позволяет показать взнос отдельных частей в общий конечный результат. График же отображает последовательность изменений числовых значений за абсолютно равные промежутки времени. Эти виды диаграмм наиболее часто используются для построений.

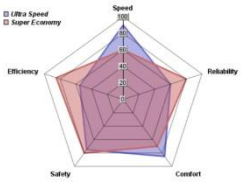
Диаграммы с областями. Основной целью такой диаграммы является акцент на величине изменения данных в течение некоторого периода, путем показа суммирования введенных значений. А также отображение доли отдельно взятых значений в общей сумме.



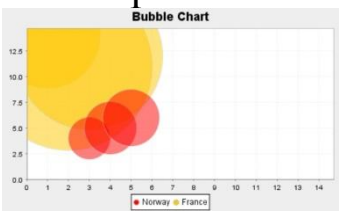
Кольцевая и круговая диаграммы. Данные виды диаграмм весьма схожи по целям. Обе они отображают роль каждого элемента в общей сумме. Их отличие заключается лишь в том, что диаграмма кольцевая имеет возможность содержать несколько рядов с данными. Каждое отдельное вложенное кольцо представляет собой индивидуальный ряд значений/данных.



Лепестковая. В данном случае каждая категория представляет индивидуальную координатную ось, исходящую от нулевой точки координат. Данный вид диаграмм позволяет сравнивать общие значения из некоторого количества введенных данных.



Пузырьковая. Одна из разновидностей точечной. Величина маркера зависит от величины третьей переменной. При предварительной подготовке располагать данные следует точно так же, как и при подготовке к построению точечной диаграммы.



Биржевая диаграмма. Использование таковой часто является неотъемлемым процессом при продаже акций или других ценных бумаг. Также возможно ее построение для наглядного определения изменения температурных режимов. Для трех и пяти значений такой вид графика может содержать в себе пару осей: первую – для столбиков, которые представляют интервал неких колебаний, вторую – для изменения ценовой категории.

Это лишь малая часть типов диаграмм, которые могут вам понадобиться. Выбор всегда зависит от целей.

Мультимедийные средства.

Понятие мультимедиа, вообще, и средств мультимедиа, в частности, с одной стороны тесно связано с компьютерной обработкой и представлением разнотипной информации и, с другой стороны, лежит в основе функционирования средств ИКТ, существенно влияющих на эффективность образовательного процесса. *Мультимедиа* - это:

- технология, описывающая порядок разработки, функционирования и применения средств обработки информации разных типов;
- информационный ресурс, созданный на основе технологий обработки и представления информации разных типов;
- компьютерное программное обеспечение, функционирование которого связано с обработкой и представлением информации разных типов;
- компьютерное аппаратное обеспечение, с помощью которого становится возможной работа с информацией разных типов;
- особый обобщающий вид информации, которая объединяет в себе как традиционную статическую визуальную (текст, графику), так и динамическую информацию разных типов (речь, музыку, видео фрагменты, анимацию и т.п.).

Таким образом, в широком смысле термин "мультимедиа" означает спектр информационных технологий, использующих различные программные и технические средства с целью наиболее эффективного воздействия на пользователя (ставшего одновременно и читателем, и слушателем, и зрителем).

Средства, используемые при создании мультимедийных продуктов:

- системы обработки статической графической информации;
- системы создания анимированной графики;
- системы записи и редактирования звука;
- системы видеомонтажа;
- системы интеграции текстовой и аудиовизуальной информации в единый проект.

Рассмотрим более подробно современные мультимедийные средства, перспективные для медицинского использования.

- *3D Очки.* Область применения – эффект присутствия для обучающихся во время реально проводимой операции, обследования больного.
- *Web-Камеры.* Web-камера - это стационарно установленная камера, имеющая встроенный web-сервер, сетевой интерфейс и подключающаяся непосредственно к LAN/ WAN/ Internet. Многие сетевые камеры имеют такие дополнительные средства как: детекторы движения, отправка сообщений по e-mail, работа с модемом, подключение внешних датчиков и пр. Пользователи могут обращаться к камере посредством стандартного web браузера.
- *мультимедиа компьютер, сканер, клавиатура.* В основе - светочувствительный сенсор - это своего рода сердце любой цифровой камеры. Именно он позволяет преобразовывать свет в электрические сигналы, доступные для дальнейшей электронной обработки. Основной принцип действия и ПЗС - и КМОП-сенсоров одинаков: под воздействием света в полупроводниковых материалах рождаются носители заряда, которые впоследствии преобразуются в напряжение. Различие между ПЗС - и КМОП-сенсорами заключается, прежде всего, в способе накопления и передачи заряда, а также в технологии преобразования его в аналоговое напряжение. Не вдаваясь в подробности конструкции различных типов сенсоров, отметим лишь, что КМОП-сенсоры являются значительно более дешевыми в производстве, но и более «шумными». Принцип работы Web-камеры схож с принципом работы любой цифровой камеры или фотоаппарата. Кроме оптического объектива и светочувствительного ПЗС - или КМОП-сенсора обязательным является наличие аналого-цифрового преобразователя (АЦП), основное назначение которого -- преобразовывать аналоговые сигналы светочувствительного сенсора, то есть напряжение в цифровой код. Кроме того, необходима система цветоформирования. Другим важным элементом камеры является схема, отвечающая за компрессию данных и подготовку к передаче в нужном формате. В Web-камерах видеоданные передаются в компьютер по USB-интерфейсу, то есть заключительной схемой камеры должен быть контроллер USB-интерфейса.

Сканер (англ. scanner) - устройство, которое, анализируя какой-либо объект (обычно изображение, текст), создаёт цифровую копию изображения объекта. Процесс получения этой копии называется сканированием. В большинстве сканеров для преобразования изображения в цифровую форму применяются светочувствительные элементы на основе приборов с зарядовой связью (ПЗС). По способу перемещения считывающей головки и изображения относительно друг друга сканеры подразделяются на ручные (англ. Handheld), рулонные (англ. Sheet-Feed), планшетные (англ. Flatbed) и проекционные. Разновидностью проекционных сканеров являются слайд-сканеры, предназначенные для сканирования фотопленок. В высококачественной полиграфии используются барабанные сканеры, в которых в качестве светочувствительного элемента используется фотоэлектронный умножитель (ФЭУ). Принцип работы однопроходного планшетного сканера состоит в том, что вдоль сканируемого изображения, расположенного на прозрачном неподвижном стекле, движется сканирующая каретка с источником света. Отраженный свет через оптическую систему сканера (состоящую из объектива и зеркал или призмы) попадает на три расположенных параллельно друг другу фоточувствительных полупроводниковых элемента на основе ПЗС, каждый из которых принимает информацию о компонентах изображения.

Динамический диапазон сканера - это показатель технических возможностей сканеров, характеризующий интервал оптических плотностей, который воспринимается сканером. Основной характеристикой любого оригинала является его оптическая плотность, определяющаяся способностью оригинала отражать или пропускать свет. Оптическая плотность лежит в пределах от 0, что соответствует белому цвету, до 4, что соответствует черному цвету и обозначается OD (Optical Density) или просто D. Динамический диапазон (Dynamic Range), или диапазон плотности (Density Range), определяется как разница между самым светлым (Dmin) и самым темным (Dmax) участками оригинала и зависит от типа оригинала и его происхождения. Применительно к сканеру, динамический диапазон определяется как разница между самым светлым (Dmin) и самым темным (Dmax) участками оригинала,

которые сканер в состоянии обработать. С увеличением динамического диапазона сканера возрастает количество вводимых градаций яркости и, следовательно, плавность переходов в смежных тонах изображения. Недостаточный динамический диапазон сканера может привести искажениям цветопередачи при сканировании изображений, содержащих плавные тоновые переходы (переходы яркости), наподобие фотоснимков голубого неба, заката, или к потере деталей в снимках светлых и темных предметов: цветов, белой одежды, облаков, “лунной дорожки” тень от здания и т.д. Напротив, сканер, имеющий высокий показатель динамического диапазона передает оригинал настолько "объемно", что, к примеру, отсканированные со слайда облака, кажется, движутся по экрану.

Мультимедийная компьютерная клавиатура, способная управлять громкостью звука и сетевым поведением компьютера. Многие современные компьютерные клавиатуры, помимо стандартного набора из ста четырёх клавиш, снабжаются дополнительными клавишами (как правило, другого размера и формы), которые предназначены для упрощённого управления некоторыми основными функциями компьютера:

· управление громкостью звука: громче, тише, включить или выключить звук;

· управление лотком в приводе для компакт-дисков: извлечь диск, принять диск;

· управление аудиопроигрывателем: играть, поставить на паузу, остановить воспроизведение, промотать аудиозапись вперёд или назад, перейти к следующей или предыдущей аудиозаписи;

· управление сетевыми возможностями компьютера: открыть почтовую программу, открыть браузер, показать домашнюю страницу, двигаться вперёд или назад по истории посещённых страниц, открыть поисковую систему;

· управление наиболее популярными программами: открыть калькулятор, открыть файловый менеджер;

· управление состоянием окон операционной системы: свернуть окно, закрыть окно, перейти к следующему или к предыдущему окну;

· управление состоянием компьютера: перевести в ждущий режим, перевести в спящий режим, пробудить компьютер, выключить компьютер.

Так как многие из этих функций (управление звуком и воспроизведением звукозаписей, управление компакт-дисками и т. п.) относятся к сфере мультимедиа, то такие клавиатуры часто называются «мультимедийными клавиатурами».

Виртуальная лазерная клавиатура. Идея реализации виртуальной клавиатуры без проводов и кнопок родилась несколько лет назад в стенах израильской компании Developer VKB Inc. Представленная на выставке CeBIT 2002 компанией Siemens Procurement Logistics Services первая виртуальная клавиатура без единого механического или электрического элемента стала первой практической реализацией этой идеи. Разработчики лазерного интерфейса виртуальной клавиатуры предполагали, что их разработка на практике может быть интегрирована в любое мобильное устройство - телефон, ноутбук, планшетный ПК и даже в стерильное медицинское оборудование. Принцип работы виртуальной лазерной клавиатуры прост и понятен без долгих объяснений. В конструкции используется два полупроводниковых диодных лазера - "красный" для создания проекции клавиатуры и невидимый инфракрасный с фотодетектором ИК-излучения для определения клавиши, к которой прикоснулся ваш палец. Пока вы непринужденно набираете текст по лазерной проекции клавиш - как на обычной клавиатуре, невидимый луч анализирует координаты положения пальцев и обрабатывает полученную информацию соответствующим образом. Добавляем к этой конструкции беспроводной интерфейс Bluetooth - и виртуальная клавиатура для любых типов стационарных и мобильных устройств - ПК, ноутбуков, карманных ПК или смартфонов, готова.

The Orbitouch. Данный агрегат выглядит как порождение злобного инопланетного разума, однако, на самом деле это тоже всего лишь клавиатура. Сразу же возникает вопрос - как же с ней работать? Ну, объяснить достаточно просто - выступы вращаются, а буквы набираются в соответствии с тем в каких позициях они стоят. Для каждой "ручки" есть восемь позиций, так что на числа и

буквы должно хватить. В клавиатуре есть встроенная мышь, так что тут покупатель может даже сэкономить.

Компьютерный руль. Компьютерный руль - игровой контроллер, имитирующий автомобильный руль. Применяется для игры в компьютерные игры - автосимуляторы. Помимо рулевого колеса и двух (трёх) педалей, в компьютерном руле могут быть такие органы управления. В медицине используется для изучения реакции человека как компонента эргатической системы. Компьютерный руль является потомком джойстика; первые рули действительно эмулировали двухосный джойстик. Существуют два рудимента того времени. Первый джойстик-руль для компьютерных игр появился в 1983 году. Это была обычная пластиковая коробка с баранкой диаметром 17 см и единственной гладкой педалью. Далее производители начали развивать идею. Постепенно они пришли к выводу, что какая-никакая отдача позволила бы игрокам прочувствовать дорогу намного лучше, да и интерактивность поднялась бы на несколько ступеней. Самый простой способ достижения подобного эффекта - установка вибромотора. Предположим, вы заехали одним колесом на обочину - игра посылает сигнал на джойстик, вибромотор начинает легко потряхивать баранку. Зачастую в прайсах модели с вибромотором обозначают как рули с обратной связью или Rumble Feedback.

