Севастопольский национальный технический
университет

МЕТОДИЧЕСКИЕ УКАЗАНИЯ
к выполнению курсовой работы

СПЕЦИАЛЬНЫЕ СТАТИСТИЧЕСКИЕ ПРИЛОЖЕНИЯ И МОДЕЛИ
по дисциплине

«СПЕЦИАЛЬНЫЕ ГЛАВЫ МАТЕМАТИКИ»

для студентов направления подготовки
09.03.02 *Информационные системы и технологии*
дневной и заочной форм обучения

Уровень подготовки *Магистр*

Севастополь
2014

УДК 519.6

Специальные статистические приложения и модели. Методические указания и индивидуальные задания к курсовой работе по дисциплине «Специальные главы математики» / Сост. Первухина Е.Л., Голикова В.В. – Севастополь: Изд-во СевНТУ, 2014. – 14 с.

Целью курсовой работы является закрепление, углубление и обобщение знаний по специальным разделам математики, необходимых при обработке экспериментальных данных, статистическом моделировании сложных систем различной физической природы, а также получение навыков работы с перспективными вычислительными средствами.

Методические указания предназначены для студентов направления подготовки 09.03.02 *Информационные системы и технологии* (квалификация (степень) "МАГИСТР")

Методические указания рассмотрены и утверждены на заседании кафедры информационных систем (протокол № 1 от 26 августа 2014 г.).

Допущено учебно-методическим центром СевНТУ в качестве методических указаний.

Рецензент: Скатков А.В., д.т.н., проф. каф. *Кибернетики и вычислительной техники*

СОДЕРЖАНИЕ

1 ЦЕЛЬ И ЗАДАЧИ КУРСОВОЙ РАБОТЫ

Целью курсовой работы является закрепление, углубление и обобщение знаний по специальным разделам математики, необходимых при обработке экспериментальных данных, статистическом моделировании сложных систем различной физической природы, а также получение навыков работы с перспективными вычислительными средствами.

Курсовая работа связана с решением инженерных задач с использованием информационных систем последнего поколения и с приобретением навыков научного исследования, в том числе при написании дипломной работы.

Курсовая работа выполняется по индивидуальным заданиям и включает следующие этапы:

- углубленное изучение методов статистического моделирования сложных систем различной физической природы;
- знакомство с вычислительными средствами, используемыми в современных информационных системах;
- построение алгоритмов и решение задач.

В качестве источника экспериментальной информации об экономических системах используется материал базы данных Федеральной службы государственной статистики РФ, Российский статистический ежегодник, реальная экспериментальная информация о параметрах технических объектов.

2 КРАТКИЕ ТЕОРЕТИЧЕСКИЕ СВЕДЕНИЯ

2.1 Основные понятия в анализе временных рядов

2.1.1 Описание математических методов.

Табличное представление временного ряда и описательные статистики чаще всего не позволяют понять характер процесса, в то время как по графику временного ряда можно сделать довольно много выводов. В дальнейшем они могут быть проверены и уточнены с помощью расчетов.

Построение графика временного ряда – важная задача. Проблему для воспроизведения графиков создают выбросы – наблюдения в несколько раз превышающие по величине большинство остальных значений ряда.

При анализе временных рядов часто используются вспомогательные графики для числовых характеристик ряда:

график выборочной автокорреляционной функции (коррелограммы) с доверительной трубкой для нулевой автокорреляционной функции;

график выборочной частной автокорреляционной функции с доверительной трубкой для частной нулевой автокорреляционной функции.

2.1.2 Автоковариационная функция

Предположим, что процесс $\{Z_t\}$ стационарен. Для стационарного гауссовского процесса имеем:

1) $E(Z_t) = \mu = const$;

2) $Var(Z_t) = E(Z_t - \mu_t)^2 = \sigma^2 = const$;

3) $Cov(Z_t, Z_s) = E(Z_t - \mu_t)(Z_s - \mu_s)$ являются функциями только модуля разности $|t - s|$.

В этом случае, очевидно, что ковариация между Z_t и Z_{t+k} является функцией только переменной k и не зависит от t:

$$\gamma_k = Cov(Z_t, Z_{t+k}) = E(Z_t - \mu)(Z_{t+k} - \mu).$$

Ковариация имеет недостатки: с трудом поддается непосредственной интерпретации, так как зависит от единиц измерения Z_t, Поэтому определяется корреляция между Z_t и Z_{t+k}:

$$\rho_k = \frac{Cov(Z_t, Z_{t+k})}{\sqrt{Var(Z_t)}\sqrt{Var(Z_{t+k})}} = \frac{\gamma_k}{\gamma_0},$$

где $\gamma_0 = Var(Z_t) = Var(Z_{t+k})$.