Проектор. Проектор - световой прибор, перераспределяющий свет лампы с концентрацией светового потока на поверхности малого размера или в малом объёме. Проекторы являются в основном оптико-механическими или оптическо-цифровыми приборами, позволяющими при помощи источника света проецировать изображения объектов на поверхность, расположенную вне прибора - экран. Появление проекционных аппаратов обусловило возникновение кинематографа, относящегося к проекционному искусству. Виды проекционных приборов:

· Диаскопический проекционный аппарат - изображения создаются при помощи лучей света, проходящих через светопроницаемый носитель с изображением. Это самый распространённый вид проекционных аппаратов. К ним относят такие приборы как:

кинопроектор, диапроектор, фотоувеличитель, проекционный фонарь, кодоскоп и др.

· Эпископический проекционный аппарат - создаёт изображения непрозрачных предметов путём проецирования отраженных лучей света. К ним относятся эпископы, мегаскоп.

· Эпидиаскопический проекционный аппарат - формирует на экране комбинированные изображения как прозрачных, так и непрозрачных объектов.

Мультимедийный проектор (также используется термин «Цифровой проектор») - с появлением и развитием цифровых технологий это наименование получили два, вообще говоря, различных класса устройств. На вход устройства подаётся видеосигнал в реальном времени (аналоговый или цифровой). Устройство проецирует изображение на экран. Возможно, при этом наличие звукового канала. Устройство получает на отдельном или встроенном в устройство носителе или из локальной сети файл или совокупность файлов («слайдшоу») - массив цифровой информации. Декодирует его и проецирует видеоизображение на экран, возможно, воспроизводя при этом и звук. Фактически, является сочетанием в одном устройстве мультимедийного проигрывателя и собственно проектора. Название «цифровой проектор» связано, прежде всего, с обычным ныне применением в таких проекторах цифровых технологий обработки информации и формирования изображения.

Лазерный проектор - выводит изображение с помощью луча лазера.

Области применения мультимедийных средств охватывают различные виды интеллектуальной деятельности: науку и технику, образование, культуру, бизнес, а также применяются в среде обслуживания при создании электронных гидов с погружением в реальную среду, мультитеках.

Одной из основных сфер применения систем мультимедиа является образование в широком смысле слова, включая и такие направления как видеоэнциклопедии, интерактивные путеводители, тренажеры, ситуационно-ролевые игры и др. Компьютер, снабженный платой мультимедиа, немедленно становится универсальным обучающим или информационным

инструментом по практически любой отрасли знания и человеческой деятельности. Очень большие перспективы перед мультимедиа в медицине: базы знаний, методики операций, каталоги лекарств и т.п. В сфере бизнеса фирма по продаже недвижимости уже используют технологию мультимедиа для создания каталогов продаваемых домов - покупатель может увидеть на экране дом в разных ракурсах, совершить интерактивную видеопрогулку по всем помещениям, ознакомиться с планами и чертежами. Технологические мультимедиа пользуется большим вниманием военных: так, Пентагон реализует программу перенесения на интерактивные видеодиски всей технической, эксплуатационной и учебной документации по всем системам вооружений, создания и массового использования тренажеров на основе таких дисков.

Весьма перспективными выглядят работы по внедрению элементов искусственного интеллекта в системе мультимедиа. Они обладают способностью "чувствовать" среду общения, адаптироваться к ней и оптимизировать процесс общения с пользователем; они подстраиваются под читателей, анализируют круг их интересов, помнят вопросы, вызывающие затруднения, и могут сами предложить дополнительную или разъясняющую информацию. Системы, понимающие естественный язык, распознаватели речи еще более расширяют диапазон взаимодействия с компьютером.

Еще одна быстро развивающаяся, совершенно уже фантастическая для нас область применения компьютеров, в которой важную роль играет технология мультимедиа - это системы виртуальной, или альтернативной реальности, а также близкие к ним системы "телеприсутствия". С помощью специального оборудования - система с двумя миниатюрными стереодисплеями, квадранаушниками, специальных сенсорных перчаток и даже костюма вы можете "войти" в сгенерированный или смоделированный компьютером мир, (а не заглянуть в него через плоское окошко дисплея) повернув голову, посмотреть налево или направо, пройти дальше, протянув руку вперед - и увидеть ее в этом виртуальном мире; можно даже взять какой либо виртуальный предмет (почувствовав при этом его тяжесть) и

переставить в другое место; можно таким образом строить, создавать этот мир изнутри.

Порядок выполнения работы

1. Изучите теоретический материал.
2. По таблице данных результатов анализа крови больных для каждой представленной характеристики постройте в Excel диаграммы различных типов и графики в разных формах (с моделями линий тренда).
3. По моделям, отражающим популяционные изменения и «хищник-жертва» постройте графики функций, изменяющиеся во времени (для модели «хищник-жертва» - график, отражающий зависимости численностей «жертвы» и «хищника»).
4. Изучите графическое отражение динамики игры «Жизнь» по электронной версии и зафиксируйте несколько «скрин-шотов» (см., например, http://www.nature.air.ru/models/models.htm).
5. По данным приложения составьте в Excel отчетную таблицу, содержащую: заголовок, наименование регистрируемых характеристик, среднее значение по каждой характеристике, диаграммы (полученные в п.2 – выборочно), выделение цветом и буквами строк, у которых большинство показателей выше средних значений (буквы формируются с помощью оператора выполнения условия. Постарайтесь совместно с цветовой гаммой использовать «смайлики».
6. Осуществите представление динамически изменяющейся картины («слайдшоу»), отражающей функционирование любого органа человека (применять любые мультимедийные средства и графические редакторы – с указанием задействованных команд). Разрешается использовать сменяющиеся слайды.
7. Оформите отчет, включающий результаты выполнения работы и краткие ответы на контрольные вопросы.
8. Составьте аннотацию не менее 2 информационных источников, найденных в сети интернет по тематике лабораторной работы (объемом 150-350 слов).

Контрольные вопросы.

1. Что такое графическое представление информации (числовой и семантической)?

2. Какие мультимедийные средства применяются в операционной?
3. Какие мультимедийные средства применяются в системах прикроватного мониторинга?
4. Для чего предназначены «смайлики»?
5. Какие графики можно отображать в различных пакетах статистической обработки?
6. Какие диаграммы можно отображать в различных пакетах статистической обработки?
7. Как осуществляется приведение реальных графических изображений в окнах определенного размера?
8. Каким образом отображаются трехмерные объекты на плоскостном экране монитора?
9. Каким образом осуществляется «слайдшоу»?
10. Каким образом используются графические изображения при имитационном моделировании?

Библиография

1. Биологические имитационные модели /URL: http: //www.nature.air.ru/models/models.htm.
2. Биометрика /URL: http://www.biometrica.tomsk.ru
3. Диаграммы. /URL: http://fb.ru/article/71308/vidyi-diagramm-i-ih-osobennosti
4. Инженерная 3D-компьютерная графика [Текст] : учебное пособие для бакалавров / под ред. А. Л. Хейфеца ; Министерство образования и науки Российской Федерации, Южно-Уральский государственный университет, 2012. - 464 с.
5. Разработка компонента визуализации биомедицинских данных на основе технологии OPENGL /URL: http: //ilab.xmedtest.net/?q=node/5870
6. Современное программное обеспечение для статистической обработки биомедицинских исследований /URL: http: //www.disser.ru/library/31/440.htm

Приложение 1

Пример математических моделей биологических систем

1. Игра «Жизнь».

Многие процессы, интересующих медиков (например, динамика возрастного состава в регионе) может быть с определенной точностью описано математической моделью. Самая простая модель известна под названием «Игра Жизнь».

Предполагается наличие прямоугольного клетчатого поля, в каждой клетке которого может «жить» существо. Если клетка пустая – то в ней никто не живет. Модель задается двумя параметрами: начальной конфигурацией (размером поля и расположением живых существ) и определенными «биологическими законами», регулирующими жизнь популяции существ. На каждой итерации осуществляется последовательный просмотр всех клеток с некоторой начальной (координаты клетки выбираются исследователем или случайным обрзом) с применением к ним биологических законов.

В качестве типовых законов, предлагаются, например, следующие:

1. Если выбранная клетка пуста, а в соседней с ней клетках находится более двух существ, то внутри клетки появляется существо («размножение»).
2. Если выбранная клетка непуста, а в соседних с ней клетках живет меньше трех или больше четырех существ, то клетка очищается (существо в ней погибает от одиночества или перенаселения).
3. Если правила 1 и 2 не выполняются, то ничего с клеткой не происходит.

(Под соседними подразумеваются восемь окружающих клеток, за границей ореала – прямоугольного поля – существ нет).

2. Модель «хищник» - «жертва».

В системе хищник-жертва ситуация моделируется следующим образом. В случае конкурирующих популяций исчезновение одной означает выигрыш для другой в борьбе за дополнительные ресурсы. Обозначим через C численность популяции хищника, N – популяцию жертвы. Наиболее популярная модель, отражающая колебания численности имеет вид:

$$N_{i+1} = N_i + (r \cdot N_i - a \cdot N_i \cdot C_i) \cdot \Delta t,$$
$$C_{i+1} = C_i + (-q \cdot C_i - a \cdot f \cdot N_i \cdot C_i) \cdot \Delta t.$$

Согласно первому уравнению при C=0 численность жертв быстро растет со скоростью r , поскольку модель не учитывает внутривидовой конкуренции. Скорость роста числа жертв ($\frac{\Delta N}{\Delta t}$) уменьшается тем больше, чем чаще происходят встречи особей видов (тогда a - коэффициент эффективности поиска).

Второе уравнение показывает, что в отсутствии жертв численность хищников быстро убывает со скоростью q: положительное слагаемое в правой части уравнения компенсирует эту убыль, f – коэффициент эффективности перехода пищи к потомству хищников.

3. Внутривидовая конкуренция в популяции с дискретным размножением.

Для популяций с дискретным размножением (некоторые виды растений, насекомые) поколения дифференцирпованно разнесены во времени и особи разныз покалений вместе не сосуществуют. Численность подобной популяции характеризуется числом N_t , а время t- дискретная величина – можно, в первом приближении, считать номером популяции. Тогда одна из моделей межвидовой конкуренции может быть описана уравнением:

$$N_{t+1} = \frac{N_t \cdot R}{1 + (a * N_t)^b}$$

где R – скорость воспроизводства попляции в отсутствии внутривидовой конкуренции (математически это соответствует a=0); a – параметр, характеризующий интенсивность внутривидовой конкуренции, при b=1 осуществляется выход численности популяции на стационарное значение при любых значениях других параметров модели.

Знаменатель в уравнении отражает наличие конкуренции, делающей скорость роста тем меньше, чем больше численность популяции. Ланная модель описывает четыре вида эволюции:

1. монотонное установление стационарной численности популяции;
2. колебательное установление стационарной численности популяции;
3. устойчивые предельные циклы изменения численности популяций;
4. случайные изменения численности популяции без наличия явных закономерностей.

4. Внутривидовая конкуренция в популяции с непрерывным размножением.

В данном случае численность популяции N(t) является непрерывной функцией во времени. В начале эволюционного процесса численность популяции невилика, а ее удельная скорость не зависит от численности: $\frac{1}{N} \cdot \frac{\Delta N}{\Delta t} = r$ - скорость роста численности популяции в отсутствии конкуренции. Далее, по мере роста численности, скорость роста начинает уменьшаться и при достижении определенного критического значения К обращаетя в ноль. Таким образом, в первом приближении, математическая модель имеет вид:

$N_{i+1} = N_i + r \cdot N_i \cdot (\frac{K - N_i}{K}) \cdot \Delta t$.

Приложение 2

Показатели крови пациентов отделения гастроэнтрологии.

эритроциты	гемоглобин	Цветовой показатель	лейкоциты	эозофилы	палочкоядерные	лимфоциты	моноциты
4,6	142	0,93	7,05	1	14	38	2
4,2	115	0,82	7,8	2	4	10	2
3,4	107	0,94	8	1	6	27	2
3	87	0,87	7,3	4	1	36	4
3,4	100	0,88	6,15	0	4	17	2
5,0	170	1,02	7,1	2	5	30	4
5	150	0,90	3,7	6	1	21	1
3,6	105	0,88	17,6	0	7	26	2
5,3	108	0,61	4,2	4	1	46	2
4,16	132	0,95	5,1	2	4	47	4
4,3	145	1,01	4,55	4	5	36	3
4,0	128	0,96	7,6	4	2	29	7
5,0	160	0,96	6,8	0	4	23	1
4,46	146	0,98	6,5	4	1	27	2
5,4	176	0,98	7,8	2	2	29	0
3,1	108	1,05	3,7	1	2	31	4
4,0	124	0,93	4,4	2	7	22	3
5,1	164	0,96	7,9	0	5	60	6
4,25	138	0,97	6,2	2	7	28	4
4,8	160	1,00	6,7	2	2	25	1
4,3	150	1,05	12	1	1	30	9
4,0	113	0,85	5,85	6	4	37	5
4,8	153	0,96	7,1	6	1	30	2
3,6	114	0,95	13,1	0	2	24	1
3,35	103	0,92	9,1	1	3	27	17

ЛАБОРАТОРНАЯ РАБОТА № 3. РАСЧЕТ КРИТЕРИЕВ КАЧЕСТВА ДИАГНОСТИЧЕСКОГО ПРОЦЕССА

Цель работы. Овладение навыками оценки вычисления типовых критериев качества классификации биологических объектов.

Краткие теоретические сведения.

В биомедицинских исследованиях основополагающую роль является правильное соотнесение биообъекта или процесса к определенному классу. Указанное соотнесение называется классификацией или диагностикой.

В конечном итоге диагностика осуществляется путем применения определенных решающих правил. В случае возможности формализации синтезируются базы знаний соответствующих автоматизированных систем поддержки принятия решений (АСППР).

АСППР диагностики состояния биообъекта находят все большее применение в медицинской практике на различных этапах лечебно-диагностического процесса, профилактике и скрининге заболеваний.

Для определения классификационных (диагностических) способностей системы поддержки принятия решений на экзаменационной выборке с заранее определенными кластерами (так называемо, «обучение с учителем») осуществляется сбор статистического материала, по которому определяются показатели качества диагностической системы – применения заложенных в базе знаний СППР решающих правил.

Диагностика определенной нозологической группы проводится либо относительно некоторого класса условно здоровых людей (класс А) и класса, характерного для определенной нозологии (класс Б). (В случае дифференциальной диагностики осуществляется сравнение между различными нозологиями – в этом случае под «классом А» понимается одна из нозологий, под «классом Б» - другая).

По результатам тестирования составляется таблица вида:

Таблица Распределения результатов диагностики

Обследуемые	«золотой стандарт» (истина)		Всего
	Болен	здоров	
Болен (положительный результат теста)	Истинно-положительный результат a	Ложно-положительный результат b	a+b
Здоров (отрицательный результат теста)	Ложно-отрицательный результат c	Истинно-отрицательный результат d	c+d
Всего	a+c	b+d	a+b+c+d

В качестве показателей качества, характеризующих статистическую достоверность медицинской экспертной системы, при первичной оценки качества ее работы, выбираются: диагностическая чувствительность (ДЧ), диагностическая специфичность (ДС), прогностическая значимость положительных ($ПЗ^{+}$) и отрицательных ($ПЗ^{-}$) результатов испытаний и диагностическая эффективность (ДЭ).

Указанные показатели качества рассчитываются в соответствии со следующими выражениями:

$$ДЧ = \frac{a}{a+c}\ ДС = \frac{d}{d+b}\ ДЭ = \frac{a+d}{a+b+c+d}\ ПЗ^{+} = \frac{a}{a+b}\ ПЗ^{-} = \frac{d}{c+d}$$

где: a – истинно положительный результат равный количеству пациентов из класса заболеваний Б правильно классифицируемых;

b – ложноположительный результат равный количеству относительно здоровых людей класса А ошибочно отнесенных экспертной системой к классу Б;

c – ложноотрицательный результат равный количеству людей из класса Б отнесенных экспертной системой к классу А;

d – истинно отрицательный результат равный количеству людей из класса А правильно классифицированных экспертной системой.

Заметим, что соотношение $\frac{a+b}{c+d}$ существенно влияет на значения показателей качества.

Порядок выполнения работы

1. Изучить теоретический материал.
2. Согласно данным приложения №2 лабораторной работы № 2 синтезировать диагностическое решающее правила вида: «ЕСЛИ большее количество значений показателей входи в доверительный интервал, ТО пациент принадлежит к данному классу, В ПРОТИВНОМ СЛУЧАЕ он принадлежит другому классу». Рассчитайте показатели качества. Сделайте выводы.
3. Изменяя значения a, b, c, d случайным образом в диапазоне [30, 60] порядка 20-30 раз, определите значения критериев качества и постройте и проанализируйте графики зависимости указанных показателей от величины $\frac{a+b}{c+d}$. (построение осуществите на одной плоскости). Сделайте выводы.
4. Оформите отчет, включающий результаты выполнения работы и краткие ответы на контрольные вопросы.
5. Составьте аннотацию не менее 2 информационных источников, найденных в сети интернет по тематике лабораторной работы (объемом 150-350 слов).

Примечания: для расчетов и их графической интерпретации используйте любые офисные средства.

Контрольные вопросы

1. Что такое диагностический процесс?
2. Какие решающие правила применяются в автоматизированных системах поддержки принятия решений?
3. Как определяются критерии качества диагностического процесса?
4. В чем заключается семантическая нагрузка показателей качества диагностического процесса?
5. Каким образом значения показателей качества связаны с ошибками первого и второго рода?

Библиография

1. Доказательная медицина для всех /URL: http://medspecial.ru/
2. Доказательная медицина. Презентация /URL: http://www.myshared.ru/slide/181227/

3. Лекции по медицинской информатике. /URL: http://kingmed.info/lektsii/Meditsinskaya_informatika_i_biostatistika/lecture_772/Dokazatelnaya_meditsina__alternativa_meditsine_mneniy
4. Разработка критериев оценки качества оказания медицинской помощи в дневном стационаре в педиатрии /URL: http://cyberleninka.ru/article/n/razrabotka-kriteriev-otsenki-kachestva-okazaniya-meditsinskoy-pomoschi-v-dnevnom-statsionare-v-pediatrii
5. Рассчитать диагностическую чувствительность и специфичность теста. /URL: http://allrefs.net/c49/3t4ln/p1/
6. Ступаков, И. Н. Доказательная медицина в практике руководителей всех уровней системы здравоохранения [Текст] / под ред. В. И. Стародубова. - М.:МЦФЭР, 2016. - 448 с.
7. Триша Гринхальх. Основы доказательной медицины Издательство ГЭОТАР-Медиа, 2018. – 336 .

ЛАБОРАТОРНАЯ РАБОТА №4. ПРОГНОЗИРОВАНИЕ РАЗВИТИЯ ЗАБОЛЕВАЕМОСТИ В РЕГИОНЕ

Цель работы: овладение навыками методами идентификации и анализа прогностических моделей уровня заболеваемости в регионе на основе данных многолетних наблюдений.

Краткие теоретические сведения.

Экстраполяционные методы прогнозирования основываются на концепции сохранения в будущем тенденций прошлых закономерностей с учетом текущей ситуации (закономерностей) - различные методы представлены, например, в работах.

В условиях невозможности детерменированного (или плохо формализуемого) применения временных рядов для прогнозирования и большой неопределенности объекта (нечеткости или грубости результатов мониторирования) применяются методы экспертных оценок.

Методы имитационного моделирования предполагают разработку математических или логических моделей будущего функционирования объекта, в том числе с применением лингвистических переменных. В этом случае, удачно применяют достижения искусственного интеллекта на основе: формального аппарата математической логики, индукции и дедукции, теория вероятностей и статистические методы, теория распознавания образов, теория нечетких множеств и нечеткого логического вывода, искусственные нейронные и иммунные сети, методы построения информационно-аналитических и концептуальных моделей, мягкие вычисления, теорию субъективного анализа и т.д.

В обобщенном виде схема прогнозирующей системы показана на рисунке 1.

К наиболее распространенным методам прогнозирования, применяемым в настоящее время, относятся:

1. Регрессионные модели синтезируются с использованием специальных методов подбора вида экстраполирующей функции и определения значений её параметров (http://prognoz.org/lib/metody-prognozirovaniya);

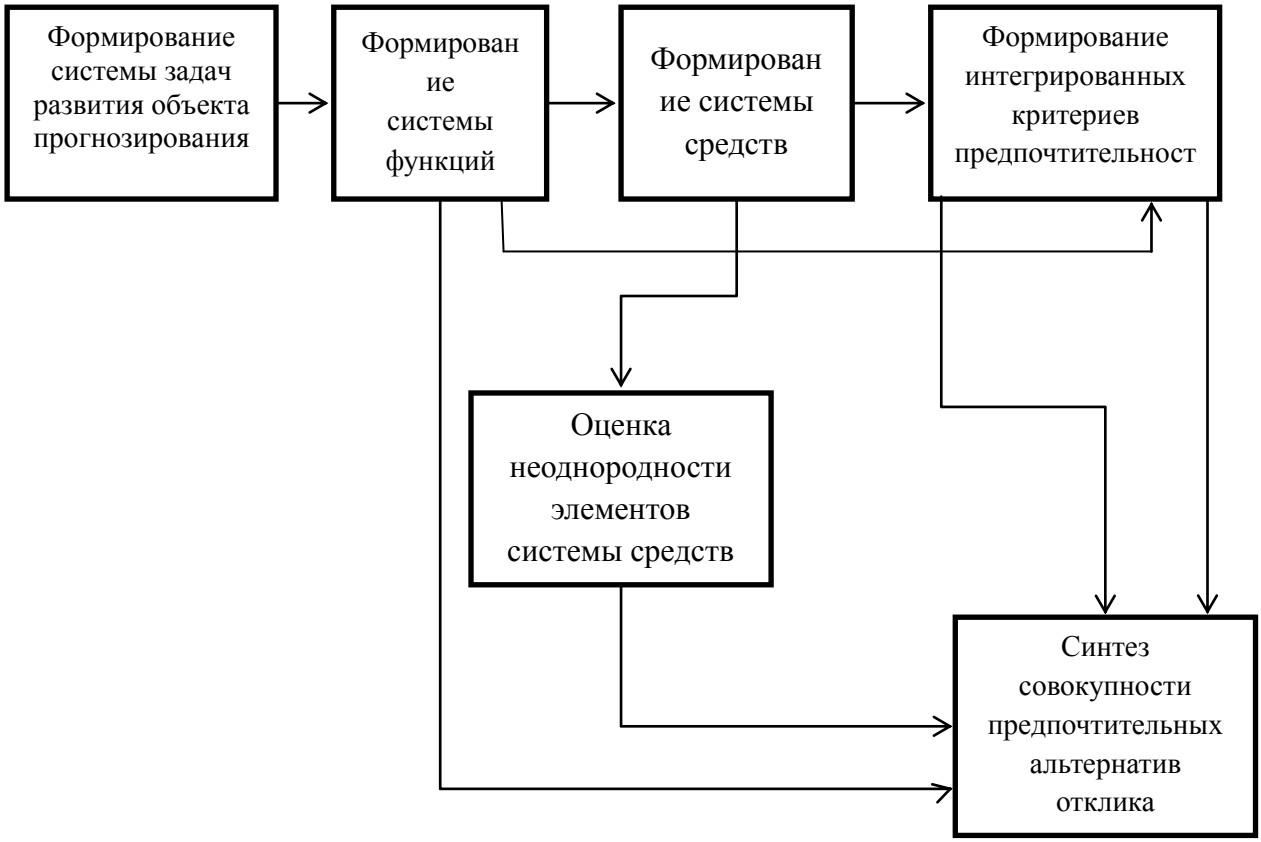


Рисунок 1 - Схема прогнозирующей системы.