В анализе временных рядов функция γ_k называется автоковариационной функцией, а функция ρ_k – автокорреляционной функцией. Такое название объясняется тем, что они представляют собой ковариацию и корреляцию, измеренную для различных значений Z_t и Z_{t+k} одного и того же временного ряда. Легко убедиться, что для стационарного процесса ρ_k и γ_k удовлетворяют следующим свойствам:

1. $\gamma_0 = Var(Z_t)$, $\rho_0 = 1$;
2. $|\gamma_k| \le \gamma_0$, $|\rho_k| \le 1$;
3. $\gamma_k = \gamma_{-k}$ и $\rho_k = \rho_{-k}$.

Последнее свойство означает, что ρ_k и γ_k являются четными функциями и, следовательно, симметричны относительно точки $k = 0$. В силу этого эти функции рассматриваются только для неотрицательных значений аргумента.

График автокорреляционной функции является важным средством идентификации модели временного ряда. Он строится только для неотрицательных значений дискретного аргумента k, называется коррелограммой.

2.1.3 Частная автокорреляционная функция

В анализе временных рядов частная автокорреляционная функция наряду с автокорреляционной функцией играет решающую роль в идентификации временного ряда. ЧАКФ измеряет корреляцию между случайными величинами Z_t и Z_{t+k}, «очищенную» от влияния случайных величин $Z_{t+1}, Z_{t+2}, ..., Z_{t+k-1}$, находящихся между ними. Фактически ЧАКФ есть условная корреляция $Corr(Z_t, Z_{t+k} | Z_{t+1}, Z_{t+2}, ..., Z_{t+k-1})$.

Если рассмотреть регрессионную модель с зависимой переменной Z_{t+k} (случайная переменная из стационарного процесса с нулевым средним) и k лаговыми переменными $Z_{t+k-1}, Z_{t+k-2}, ..., Z_t$, используемыми в качестве независимых переменных, то:

$$Z_{t+k} = \Phi_{k1} Z_{t+k-1} + \Phi_{k2} Z_{t+k-2} + \cdots + \Phi_{kk} Z_t + e_{t+k},$$

где Φ_{ki} — i-ый регрессионный параметр, e_{t+k} – случайная ошибка, распределенная по нормальному закону, некоррелированная с Z_{t+k-j} для $j \geq 1$. Умножая на Z_{t+k-j} обе части данного регрессионного уравнения и применив оператор математического ожидания, получаем $\gamma_j = \Phi_{k1}\gamma_{j-1} + \Phi_{k2}\gamma_{j-2} + \cdots + \Phi_{kk}\gamma_{j-k}$, откуда

$$\rho_j = \Phi_{k1}\rho_{j-1} + \Phi_{k2}\rho_{j-2} + \cdots + \Phi_{kk}\rho_{j-k}.$$

При одной реализации естественной оценкой среднего $\mu = E(Z_t)$ процесса является выборочное среднее – среднее из n наблюдений по времени:

$$\bar{Z} = \frac{1}{n}\sum_{t=1}^{n} Z_t.$$

2.1.4 Выборочная автоковариационная и автокорреляционная функция

Располагая единственной реализацией случайного процесса, можно использовать усреднение по времени для оценивания автоковариационной функции γ_k:

$$\hat{\gamma}_k = \frac{1}{n}\sum_{t=1}^{n-k}(Z_t - \bar{Z})(Z_{t+k} - \bar{Z})$$ или

$$\hat{\hat{\gamma}}_k = \frac{1}{n-k}\sum_{t=1}^{n-k}(Z_t - \bar{Z})(Z_{t+k} - \bar{Z}).$$

Обе эти оценки являются смещенными. В общем случае оценка $\hat{\gamma}_k$ имеет большее смещение, чем $\hat{\hat{\gamma}}_k$, особенно, когда k велико относительно n. Последнее обстоятельство является причиной того, что в анализе временных рядов предполагается вычисление оценок автоковариаций и автокорреляций для значений k не больше $n/4$.