2. Адаптивное сглаживание предполагает итерационный процесс пересмотра выбранных значений весовых коэффициентов: веса пересматриваются по завершении каждого прогнозного периода и селектируется те значения, при котором прогноз был бы наименее ошибочным (бизнес-учебники.рф/logist/metodyi-pronozirovaniya.html);

3. Факторный анализ позволяет на основе натурных данных экспериментального наблюдения за изменениями значений характеризующих биообъект признаков сформировать определенное (конечное) множество группу показателей (например, латентных), определяющих корреляционную взаимосвязь между признаками (http://prognoz.org/lib/faktornyi-analiz);

4. Многомерная фильтрация;

5. В процессе имитационного моделирования синтезируются структуры, описывающие объекты или процессы с определенными коэффициентами подобия (как во времени, так и в пространстве) действительности с задаваемыми ограничениями

(http://ru.wikipedia.org/wiki/Имитационное_моделирование);

6. Метод группового учета аргументов (МГУА) использует семейство различных индуктивных алгоритмов для математического моделирования мультипараметрических данных, классификации, оценки информативности, решения задач интерполяции и экстраполяции временного ряда на основе различных аппроксимационных функций. Метод основан на рекурсивном селективном отборе моделей по иному от идентификации модели критериям и иной выборке, на основе которых постепенно усложняются структуры моделей, до момента «вырождения» значений критериев селекции. (http://ru.wikipedia.org/wiki/Метод_группового_учета_аргументов);

7. Экспоненциальное сглаживание тренда – моделирование путем использования экспоненциальных функций, обладающих свойствами непрерывности производных, в окне задаваемого пользователем размера анализируемого временного ряда (http://ru.wikipedia.org/wiki/Экспоненциальное_сглаживание);

8. Спектральные методы – методы, основанные на изучении спектров излучения, поглощения и рассеивания (не путать с Фурье-анализом временного ряда);

9. Метод скользящей средней применяется для выравнивания временного ряд на основе вычисления средневзвешанных характеристик «окна» определенного размера;

10. Применение сплайн-функций предполагает идентификацию некоторых элементарных математических функций, область определения которых разбита на конечное число, как правило, равноотстоящих отрезков, на каждом из которых сплайн-аппроксимация совпадает с алгебраическим полиномом или гармонической функцией (http://ru.wikipedia.org/wiki/Сплайн);

11. Оптимальные фильтры;

12. Метод Бокса-Дженкинса предполагает прогнозирование путем применения авторегрессионных структур моделей интегрированного скользящего среднего (http://bizness-gruppa.ru/box-jenkins-metod-boksa-dzhenkinsa);

13. Метод Марковских цепей основывается на анализе последовательности случайных событий с конечным или счётным числом исходов при фиксированном настоящем зависимым от

конкретного «прошлого» (обычно с единичным временным запаздыванием (http://ru.wikipedia.org/wiki/Цепь_Маркова);

14. Модели разностных уравнений используются, как правило, для исследования динамических характеристик импульсных систем (например, стабилизаторов напряжения, цепочек импульсов, передаваемых в нейронных сетях)

15. Авторегрессионная модель временных рядов основывается на автокорреляции – т.е., на построении линейных (или нелинейных) структур с запаздыванием (зависящих от «прошлых» значений ряда с определенным временным шагом) (http://ru.wikipedia.org/wiki/Авторегрессионная_модель);

16. Вероятностный метод позволяет с приемлемой (задаваемой) точностью определить, в каких пределах будет изменяться искомая величина и-или с какой вероятностью следует ожидать наступление определенного события (значения временного ряда) (http://pictoris.ru/1/4/index.html).

В качестве статических критериев, характеризующих адекватность и приемлемость модели, предлагается применять: тип распределения, оценку надежности связи между системообразующими поведение объекта характеристиками, однородность и репрезентативность динамического ряда, оценку уровня (в определенном интервале) гармонических (или иных ритмических) волновых составляющих.

При анализе поведения системы синтезируются модели, отражающие динамику поведения каждого ее элемента во времени и связей между ними, по которым осуществляется прогноз перехода системы в определенное. Для чего используются иногда функции когерентности.

При этом, возникают проблемы оценки качества прогноза до его реализации и оценка достоверности прогноза, который еще не осуществлен (аналогично проблеме необходимости наличия постаприорной вероятности в формулах Байеса при распознавании образов.

Применяемые в настоящее время методы верификации прогноза в основном оперируют статическими процедурами оценки доверительных интервалов прогнозных значений. Ошибки

возникают в двух случаях: информационные ошибки описания объекта и ошибки применяемого метода прогнозирования.

К наиболее эффективным относят следующие принципы прогнозирования:

1. Активное управление активно при пассивном прогнозе – эффективность прогноза реализуется механизмами управления;
2. Эффективность прогноза определяется мерой ее детерминированности;
3. Эффективность прогноза функционально зависит от параметров описания объекта, системно связанных друг с другом;
4. Эффективность прогноза для различных уровней описания объекта структурирования ограничена неуправляемыми факторами развития объекта, которые невозможно учесть на начальном этапе моделирования.

Обобщаемая структура взаимосвязи методов прогнозирования с исходной информацией показана на рисунке 2.

Экстраполяция, проводимая в будущее, - это перспектива, а в прошлое, - ретроспектива. К предпосылкам использования экстраполяционной методологии относятся:

- высокая вероятность того, что развитие исследуемого явления, характеризуемого рассматриваемым временным рядом, в целом описывать некоторой плавной кривой (дифференцируемой и интегрируемой);
- общая тенденция развития явления в прошлые и текущие моменты времени с высокой степенью вероятности не изменяются в ближайшем будущем.

При этом могут использоваться различные методы анализа структур данных и моделирования временных рядов в зависимости от исходной информации.

Упрощенные приемы целесообразны при недостаточной информации о предыстории развития явления (нет достаточно «длинного» ряда – у исследователя имеется обучающая выборка небольшого объема с точки зрения статистической репрезентативности рассматриваемого процесса или поведения объекта – эта ситуация характерна для биообъекта).

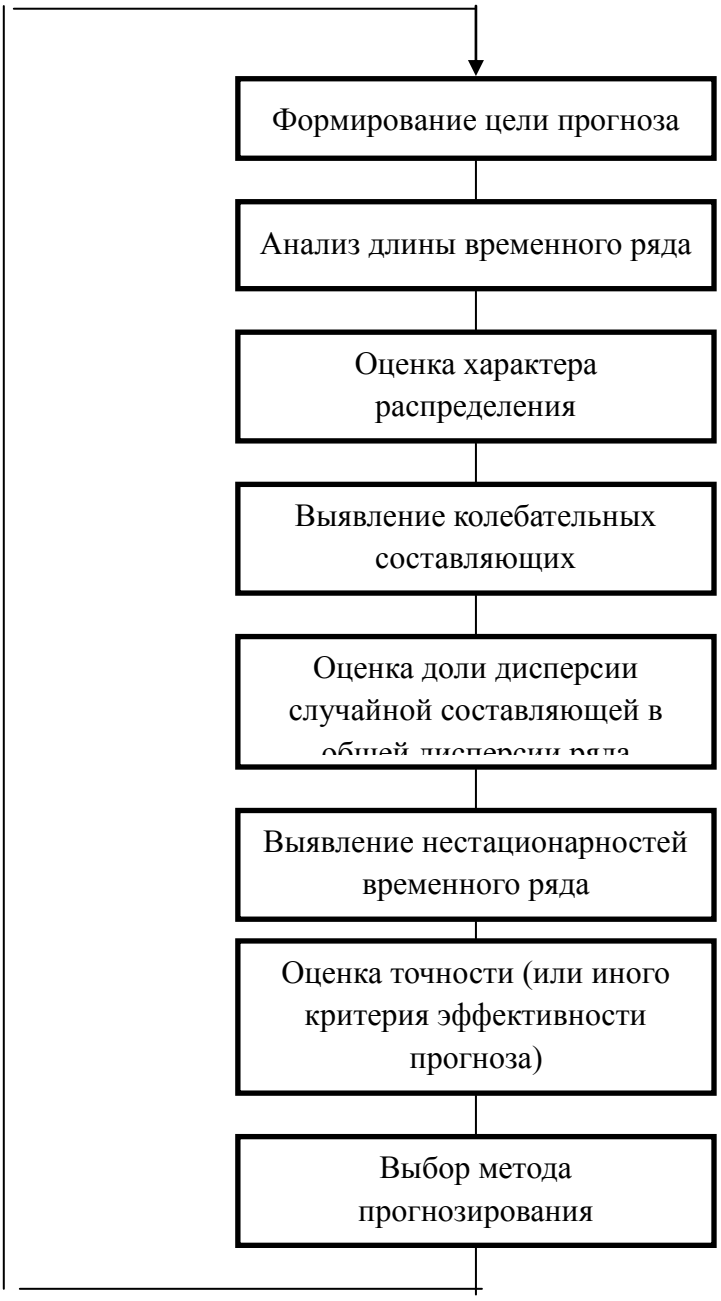


Рисунок 2 - Обобщенная структура связи исходной информации и метода прогнозирования.

Применяемы в этом случае, упрощенные приемы математического моделирования основываются на усредненных значениях показателей рассматриваемого динамического ряда. Например, для эконометрических показателей (подобными во многом показателям биообъектов, если в качестве последних рассматриваются группы людей, заболеваемость и т.п. в регионе), выделяются:

1. *Метод среднего абсолютного прироста*. Для нахождения искомого аналитического выражения, определяющего тенденцию на любое время (дату), определяется средний абсолютный прирост и осуществляется его накопление (последовательное прибавление его значения к последнему уровню ряда столько раз, на сколько периодов осуществляется экстраполяция ряда). Применение в экстраполяции среднего абсолютного прироста предполагает, что развитие явления происходит по арифметической прогрессии и относится в прогнозировании к классу «наивных» моделей, ибо чаше всего развитие явления следует по иному пути, чем арифметическая прогрессия. Поэтому данный метод относится к процедурам «разведочного» анализа и находит применение в качестве предварительного прогноза, когда у модельера нет достаточно репрезентативного динамического ряда (например, имеется только информация в начале и конце рассматриваемого периода (например, данные на начало и окончание года).

2. *Метод среднего темпа роста.* Данный метод осуществляется в случае, когда общая тенденция характеризуется показательной кривой (т.е. степенной или экспоненциальной или логарифмической).

3. *Выравнивание рядов по определенной аналитической формуле*. Определение формулы в данном случае обычно осуществляется с помощью экспертного анализа.

4. *Аналитическое сглаживание* позволяет определить общую тенденцию изменения явления на определенном временном отрезке и выполнить расчеты для периодов, в отношении которых нет исходных данных. Характерной особенностью данного метода является неглубокая величина прогноза (достаточно хорошо он работает при краткосрочном прогнозе).

5. *Адаптивные методы* используются в условиях большой колебательности значений временного ряда и позволяют при изучении тенденции учитывать степень влияния предыдущих уровней на последующие значения. К адаптивным методам относятся, например, алгоритмы, реализующие ранее упомянутые процедуры скользящих и экспоненциальных средних, метод гармонических весов, методы авторегрессионных преобразований.

Основная цель адаптивных методов заключается в структурно-параметрической идентификации самонастраивающихся моделей, способных учитывать информационную ценность (веса) различных членов (термов) математических структур моделей временного ряда и осуществлять оценки с приемлемой точности прогнозируемым значениям ряда.

6. Особенность метода *экспоненциального сглаживания* заключается в том, что в процедуре выравнивания каждого наблюдения используется только значения предыдущих уравнений, взятых с определенным весом. Смысл экспоненциальных средних состоит в нахождении таких средних, в которых влияние прошлых наблюдений затухает по мере удаления от момента, для которого определяется средние («прошлое забывается по мере отдаления от настоящего»). Главный недостаток, который проявляется особенно при анализе биомедицинских сигналов, это не учет ритмических составляющих, присущих большинству саморганизующих объектов и-или процессов в силу наличия у них автоколебаний.

Любой статистический прогноз носит приближенный характер, поэтому целесообразно определение доверительных интервалов прогноза.

Для временных рядов главный интерес представляет описание или моделирование их структуры. Цель таких исследований, как правило, шире моделирования, хотя некоторую информацию можно получить и непосредственно из модели, делая выводы о выполнении определенных физических или биологических законов и проверяя различные гипотезы. Построенные модели могут использоваться для статистического моделирования длинных рядов наблюдений при исследовании больших систем, для которых временной ряд рассматривается как входная информация.

Основной тенденцией развития процесса или поведения объекта называется плавное и устойчивое изменение уровня характеризующего наблюдаемого показателя во времени, свободное от случайных колебаний. В этом случае задача состоит в выявлении общей тенденции в изменении уровней ряда, освобожденной от действия различных факторов.

Изучение тренда включает два основных этапа:

- ряд динамики проверяется на наличие тренда;
- производится выравнивание временного ряда и непосредственно выделение тренда с экстраполяцией полученных результатов.

С этой целью временные ряды обрабатываются методами укрупнение интервалов, скользящей средней и-или аналитического выравнивания:

1. *Метод укрупнения интервалов.* Одним из наиболее простых способов изучения общей тенденции временного ряда является укрупнение интервалов - он основывается на укрупнении периодов, к которым относятся уровни временного ряда динамики.

2. *Метод скользящей средней.* Выявление общей тенденции ряда динамики осуществляется путем сглаживания значений временного с помощью скользящей средней (рассматривался ранее). Скользящая средняя – это «подвижное» значение некоторой средней динамической величины, рассчитываемой по массиву элементов временного ряда путем последовательного передвижения, как правило, на один временной интервал. То есть, первоначально вычисляется некоторый средний (или медианный) уровень из определенного числа первых по порядку уровней ряда, затем - средний уровень из такого же числа членов, начиная со второго. Таким образом, «скользящая средняя» как бы передвигается по вектору значений временного ряда от начала к концу, последовательно раз отбрасывая один уровень в начале и добавляя один в следующий. При этом посредством осреднения эмпирических данных индивидуальные колебания погашаются, и общая тенденция развития явления выражается в виде определенной плавной линии.

«Скользящая средняя» обладает достаточной гибкостью. Однако, существенным недостатком метода является укорачивание сглаженного ряда по сравнению с фактическим, что ведет к потере информации. В связи с этим, «скользящая средняя» не позволяет получить аналитического выражения для тренда и, следовательно, осуществлять имитационное моделирование.

Период скользящей может быть четным и нечетным. Практически удобнее использовать нечетный период, так как в

этом случае скользящая средняя будет отнесена к середине периода скольжения. Полученные средние соотносятся к соответствующему срединному интервалу.

Особенность сглаживания по четному числу уровней состоит в том, что каждая из численных (например, четырехчленных) средних относится к соответствующим промежуткам между смежными периодами. Для получения значений сглаженных уровней соответствующих периодов необходимо произвести центрирование расчетных средних.

Недостатком способа сглаживания рядов динамики является то, что полученные средние не дает теоретических рядов, в основе которых лежала бы математически выраженная закономерность.

3. *Метод аналитического выравнивания.* Более совершенным приемом изучения общей тенденции в рядах динамики является аналитическое выравнивание. При изучении общей тенденции методом аналитического выравнивания исходят из того, что изменения уровней ряда динамики могут быть с той или иной степенью точности приближения выражены определенными математическими функциями. Вид уравнения определяется характером динамики развития конкретного явления. Логический анализ при выборе вида уравнения может быть основан на рассчитанных показателях динамики, а именно:

- если относительно стабильны абсолютные приросты (первые разности уровней приблизительно равны), сглаживание может быть выполнено по прямой;
- если абсолютные приросты равномерно увеличиваются (вторые разности уровней приблизительно равны), можно принять параболу второго порядка;
- при ускоренно возрастающих или замедляющихся абсолютных приростах - параболу третьего порядка;
- при относительно стабильных темпах роста- показательную функцию.

Для аналитического выравнивания наиболее часто используются следующие виды трендовых моделей: прямая (линейная), парабола второго порядка, показательная (логарифмическая) кривая, гиперболическая.

Цель аналитического выравнивания - определение аналитической или графической зависимости. На практике по имеющемуся временному ряду задают вид и находят параметры функции, а затем анализируют поведение отклонений от тенденции. Чаще всего при выравнивании используются следующие зависимости; линейная, параболическая и экспоненциальная.

После выяснения характера кривой развития необходимо определить ее параметры, что можно сделать различными методами:

1. решением системы уравнений по известным уровням ряда динамики;
2. методом средних значений (линейных отклонений), который заключается в следующем: ряд расчленяется на две примерно равные части, и вводятся преобразования, чтобы сумма выровненных значений в каждой части совпала с суммой фактических значений, например, в случае выравнивания прямой линии

$$\sum(\gamma - a_0 - a_1 t) = 0; \qquad (1)$$

3. выравниванием ряда динамики с помощью метода конечных разностей;
4. методом наименьших квадратов: это некоторый прием получения оценки детерминированной компоненты $f(t)$, характеризующих тренд или ряд изучаемого явления.

Во многих случаях моделирование рядов динамики с помощью полиномов или экспоненциальной функции не дает удовлетворительных результатов, так как в рядах динамики содержатся заметные периодические колебания вокруг общей тенденции. В таких случаях следует использовать гармонический анализ.

Порядок выполнения работы.

1. Самостоятельно изучите теоретический материал.
2. Из приведенного в приложении временного ряда сформируйте последовательность, начиная с номера A1 в количестве A2 измерений. A1 определяется как остаток от деления порядкового

номера в группе на семь плюс 1. А2 определяется как утроенное значение количества букв в Вашей фамилии.

3. С помощью инструментария Excel идентифицируйте трех аппроксимантов полученных в п.2 временных трендов: лучшая по критерию детерминированности модель в режиме «построить линию тренда»; авторегрессионную модель первого порядка; гармоническую модель по двум-трем частотам (частоты определите путем подбора соответствующих периодов циклов, полученных по анализу первой производной временного тренда согласно численному дифференцированию.

4. Постройте графики временного тренда и полученных моделей на одной плоскости.

4* (повышенной сложности). Постройте модель спектра Фурье.

5. Оцените средние значения относительных ошибок аппроксимации: интерполяции, экстраполяции (до и после интерполяционного интервала. Сделайте выводы.

6. Оформите отчет, включающий в себя результаты работы (возможны скриншоты), выводы, краткие ответы на контрольные вопросы, аннотацию одного из информационных источников, указанных в библиографии или иных (найденного самостоятельно).

Контрольные вопросы:

1. Что называется интерполяцией?
2. Что называется экстраполяцией?
3. Каким образом строятся гармонические модели?
4. Охарактеризуйте виды прогнозов (сиюминутный, краткосрочный, среднесрочный, долгосрочный)?
5. Как осуществляется проверка качества прогностической модели?
6. Могут ли прогностические модели быть логическими?
7. Как осуществляется прогноз во времени и пространстве?
8. Для чего необходимо прогнозировать заболеваемость в регионе?
9. Какие заболевания населения носят ритмический характер?
10. Какие природные циклы оказывают влияние на региональную заболеваемость (и почему)?

Библиография

1. Демографические прогнозирование. Презентация. http://900igr.net/prezentatsii/geografija/Demograficheskoe-prognozirovanie/008-Vidy-i-metody-prognozirovanija.html

2. Дуброва, Т. А. Прогнозирование социально-экономических процессов [Текст]: учебное пособие / Т. А. Дуброва. - 2-е изд., испр. и доп. - М.: Маркет ДС, 2010. - 192 с.

4. Методы прогнозирования. Презентация. http://nashaucheba.ru/v24021

5. Методы прогнозирования. Презентация. http://ppt-online.org/5165

7. Статистика [Текст]: учебник для бакалавров / под ред. И. И. Елисеевой; Санкт-Петерб. гос. экон. ун-т. - 3-е изд., перераб. и доп. - Москва : Юрайт, 2014. - 558 с

Приложение

Динамика психических заболеваний в регионе

t	0	1	2	3	4	5	6	7	8	9	10
Ур-нь	19,4	24,09	27,95	32,61	36,89	40,33	44,8	48,7	50,88	52,78	54,56

11	12	13	14	15	16	17	18	19	20
56,42	58,15	58,94	60,03	60,39	58,68	57,43	56,83	55,37	53,70
21	22	23	24	25	26	27	28	29	30
51,49	48,67	45,33	40,75	36,89	34,00	29,59	24,72	20,06	14,38

ЛАБОРАТОРНАЯ РАБОТА №5. СИНТЕЗ ДИАГНОСТИЧЕСКИХ РЕШАЮЩИХ ПРАВИЛ

Цель работы: овладение навыками структурно-параметрической идентификации диагностических правил с применением информационных технологий средств вычислительной техники на основе биометрической информации.

Краткие теоретические сведения.

Под диагностическими правилами понимается процедура вывода заключения о соотнесении состояния анализируемого объекта или процесса к определенному классу или области на основании временно-пространственной регистрации существенных характеристик.

Любой объект (процесс) с точки зрения диагностики подвергается анализу со стороны исследователя, который, как правило, априори знает, какие существенные характеристики ему следует регистрировать для решения диагностической задачи. То есть, в этом случае, исследователь уже владеет набором диагностических правил, которые либо опровергают, либо подтверждают выдвинутую им рабочую гипотезу о состоянии объекта. Так как о каждом состоянии объекта может выдвигаться различное количество гипотез, то, следовательно, диагностические правила каждой из них не должны в случае объединения поглощать друг друга, и, вообще говоря, должны иметь минимальное количество пересечений как по регистрируемым параметрам, так и по диапазонам их изменений.

В общем случае диагностическое правило имеет вид, например, продукции: если значение $P=P_0$, то состояние $S=S_0$.

$$P=F(S, t, dS), \qquad (1)$$

где S - состояние; t - время; dS - диагноз изменения характеристик состояния.

Если зависимость (1) достаточно хорошо идентифицирована (с заданной степенью точности или неопределенности), то нетрудно построить эксперто-диагностическую систему продукционного типа с указанием исследователю технологии реализации необходимой информации для достаточно достоверной диагностики гипотетического состояния.

Рассмотрим **логический механизм** синтеза правила (1).

1 этап. Организация мониторинга состояния заданной глубины и полноты.

2 этап. Выделение множества ортогональных и информативных признаков с точки зрения вариативности. То есть, с одной стороны, селектируем сильно коррелированные характеристики, с другой стороны, отбираем те из них, вариативность которых (отношение дисперсии к среднему значению) выше определенного порогового уровня (например 10%).

3 этап. Кодируем состояния (лучше в двоичном коде): с учителем - то есть исследователь знает состояния, без учителя - выполняется кластер-анализ и задаются состояния или вводится пороговый принцип. Таким образом, получаем значения «логической» функции $Y=(Y_{i1},Y_{i2},Y_{il})$. Если состояний не много, то рекомендуется применять унитарное кодирование с минимизацией Хеменгового расстояния соседних состояний.

4 этап. Кодируем значение признакового пространства, следующим образом (во всех случаях рекомендуется унитарный код). По каждому оставленному признаку выделяем определенный набор состояний, как попадание значения признака в определенный диапазон. Диапазон определяется либо:

1) Экспертом, исходя из его знаний и жизненного опыта.

2) Исследователем, по анализу частоты распределений значений и личного опыта. При достаточно небольшом количестве признаков анализ гистограммы рекомендуется проводить визуально, наблюдая все признаки одновременно (в концепции системный подход).

3) Автоматически (с применением ЭВМ) по следующему алгоритму.

Исследователь задает количество состояний по каждому признаку n_i (каждое из них кодируется, желательно в унитарном коде). Определяется медиана Mo и дисперсия G_o. Определяется удельное отклонение как $Gy = \sqrt{Go/(n-1)}$. В качестве первого диапазона (состояния) выбирается величина внутри диапазона $M_o \pm G_y$. Все значения X_i попавшие в данный диапазон кодируются определенным состоянием S_o. Величина n_i декрементируется и повторяется описанный процесс над «оставшимися» данными. Так

продолжается до тех пор, пока n_i не станет равно 0 и всем оставшимся значениям будет присвоено состояние S_n. Граничные значения $M_o \pm G_y$ либо включаются в одно из состояний, либо, что более оптимально, кодируются знаком переходной функции.

5 этап. Определяем функциональные зависимости между полученными булевыми функциями (парные и множественные) и парное Хеминговое расстояние. Те признаки, у которых это расстояние равно нулю, селектируются путем оставления одного из них с наибольшей вариативностью.

Явный вид логической зависимости между булевыми переменными X_k, $k=1,m$ определяются следующим образом. На первом шаге проверяются условия независимости: поскольку каждая булева функция может иметь два значения истинности, то m булевых функций может образовывать 2^m комбинаций значений истинности. Согласно определению m-булевых функций независимы, если в совокупности при всех возможных значениях аргументов они могут принимать 2^m комбинаций значений истинности. Т.е., для проверки независимости необходимо вычислить их изображающие числа и проверить, образуют ли они полный набор чисел. Если да, то функции независимы, в противном случае - зависимы.

На втором шаге в базисе булевых функций выписывают в последовательные строки изображающие числа и определяют какие числа отсутствуют в наборе столбцов (повторяющиеся значения чисел считают один раз). Столбцы набора представляют собой комбинации значений истинности функций $X_i,\ldots, X_m$, при которых соответствующие элементарные произведения составленные из $X_i,\ldots, X_m$ истинны.

Таким образом, если идентифицируется зависимость:

$$F(X_i,\ldots,X_n)=1 \qquad (2),$$

то, следовательно, имеющиеся в наборе столбцы указывают номера тех колонок базиса в $(X_1,\ldots,X_n)$, которые совпадают с номерами изображающего числа $\#F(X_1,\ldots,X_m)$, на которых функция F истинна.

Например, пусть задан протокол мониторинга трех логических функций:

X_1 11001010
X_2 10101100
X_3 11001100

Выпишем последовательно все столбцы в этом наборе изображающих чисел как строки и укажем справа их десятичные значения:

111=7, 101=5, 010=2, 000=0, 111=7, 110=6, 001=1, 000=0

Видно, что десятичные эквиваленты 3 и 4 отсутствуют, а это означает, что по отношению и в (X_1, X_2, X_3) изображающее число связи $F(X_1, X_2, X_3) = 1$ имеет вид $\#F(X_1, X_2, X_3) = 1$.