Выборочная автокорреляционная функция вычисляется на основе автоковариационной функции по формуле:

$$\hat{\rho}_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_0} = \frac{\sum_{t=1}^{n-k}(Z_t - \bar{Z})(Z_{t+k} - \bar{Z})}{\sum_{t=1}^{n}(Z_t - \bar{Z})^2}.$$

Заметим, что выборочная АКФ $\hat{\rho}_k$ симметрична относительно начала $k = 0$, как и истинная АКФ ρ_k

$$\hat{\rho}_k = \frac{\sum_{t=1}^{n-k}\left(Z_t-\overline{Z}\right)\left(Z_{t+k}-\overline{Z}\right)}{\sum_{t=1}^{n}\left(Z_t-\overline{Z}\right)^2} = \frac{\sum_{t=k+1}^{n}\left(Z_t-\overline{Z}\right)\left(Z_{t-k}-\overline{Z}\right)}{\sum_{t=1}^{n}\left(Z_t-\overline{Z}\right)^2} = \hat{\rho}_{-k}.$$

2.1.5 Выборочная частная автокорреляционная функция

Существует рекурсивный метод для вычисления $\hat{\Phi}_{kk}$, позволяющий избежать сложного вычисления определителей матриц больших размерностей. Метод заключается в последовательном вычислении $\hat{\Phi}_{kk}$ по следующим формулам:

$$\hat{\Phi}_{11} = \hat{\rho}_1,$$

$$\hat{\Phi}_{k+1,k+1} = \frac{\hat{\rho}_k - \sum_{j=1}^{k}\hat{\Phi}_{kj}\rho_{k+1-j}}{1-\sum_{j=1}^{k}\hat{\Phi}_{kj}\hat{\rho}_j},$$

$$\hat{\Phi}_{k+1,j} = \hat{\Phi}_{kj} - \hat{\Phi}_{k+1,k+1}\hat{\Phi}_{k,k+1-j}, \quad j=1,\ldots,k$$

Этот метод справедлив так же для вычисления теоретической ЧАКФ Φ_{kk} как функции ρ_i.

2.1.6 Задания

Для заданного временного ряда:

1. Получить дескриптивные статистики;
1. Построить график значений временного ряда;
2. Построить графики выборочных АКФ и ЧАКФ.

2.2 Исследование свойств нестационарных временных рядов. Методология Бокса-Дженкинса

2.2.1 Линейные стохастические модели ARIMA

Методика идентификации временных рядов включает два основных шага: идентификацию порядка разности и идентификацию результирующего стационарного процесса АРСС.

Переход от нестационарных процессов к стационарным осуществляется с помощи взятия разностей порядка d: $\nabla^d = (1-U)^d Y_i$. При d=1 $\nabla Y_i = Y_i - Y_{i-1} = (1-U)Y_i$, при d=2 $\nabla^2 Y_i = (1-U)^2 Y_i$.

Такие модели называются процессами авторегрессии – проинтегрированного скользящего среднего (АРПСС, или ПАРСС, в английской аббревиатуре ARIMA – *autoregresive integrated moving average*).

Рассматриваемый динамический процесс удовлетворяет следующей разностной схеме

$$\Phi(U)\nabla^d Y_i = \Theta_0 + \Theta(U)\varepsilon_i, \qquad i=0, \pm 1, \pm 2, \ldots \quad (1)$$

Здесь ∇ - оператор взятия разности, U - оператор сдвига назад: $UY_i = Y_{i-1}$, $U^j Y_i = Y_{i-j}$ (F - оператор сдвига вперед, такой что $FY_i = Y_{i+1}$, $F^j Y_i = Y_{i+j}$, связан с оператором сдвига назад отношением $F = U^{-1}$). В выражении (1)

$$\Phi(U) = 1 - \phi_1 U - \phi_2 U^2 - \ldots - \phi_p U^p, \quad (2)$$

$\phi_1 \ldots \phi_p$ - конечное множество коэффициентов авторегрессии, p - порядок авторегрессии,

$$\Theta(U) = 1 - \theta_1 U - \theta_2 U^2 - \ldots - \theta_q U^q, \quad (3)$$

$\theta_1, \ldots, \theta_q$ - конечное множество коэффициентов скользящего среднего, q - порядок скользящего среднего, ε_i - белый шум (т.е. последовательность независимых одинаково распределенных случайных величин с конечной дисперсией). Если в (1) оператор авторегрессии имеет порядок p, взята d-я разность и оператор скользящего среднего имеет порядок q, говорят, что модель имеет порядок (p, d, q).

Условие стационарности предполагает расположение корней полинома (2) вне единичного круга.

Для стационарной модели, у которой ни один из корней не лежит близко к границе единичного круга, автокорреляционная функция быстро затухает при средних и больших k.

Отсутствие у автокорреляционной функции тенденции к затуханию может рассматриваться как свидетельство того, что существует корень, близкий к 1.

Нестационарность подсказывается отсутствием быстрого спада выборочной автокорреляционной функции. При этом не обязательно, чтобы выборочные корреляции при малых задержках были велики.