Минимизируя полученную функцию, получаем:

$$\#F=\bar{X}1\bar{X}3+X2X3+X1\bar{X}2=1$$

Проверяем:

X1	X2	$\bar{X}3$	$\bar{X}1X3$	X2X3	$X1\bar{X}2$	F
1	1	1	0	1	0	1
1	0	1	0	0	1	1
0	1	0	1	0	0	1
0	0	0	1	0	0	1
1	1	1	0	1	0	1
0	1	1	0	1	0	1
1	0	0	0	0	1	1
0	0	0	1	0	0	1

Таким образом, определяется как логические функции связаны между собой.

6 этап. Идентифицируем логические функции *Y=F(X)* - парная зависимость и/или *Y=F({X})* (З) - множественная зависимость. Заметим, что возможен вариант отсутствия тех или иных функциональных зависимостей.

7 этап. Переходим от полученных булевских функций либо к продукционным диагностическим правилам, либо к схемотехническому решению идентификационного диагностического устройства. Однако, второй вариант менее устойчив и мобилен в случае достаточно быстрого изменения окружающей среды, приводящего к изменению в функционировании анализируемого объекта (системы, процесса), а, следовательно, и вида идентифицированных функций.

Как и во множественном регрессионном анализе, при синтезе зависимостей (3) для получения более строгого результата (минимизации пересечений понятий в диагностических, классификационных правилах каждого состояния) рекомендуется руководствоваться правилом максимальной организации (независимости) факторного пространства. Для этого необходимо добиться максимальной независимости X между собой, т.е. в идеале не должно существовать функциональных зависимостей между X_i. Т.е., если на пятом этапе идентифицируется F(x)=1 , то необходимо изменить множество X: либо путем исключения переменных (по критерию вариативности), что чревато в общем случае, потерей информации; либо изменить кодирование вводимых сигналов путем уменьшения количества состояний и/или изменения (экспертным путем) диагностических классов состояний. При достаточно мощной вычислительной технике и сравнительно небольшом размере факторного пространства (до 100 признаков) эти проблемы могут быть решены переборным путем. В противном случае, следует применять методы целенаправленного случайного поиска.

Как и в авторегрессионном анализе возможно формирование продукционных диагностических правил с учетом фактора запаздывания.

Порядок выполнения работы.

1. Самостоятельно изучите теоретический материал.
2. Согласно номеру варианта (равен порядковому номеру студента в журнале группы) N сформируйте протокол мониторинга наблюдения за процессом X, следующим образом:

Индекс переменной X	Индекс протокола Z (см. таблицу 1)
1	mod(n,8)+1
2	mod(n,8)+2
3	mod(n,8)+3
4	mod(n,8)+4

3. Выберите в качестве выходной величины, определяющей состояние процесса переменную мониторинга с индексом 4.
4. Зададитесь числом состоянием по Y- 3, X1-5, X2- 3, X3 -2.
5. Определите диапазон изменений состояний, причем для Y_1,X_1,- автоматически по дисперсии (здесь и далее рекомендуется

использовать интегрированные среды типа EXCEL, STATISTICA), X_2 - экспертным путем анализа гистограммы, X_3 - экспериментальным путем заданием одного порога (<,>=).

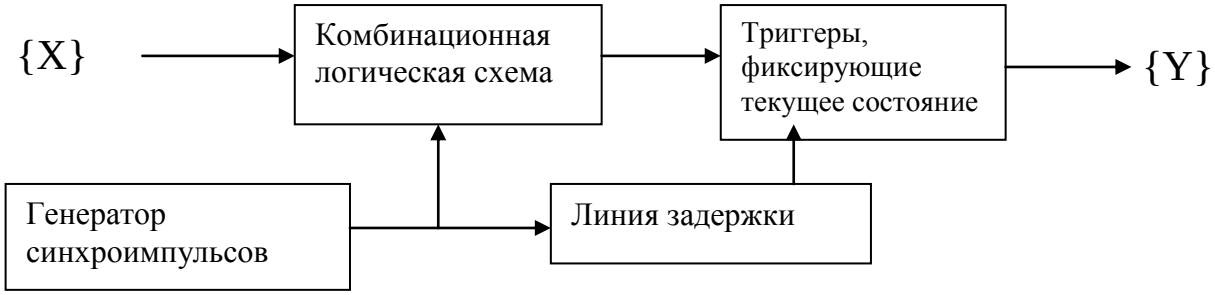
6. Закодируйте состояние по X, У - т.е. получите характеристические числа (вид) булевых функций $X_б$ и $У_б$.

7. Проанализируйте взаимозависимость между булевскими факторами X и сформируйте наиболее ортогональное факторное пространство (векторы которого в наименьшей степени зависимы между собой).

8. Идентифицируйте в минимальном виде (с помощью карт Карно) логические функции $У_б=F(X_б)$.

9. Перейдите от булевского представления к логико-семантическому и сформулируйте диагностические правила продукционного типа.

10*. Составьте схемотехническое решение диагностических правил по следующей структуре:



Рассчитайте скважность синхроимпульсов и их характер для обеспечения устойчивой работы схем, считая время срабатывания любого логического элемента 10мс, время срабатывания триггера 50мс.

11. Оформите отчет с указанием последовательности своих действий, необходимых комментариев и выводов, кратких ответов не менее чем на 4 вопроса, аннотацию информационных источников, указанных в библиографии.

Примечание. П.10 – является заданием повышенной сложности и выполняется по желанию студента.

Контрольные вопросы:

1. Что определяет решающее правило?
2. Какие типы решающих правил применяют в диагностическом процессе при обработке результатов мониторинга?
3. В чем заключается логический способ синтеза решающего правила?
4. Каким образом осуществляется бинарное кодирование признакового пространства при синтезе логических решающих правил?
5. Как формулируется решающее правило продукционного типа?
6. Как осуществляется семантическое описание решающего правила?
7. Каким образом реализуется схемотехническая реализация решающего правила на определенной электронной базе?
8. Как проверяется качество применения решающего правила?

ПРИЛОЖЕНИЕ

Таблица 1. Результаты регионального мониторинга

год	Всего родилось	Всего заболело	Врожденные пороки (ВП)	асфаксия	Умерло всего	Умерло от ВП	Умерло от асфиксии
1	1657	90	22	6	17	4	5
2	2081	170	24	9	32	3	9
3	2173	201	20	25	32	7	5
4	2676	198	41	17	34	5	5
5	2557	191	51	47	21	3	10
6	2522	586	83	78	23	1	3
7	2893	252	30	19	31	3	7
8	2956	270	45	8	32	6	5
9	2650	197	38	12	25	3	6
10	3036	213	42	36	36	4	12
11	3165	230	37	32	21	2	5
12	3181	218	61	42	27	8	3
13	2930	216	65	58	18	1	6
14	2491	202	41	55	20	2	5
15	2964	185	37	39	25	2	1
16	2425	290	65	87	22	2	7
17	2432	238	50	53	19	3	4
18	2388	196	34	65	19	0	9
19	2290	197	34	58	22	4	9
20	2995	193	45	27	28	5	7

Год	Ревм. пораж. сердца	Инфаркт миокарда	Гипертан. болезнь	стенокардия	Септический эндокартит	летальность
1	107	8	171	30	5	18
2	147	4	151	38	9	20
3	146	22	124	42	4	17
4	122	27	145	56	10	20
5	104	37	134	83	8	23
6	77	37	110	33	11	23
7	82	24	156	38	8	20
8	104	37	100	40	4	13
9	88	21	87	45	35	18
10	75	17	111	41	6	2
11	44	28	71	29	3	17
12	56	4	120	38	8	12
13	44	16	100	31	7	7
14	44	12	104	28	5	24
15	57	15	97	35	15	10
16	32	17	84	33	14	5
17	36	15	116	40	20	16
18	44	20	176	44	2	28
19	51	23	157	37	11	13
20	48	18	162	43	17	17

ЛАБОРАТОРНАЯ РАБОТА №6. АНАЛИЗ ДИНАМИКИ ЭКОЛОГИЧЕСКОЙ СИТУАЦИИ В РЕГИОНЕ

Цель работы: овладение навыками математического моделирования экологической ситуации в регионе как одного из основных факторов влияющих на заболеваемость на основе результатов биометрического мониторинга.

Краткие теоретические сведения.

Ухудшение экологической обстановки и социальной среды существенно отражается на состоянии здоровья человека. Здоровье человека и биосферы неразделимо связаны и определяются множеством компонент. Взаимодействуя с миром в ходе своей деятельности, человек ощущает на себе ответную реакцию окружающей среды.

Кроме неизбежных природных явлений (таких как изменение солнечной активности), на здоровье человека могут влиять экологические факторы, вызываемые им самим в ходе своей деятельности. Указанное обусловливает необходимость управления деятельностью основных загрязнителей окружающей среды на основе математических методов (моделей), позволяющих достаточно адекватно оценить влияние различных факторов на рассматриваемый класс заболеваний с целью прогноза и (или) управления динамикой последнего.

Региональные экологические проблемы, сформировавшиеся в результате загрязнения окружающей среды из-за деятельности человека, требуют для своего решения использования региональных информационных систем (РИС). Одна из задач таких систем должна состоять в своевременном определении воздействия загрязняющих веществ на здоровье человека на основании анализа накопленной информации о состоянии окружающей среды и медико-биологической информации и прогнозирования уровня региональной заболеваемости. Исследование и прогноз выполняются на основе имитационного моделирования.

Для характеристики уровня экологической напряженности региона используют понятие экологической нормы, которое отражает определенные параметры сохранения приспособительных структур и функций экосистемы определенного иерархического

уровня. Такое определение нормы может указывать на степень максимально допустимого воздействия человека и общества на окружающую среду, которое обеспечивает функционирование и сохранение структуры и динамических качеств экосистемы в целом.

Поражение городского населения возможно главным образом через атмосферу, экологических природных систем - через все природные среды.

Таким образом, выделяются два аспекта проблемы для изучения влияния антропогенной деятельности человека на заболеваемость:

а) воздействие на импактном уровне (на относительно небольшой территории);

б) массовое воздействие на природу, природные экосистемы на фоновом (как правило, невысоком) уровне, но на обширных территориях, практически по всей территории земного шара.

В концептуальном моделировании принято рассматривать три этапа:

- сбор и анализ априорной информации о предметной области и проблемной среде;
- концептуальный анализ предметной области с учетом требований пользователей;
- концептуальный синтез или собственно построение концептуальной модели предметной области.

Общая технология экологического управления в регионе состоит из трех этапов.

Целью первого этапа является получение информации о фактическом загрязнении сред региона. Учитывается как анализ источников антропогенного загрязнения региона, так и анализ естественных процессов, приводящих к фоновым концентрациям загрязняющих веществ в средах региона.

На втором этапе оценивается влияние состояния среды на заболеваемость населения.

На третьем этапе строится прогноз заболеваемости населения в зависимости от состояния среды и изменение самой среды, с последующей выдачей рекомендаций планирующим, природоохранным и хозяйственным органам.

Разработка пути возможного оздоровления и профилактика уровня заболеваемости в регионе в автоматизированной системе основывается на оценке влияния выбросов отдельных предприятий на ту или иную заболеваемость с последующей выдачей рекомендаций планирующим, природоохранным и хозяйственным органам о проведении мероприятий, призванных скорректировать выбросы соответствующих предприятий. Таблица связей, полученная при помощи автоматизированной системой моделирования, призвана ставить в соответствие предприятиям региона мероприятия по снижению выбросов.

В общем виде методика анализа вклада выбросов отдельного предприятия на уровень заболеваемости населения выглядит следующим образом:

1) Определяется список экологических факторов, обусловленных выбросами V_I -V_p в окружающую среду, влияющих на уровень заболеваемости по нозологии N_k.
2) Используя полученную математическую модель влияния факторов окружающей среды на уровень заболеваемости осуществляется прогноз о показателях заболеваемости в конкретном регионе.
3) На основании полученных показателях заболеваемости оценивается вклад каждого выброса в рост уровня заболеваемости населения по определенной нозологии в регионе.
4) Определяется вклад каждого выброса конкретного предприятия административного района в рост уровня заболеваемости путем нахождения доли выброса этого предприятия к общим выбросам по району.
5) Определяется общий вклад предприятия в рост уровня заболеваемости по данной нозологии.
6) Определяется экономический ущерб U_k, вызванный влиянием деятельности предприятия на рост данной заболеваемости.
7) Шаги 1-6 повторяются для других нозологии, тем самым определяется общий экономический ущерб, вызванный влиянием деятельности предприятия на рост всей заболеваемости населения и определяется размер штрафа R_m для этого предприятия (т), эквивалентный сумме ущерба.

Далее из списка мероприятий для уменьшения показателей выбросов автоматически выбираются конкретные мероприятия, которые необходимо провести на данных предприятиях для уменьшения показателей выбросов, что в свою очередь должно привести к уменьшению заболеваемости системы кровоснабжения у населения.

Методика определения снижения значений выбросов загрязняющих веществ конкретными предприятиями состоит в следующем.

На первом этапе строится прогноз заболеваемости населения региона на следующий год при помощи автоматизированной системы математического моделирования.

Далее, определяется рост заболеваемости в процентном отношении для каждого района и для каждой нозологии.

На третьем этапе осуществляется коррекция значений выбросов предприятий региона с целью уменьшения уровня заболеваемости по следующему алгоритму. Задается порог роста заболеваемости h_i, при котором считается целесообразным принятие мер. Также определяется шаг снижения выбросов предприятием h.

После определения заболеваемости и района, в котором рост заболеваемости ожидается выше порогового уровня, определяются выбросы, значимо влияющие на эту заболеваемость по заданному критерию. Затем определяются предприятия региона, вносящие наибольший вклад в сброс этих веществ, и моделируется снижение уровня выбросов на заболеваемость.

Далее прогнозные значения уровня заболеваемости пересчитываются с новыми значениями выбросов загрязняющих веществ и проверяется условие роста заболеваемости в этом районе. Операция повторяется до тех пор, пока рост заболеваемости не станет ниже hi.

Порядок выполнения работы.

1. Определить вариант задания по формуле *Nвар=N mod (4)+1*, где *N* - порядковый номер в группе. Согласно приложению и *Nвар* сформировать протокол мониторинга экологической ситуации города *{XNвар, Y1, Y2, Y3}*, где *X* - показатель уровня

заболеваемости населения, Y - показатель загрязненности города определенным веществом.

2. Построить линейные и-или нелинейные регрессионные модели вида (отобрать лучшие по критерию детерменированности):

2.1. $X=f(Y_1)$; $X=f(Y_2)$; $X=f(Y_3)$; $X=f(Y_1,Y_2)$; $X=f(Y_1,Y_3)$; $X=f(Y_2,Y_3)$; $X=f(Y_1,Y_2,Y_3)$;

2.2 Повторить пункт 2.1, используя в качестве Y_i, $i=1..3$ значения $Y_{i,j}$ с нарастающим шагом, т.е.

$$Y'_{ij} = \sum_{k=1}^{j-1} Y_{ik}$$

3. На основе информационных источников изучить, что влияет на уменьшение Y_i.

4. На основании анализа математических моделей, индентифицированных в пункте 2 и результатов «экспертного» анализа пункта 3, сформулировать предложения управляющего воздействия на факторы Y с целью улучшения показателя отклика X. Оцените доминантность влияния уровня определенного загрязнения на параметр здоровья. Сделайте вывод о структуре наиболее адекватной модели по критериям корреляции и СКО. Сравните частоты ритмической модели (если она адекватна) с внешними космогеологическими частотами (см. Приложение) и сделайте выводы. Ритмическую модель в данной работе предлагается получить путем анализа автокорреляционной функции или визуального анализа динамики заболеваемости во времени или визуального анализа поведения первой производной, вычисленной численными методами.

5. Оформите отчет, включающий в себя результаты работы (возможны скриншоты), выводы, краткие ответы на контрольные вопросы, аннотацию одного из информационных источников, указанных в библиографии или иных (или найденного самостоятельно).

Контрольные вопросы:

1. С какими природными циклами наиболее коррелирует динамика определенных заболеваний?

2. Каким образом связаны между собой уровни заболеваемости населения и уровни антропогенного воздействия на окружающую среду (на примере уровней загрязнителей)?
3. Почему антропогенное воздействие следует учитывать с нарастающим эффектом?
4. Каким образом используются регрессионные и авторегрессионные математические модели для прогнозирования заболеваний?
5. Как осуществляется прогнозирование в Excel с помощью линии тренда?
6. Каким образом можно прогнозировать ритмические тенденции региональной заболеваемости?
7. Каким образом можно использовать логические функции (модели) для прогнозирования заболеваний?
8. Можно ли использовать искусственные нейронные сети для прогнозирования заболеваемости?
9. Каким образом можно использовать прогностические модели для удаления артефактов и восстановления пропущенных значений в мониторинге заболеваемости или состояния пациента в процессе терапевтического воздействия?

Приложение к лабораторной работе №6

Таблица 2. уровни показателей здоровья городского населения

Годы	Смертность	рождаемость	Прививки от дифтерии	Прививки от столбняка	Прививки от кори	Прививки от гепатита
1	13,2	14,99	10,2	24,6	8,9	15,8
2	11,9	13,36	11,0	30,5	7,7	34,3
3	12,8	11,4	9,5	29,3	9,6	24,9
4	13,5	12,23	10,3	27,7	10,6	28,5
5	13,4	12,92	11,3	43	11,7	18,9
6	13,0	11,67	11,2	44,7	11,5	18,2
7	13,7	11,55	12,4	47,5	10,7	13,0
8	13,0	13,3	12,6	44,1	11,0	19,1
9	14,2	14,7	11,7	44,4	11,9	24,2
10	14,6	14,1	12,1	46,6	18,2	28,2
11	13,0	14,6	13,2	39,5	16,6	12,3
12	13,3	14,8	13,0	40,0	16,3	19,51
13	13,6	13,8	13,5	38,7	17,4	10,12
14	13,6	11,8	14,5	40,9	26,8	19,91
15	14,6	10,8	13,8	34,2	25,1	12,54
16	14,7	9,7	11,6	28,3	12,1	4,2
17	16,5	9,0	11,0	31,1	11,0	85,7
18	18,0	9,3	7,3	16,6	5,6	72,2
19	16,7	8,6	11,5	38,7	3,7	2,02
20	17,0	8,3	11,1	52,5	11,4	22,6

Таблица 2. -уровни загрязнения города (условные единицы)

годы	Пыль	Оксид углерода	Диоксид азота	фенол	формальдегид	марганец
1	6	1,5	2,667	8	3,5	0,8
2	4	0,667	2,222	8	2,75	0,16
3	6	1,45	1,333	6	2,25	1,7
4	2	1,067	4,444	6	2,75	1,7
5	4	1,833	0,889	6	3,25	2,5
6	4	1,833	0,889	4	3,25	1,1
7	6	1,667	1,333	6	3,75	2,0
8	6	1,667	1,333	10	3,0	0,7
9	8	2	1,156	8	2,75	0,6
10	4	1,667	0,889	7,4	2,0	1,2
11	4	1,667	1,333	8	3,0	3,5
12	2	1,667	1,33	8	3,75	1,4
13	2	1,667	1,778	10	3,5	3,2
14	2	1,667	1,333	6	3,25	1,4
15	2	1,663	2,222	4	3,0	0,2
16	2	1,667	1,333	4	3,0	0,6
17	4	1,667	1,333	2	2,5	0,4
18	5	2,333	1,333	8	2,25	0,5
19	1,8	2,5	1,333	4	3,0	0,6
20	1,8	2,67	1,3	4	3,25	0,1

ЛАБОРАТОРНАЯ РАБОТА № 7. КОРРЕЛЯЦИОННЫЙ И АВТОКОРРЕЛЯЦИОННЫЙ АНАЛИЗЫ В БИОМЕДИЦИНСКОЙ ПРАКТИКЕ

Цель работы: овладение навыками использования инструментария ПЭВМ для проведения исследований в области изучения связей между регистрируемыми характеристиками состояния биообъекта методами корреляционного и регрессионного анализов.

Краткие теоретические сведения

В большинстве случаев результаты медико-биологических исследований представляют собой либо вектор, характеризующий изменения биологических сигналов во времени – ЭКГ, реограммы и т.п. (временной ряд) либо их совокупность либо вектор, характеризующий значения различных характеристик зарегистрированных практически одномоменто.

При моделировании временных рядов модели в виде эмпирических формул, полиномов или рядов Фурье являются функциональными моделями, так как представляют собой аналитическое выражение или функцию, которая отражает функциональную связь между зависимым признаком (амплитудой сигнала или артериальным давлением) и независимым признаком – временем.

Вообще отличительной чертой биосистем является многообразие признаков, характеризующих эту биосистему (возраст, пол, рост, вес, пульс, давление и так далее). Часто между вариациями этих признаков существует связь, (например: обычно чем больше рост, тем больше вес, но могут быть и исключения).

Эта связь не является строго функциональной, так как возможны отклонения от общей тенденции, она называется корреляционной (от английского correlation – соотношение, взаимосвязь, взаимодействие). Корреляция между двумя признаками (с точки зрения статистических догматов) состоит в том, что среднее значение одного признака (результативного или зависимого) изменяется в зависимости от изменения среднего значения другого признака (факторного или независимого). Поэтому, если нужно построить модель для связи средних

значений каких-либо признаков (например, среднегодовым числом сердечно-сосудистых заболеваний и среднегодовым значением метео-осадков), то для этого используют методы корреляционного анализа. Предметом корреляционного анализа является выявление корреляционной связи между признаками.

Корреляционный анализ включает в себя четыре этапа:

1. выявление наличия корреляционной связи,
2. определение формы корреляционной связи,
3. вычисление тесноты корреляционной связи,
4. оценка статистической достоверности результатов п.3.

Пусть биосистемой является пациент. У него измерено два признака: вес (Р) и артериальное давление (АД). Данные измерений занесены в таблицу. Цель исследования – найти корреляционную связь между этими признаками.

Применим поэтапно корреляционный анализ.

Этап 1. Выявление наличия корреляционной связи. Этот этап проводится при помощи построения корреляционного поля или диаграммы рассеивания. Чтобы построить корреляционное поле, нужно в координатной плоскости признаков нанести все объекты. Каждый объект будет изображен в виде точки с координатами, соответствующими значениям признаков. Множество таких точек и будет *корреляционным полем* или диаграммой рассеивания.

Для построения диаграммы рассеивания необходимо использовать в электронной таблице режим «Диаграмма точечная». Чтобы выявить наличие или отсутствие корреляционной связи, нужно проанализировать форму корреляционного поля (диаграммы рассеивания). При наличии корреляционной связи форма корреляционного поля близка к элипсоидной. Оси эллипса расположены по диагонали по отношению к осям координат в координатной плоскости. Главная ось эллипса наклонена к оси абсцисс по острым углом. Если наклон «направо», то корреляция положительная, «налево» - отрицательная.

Этап 2.Определение формы корреляционной связи**.** Если ранжировать выборку из двух признаков по возрастанию одного из них и отобразить график зависимости между признаками, то, наблюдая за полученной кривой можно предположить вид корреляционной функции между признаками. Вид этой функции

называется линией регрессии. (Заметим, что если каждому значению одного признака соответствует только одно значение другого – то связь называется функциональной). Полученную линию регрессии можно промоделировать аналитической функцией, уравнение которой называется уравнением регрессии. Оно определяет форму корреляционной связи. Если линия регрессии моделируется прямой, то корреляционная связь между признаками называется линейной регрессией. Это частный, но весьма распространённый случай. Для выявления уравнения регрессии в Excel на графике, отражающем связь между характеристиками необходимо построить линию тренда, отобразить на графике коэффициент детерминации (квадрат парной корреляции в данном случае) и график зависимости, выбрать регрессионное уравнение соответствующее максимальному значению коэффициента детерминации.

Этап 3. Вычисление тесноты корреляционной связи. Тесноту или степень связи можно вычислить при помощи коэффициентов корреляции. В случае линейной регрессии между признаками x и y тесноту связи характеризует линейный коэффициент парной корреляции $r_{y/x}$. Величина $r_{y/x}$ изменяется в пределах от –1 до 1 ($-1< r_{y/x}<1$). Если $r_{y/x} > 0$, то корреляционная связь между признаками прямая, причём, чем ближе к 1, тем связь теснее, тем больше она приближается к строгой (функциональной) линейной связи. Если $r_{y/x} < 0$, то связь обратная, причём, чем ближе $r_{y/x}$ к -1,тем теснее обратная связь. Если $r_{y/x} = 1$ (или $r_{y/x}= -1$), то связь между признакам - строгая функциональная (прямая или обратная).

Оценка статистической значимости полученного значения коэффициента парной корреляции (с применением соответствующей функции в Excel) осуществляется, например, путем вычисления коэффициента Стьюдента и оценки его статистической значимости путем вычисления значения ошибки первого рода и сравнения с пороговым значением. Коэффициент Стьюдента вычисляется по формуле: $\sqrt{\boldsymbol{t = \frac{ry\,/x^2}{\sqrt[2]{1-|ry\,/x|}}}}\sqrt[2]{\boldsymbol{n-1}}$. Где n – количество значений пар x, y в выборке.

Изменение регистрируемой характеристики во времени представляется в виде одномерного массива, называемого «временным рядом».

Временной ряд является нестационарным, если он содержит такие систематические составляющие как тренд и цикличность. Нестационарные временные ряды характеризуются тем, что значения каждого последующего уровня временного ряда корреляционно зависят от предыдущих значений.

Для исследования временного ряда на первоначальном этапе осуществляется построение и изучение автокорреляционной функции. Автокорреляцией уровней временного ряда называется корреляционная зависимость между настоящими и прошлыми значениями уровней данного ряда.

Лагом называется величина сдвига между рядами наблюдений. Лаг временного ряда определяет порядок коэффициента автокорреляции. Например, если уровни временного ряда x_t и x_{t-1} корреляционно зависимы, то величина временного лага равна единице. Следовательно, данная корреляционная зависимость определяется коэффициентом автокорреляции первого порядка между рядами наблюдений $x_1 \ldots x_{n-1}$ и $x_2 \ldots x_n$. . Если лаг между рядами наблюдений равен двум, то данная корреляционная зависимость определяется коэффициентом автокорреляции второго порядка и т. д.

При увеличении величины лага на единицу число пар значений, по которым рассчитывается коэффициент автокорреляции, уменьшается на единицу. Поэтому максимальный порядок коэффициента автокорреляции рекомендуется брать равным n/4, где n – количество уровней временного ряда.

Автокорреляция между уровнями временного ряда оценивается с помощью выборочного коэффициента автокорреляции.

Анализ структуры временного ряда с помощью коэффициентов автокорреляции стоится на следующих правилах:

1) исследуемый временной ряд содержит только трендовую компоненту, если наибольшим является значение коэффициента автокорреляции первого порядка r_{l-1};

2) исследуемый временной ряд содержит трендовую компоненту и колебания периодом l, если наибольшим является коэффициент автокорреляции порядка l. Эти колебания могут быть как циклическими, так и сезонными;
3) если ни один из коэффициентов автокорреляции $r_l(l=1,L)$ не окажется значимым, то делается один из двух возможных выводов:
а) данный временной ряд не содержит трендовой и циклической компонент, а его колебания вызваны воздействием случайной компоненты, т. е. ряд представляет собой модель случайного тренда;
б) данный временной ряд содержит сильную нелинейную тенденцию, для выявления которой необходимо провести его дополнительный анализ.

Графическим способом анализа структуры временного ряда является построение графиков автокорреляционной и частной автокорреляционной функций.

Автокорреляционной функцией называется функция оценки коэффициента автокорреляции в зависимости от величины временного лага между исследуемыми рядами. Графиком автокорреляционной функции является коррелограмма.

Частная автокорреляционная функция отличается от автокорреляционной функции тем, что при её построении устраняется корреляционная зависимость между наблюдениями внутри лагов.

Формула для расчета коэффициента автокорреляции имеет вид:

$$r_1 = \frac{\sum_{t=2}^{n}(y_t - \bar{y}_1)(y_{t-1} - \bar{y}_2)}{\sqrt{\sum_{t=2}^{n}(y_t - \bar{y}_1)^2 \sum_{t=2}^{n}(y_{t-1} - \bar{y}_2)^2}}, \quad (1)$$

где: $\bar{y}_1 = \frac{1}{n-1}\sum_{t=2}^{n} y_t, \quad \bar{y}_2 = \frac{1}{n-1}\sum_{t=2}^{n} y_{t-1}.$

Эту величину называют коэффициентом автокорреляции уровней ряда первого порядка, так как он измеряет зависимость между соседними уровнями ряда t и y_{t-1}.

Аналогично можно определить коэффициенты автокорреляции второго и более высоких порядков. Так, коэффициент автокорреляции второго порядка характеризует тесноту связи между уровнями y_t и y_{t-2} и определяется по формуле:

$$r_2 = \frac{\sum_{t=3}^{n}(y_t - \bar{y}_3)(y_{t-2} - \bar{y}_4)}{\sqrt{\sum_{t=3}^{n}(y_t - \bar{y}_3)^2 \sum_{t=3}^{n}(y_{t-2} - \bar{y}_4)^2}}, \quad (2)$$

где

$$\bar{y}_3 = \frac{1}{n-2}\sum_{t=3}^{n} y_t, \qquad \bar{y}_4 = \frac{1}{n-2}\sum_{t=3}^{n} y_{t-2}.$$

Число периодов, по которым рассчитывается коэффициент автокорреляции, называют лагом. С увеличением лага число пар значений, по которым рассчитывается коэффициент автокорреляции, уменьшается. Считается целесообразным для обеспечения статистической достоверности коэффициентов автокорреляции использовать правило – максимальный лаг должен быть не больше $n/4$.

Свойства коэффициента автокорреляции.

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.

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

Последовательность коэффициентов автокорреляции уровней первого, второго и т.д. порядков называют автокорреляционной функцией временного ряда. График зависимости ее значений от величины лага (порядка коэффициента автокорреляции)

называется коррелограммой. Анализ автокорреляционной функции и коррелограммы позволяет определить лаг, при котором автокорреляция наиболее высокая, а следовательно, и лаг, при котором связь между текущим и предыдущими уровнями ряда наиболее тесная, т.е. при помощи анализа автокорреляционной функции и коррелограммы можно выявить структуру ряда.

Если наиболее высоким оказался коэффициент автокорреляции первого порядка, исследуемый ряд содержит только тенденцию. Если наиболее высоким оказался коэффициент автокорреляции порядка τ, то ряд содержит циклические колебания с периодичностью в τ моментов времени. Если ни один из коэффициентов автокорреляции не является значимым, можно сделать одно из двух предположений относительно структуры этого ряда: либо ряд не содержит тенденции и циклических колебаний, либо ряд содержит сильную нелинейную тенденцию, для выявления которой нужно провести дополнительный анализ. Поэтому коэффициент автокорреляции уровней и автокорреляционную функцию целесообразно использовать для выявления во временном ряде наличия или отсутствия трендовой компоненты и циклической (сезонной) компоненты.

Моделирование тенденции временного ряда

Распространенным способом моделирования тенденции временного ряда является построение аналитической функции, характеризующей зависимость уровней ряда от времени, или тренда. Этот способ называют аналитическим выравниванием временного ряда.

Поскольку зависимость от времени может принимать разные формы, для ее формализации можно использовать различные виды функций. Для построения трендов чаще всего применяются следующие функции: линейный тренд $\hat{y}_t = a + b \cdot t$; гипербола: $\hat{y}_t = a + b/t$; экспоненциальный тренд: $\hat{y}_t = \mathrm{e}^{a+b \cdot t}$ (или $\hat{y}_t = a \cdot b^t$); степенная функция: $\hat{y}_t = a \cdot t^b$; полиномы различных степеней: $\hat{y}_t = a + b_1 \cdot t + b_2 \cdot t^2 + \ldots + b_m \cdot t^m$.

Параметры каждого из перечисленных выше трендов можно определить обычным МНК, используя в качестве независимой

переменной время $t = 1, 2, ..., n$, а в качестве зависимой переменной – фактические уровни временного ряда $\hat{y}_t$. Для нелинейных трендов предварительно проводят стандартную процедуру их линеаризации.

Существует несколько способов определения типа тенденции. К числу наиболее распространенных способов относятся качественный анализ изучаемого процесса, построение и визуальный анализ графика зависимости уровней ряда от времени.

В этих же целях можно использовать и коэффициенты автокорреляции уровней ряда. Тип тенденции можно определить путем сравнения коэффициентов автокорреляции первого порядка, рассчитанных по исходным и преобразованным уровням ряда.

Если временной ряд имеет линейную тенденцию, то его соседние уровни $\hat{y}_t$ и $\hat{y}_{t-1}$ тесно коррелируют. В этом случае коэффициент автокорреляции первого порядка уровней исходного ряда должен быть высоким.

Если временной ряд содержит нелинейную тенденцию, например, в форме экспоненты, то коэффициент автокорреляции первого порядка по логарифмам уровней исходного ряда будет выше, чем соответствующий коэффициент, рассчитанный по уровням ряда. Чем сильнее выражена нелинейная тенденция в изучаемом временном ряде, тем в большей степени будут различаться значения указанных коэффициентов.

Выбор наилучшего уравнения в случае, когда ряд содержит нелинейную тенденцию, можно осуществить путем перебора основных форм тренда, расчета по каждому уравнению скорректированного коэффициента детерминации и средней ошибки аппроксимации. Этот метод легко реализуется при компьютерной обработке данных.

Порядок выполнения.

1. Изучить теоретический материал.
2. Построить корреляционные зависимости, определить коэффициенты корреляции (с оценкой значимости – ошибки первого рода), наиболее адекватные уравнения регрессии между

характеристиками – показателями крови в каждом из альтернативных классов (см. приложение к лабораторной работе 1). Рассмотреть не менее 4 пар (формат пар определяет преподаватель).
3. Построить автокорреляционные функции (порядка 10-15 первых значений) для ФПГ до и после нагрузки (см. Приложение).
4. Оформите отчет, включающий в себя результаты выполнения (скрин-шоты), выводы, ответы на контрольные вопросы (не менее пяти, из них №№12,13,14 – обязательны).

Контрольные вопросы.
1. Чем отличается функциональная и корреляционная связь между признаками?
2. Что такое временной ряд биофизиологического сигнала?
3. Что такое ранжирование выборки?
4. В каком случае регрессия будет линейной?
5. В каком случае линии регрессии совпадают?
6. Что можно сказать о корреляционной связи между признаками X и Y если значение коэффициента корреляции равно 0,3?
7. Что можно сказать о корреляционной связи между признаками, если корреляционное поле имеет форму круга?
8. Какой метод применяется для нахождения коэффициентов уравнения линейной регрессии?
9. Что такое автокорреляционная функция? Как она определяется?
10. Каким образом оценивается значимость коэффициента корреляции?
11. Как строится коррелограмма?
12. Каким образом в электронной таблице осуществляется корреляционный анализ?
13. Как применяются результаты корреляционного анализа в медицине?
14. Как применяется автокорреляция в медицине?

Библиография
1. Автокорреляция.Презентация 1. /URL: http://www.myshared.ru/slide/82380/

2. Автокорреляция.Презентация 2. /URL: http://present5.com/prezentaciya-kafedra-avtomatizacii-obrabotki-informacii2/
3. Автокорреляция уровней временного ряда. /URL: http://be5.biz/ekonomika/e011u/80.htm
4. Корреляция. Презентация. /URL: http://informslide.ru/korrelyacionnye-zavisimosti/
5. Корреляционные зависимости. Презентация. /URL: http://www.myshared.ru/slide/84314/
6. Статистическая обработка данных. Презентация. /URL: http://informslide.ru/statisticheskaya-obrabotka-dannyx/

Приложение к лабораторной работе № 7:
ФПГ (до и после нагрузки)

№ п.п.	До нагрузки	После нагрузки
1	0,395	0,00
2	0,37	0,03
3	0,32	0,04
4	0,30	0,17
5	0,30	0,46
6	0,34	0,84
7	0,43	1,25
8	0,53	1,62
9	0,71	1,93
10	0,96	2,23
11	1,21	2,45
12	1,47	2,65
13	1,75	2,80
14	2,02	2,92
15	2,23	2,98
16	2,49	3,02
17	2,68	3,02
18	2,83	2,99
19	2,98	2,96
20	3,07	2,89
21	3,14	2,80
22	3,19	2,72
23	3,20	2,57
24	3,23	2,45
25	3,22	2,31
26	3,19	2,15
27	3,16	2,02
28	3,10	1,92
29	3,06	1,78
30	2,99	1,69
31	2,89	1,62
32	2,82	1,58
33	2,75	1,54
34	2,67	1,55
35	2,59	1,53
36	2,51	1,55
37	2,44	1,56
38	2,39	1,58
39	2,31	1,54
40	2,25	1,52
41	2,24	1,52

42	2,19	1,50
43	2,15	1,44
44	2,13	1,42
45	2,08	1,35
46	2,04	1,28
47	2,01	1,22
48	1,98	1,15
49	1,92	1,05
50	1,87	0,97
51	1,82	0,88
52	1,74	0,77
53	1,69	0,68
54	1,63	0,58
55	1,54	0,47
56	1,46	0,39
57	1,39	0,28
58	1,30	0,17
59	1,21	0,07
60	1,16	0,02