По приведенным причинам предполагается, что необходимая для получения стационарности степень разности d достигнута, если автокорреляционная функция ряда $\nabla^d Y_i$ быстро затухает. На практике d обычно равно 0, 1 или 2, и достаточно просмотреть примерно 20 первых значений автокорреляции исходного ряда, его первых и вторых разностей.

2.2.2 Идентификация результирующего стационарного процесса АРСС

После принятия предварительного решения о величине d, изучают общий вид выборочных автокорреляционной и частной автокорреляционной функций соответствующего разностного ряда, чтобы выбрать порядок p и q операторов авторегрессии и скользящего среднего. При этом учитываются характерные особенности поведения теоретической автокорреляционной функции и теоретической частной автокорреляционной функции для процессов авторегрессии, скользящего среднего и смешанного процесса.

В то время как автокорреляционная функция процесса авторегрессии порядка p спадает плавно, ее частная автокорреляционная функция имеет обрыв после p-й задержки. Обратно, автокорреляционная функция процесса скользящего среднего порядка q обрывается после задержки q, в то время как ее частная автокорреляция плавно спадает с ростом задержки. Далее, автокорреляционная функция смешанного процесса, содержащая компоненту авторегрессии порядка p и компоненту скользящего среднего порядка q, после первых q - p задержек представляется в виде суммы экспонент и затухающих синусоид. Обратно, частная автокорреляционная функция смешанного процесса приближенно представляется суммой экспонент и затухающих синусоид после p - q задержек.

В общем, поведение автокорреляционной функции процесса авторегрессии похоже на поведение частной автокорреляционной функции процесса скользящего среднего и наоборот. Например, автокорреляционная функция процесса авторегрессии первого порядка экспоненциально затухает, в то время как частная автокорреляционная функция обрывается после первой задержки. Соответственно для процесса скользящего среднего первого порядка автокорреляционная функция обрывается после первой задержки. Частная автокорреляционная функция не точно экспоненциальная, но в ней преобладают экспоненциальные члены, определяющие ее общий вид.

Особенно важны процессы авторегрессии и скользящего среднего первого и второго порядков и простой смешанный процесс (1, *d*, 1).

2.2.3 Проверка адекватности подученных моделей

После идентификации модели и оценки параметров подгоняемая модель подвергается диагностической проверке. Нужно ответить на вопрос, адекватна ли модель. Ели будут обнаружены свидетельства серьезной неадекватности, возникнет необходимость узнать, как нужно изменить модель на следующем итеративном цикле.

Для этого выбран метод, заключающийся в анализе ошибок оценивания, в виде вычисления статистики Дурбина-Уотсона:

$$DW = 2 - 2\rho$$

где ρ - коэффициент корреляции между значениями случайной переменной, т.е. $\rho = \mathrm{cov}[\varepsilon(k)] = E[\varepsilon(k)\varepsilon(k-1)]$. При полном отсутствии корреляции между ошибками $DW = 2$.

Для оценки адекватности модели дополнительно можно использовать статистику Льюнга-Бокса-Пирса. Для первых известных m автокорреляций: $\rho_1,\ldots,\rho_m$ статистика Льюнга-Бокса-Пирса имеет вид:

$$Q(\rho) = T(T+2)\sum_{k=1}^{m}\frac{\rho_k^2}{T-k}$$

и асимптотическое распределение χ_m^2. Нулевая гипотеза в Q-критерии заключается в том, что случайный процесс представляет собой белый шум. Используем стандартную процедуру проверки: если расчетное значение Q-статистики больше заданного квантиля распределения χ_m^2, то нулевую гипотезу отвергаем и признаем наличие автокорреляции до m-го порядка в исследуемом процессе.

Если же есть сразу несколько «достаточно хороших», адекватных моделей, то нужно выбрать одну из них, исходя из принципа «экономии»: выбрать модель с наименьшим числом параметров.

Для этого можно использовать два теста:

Информационный критерий Акаике AIC (*Akaike information criterion*, Akaike, 1973):

$$AIC = 2 \cdot \frac{p+q}{n} + \ln\left(\frac{\sum_{t=1}^{n} e_t^2}{n}\right)$$

Из двух моделей следует выбрать ту, у которой AIC меньше.

Байесовский информационный критерий Шварца BIC (*Schwarz criterion*, Schwarz, 1978):

$$BIC = \frac{(p+q)\ln n}{n} + \ln\left(\frac{\sum_{t=1}^{n} e_t^2}{n}\right)$$

Критерий Шварца предпочтительнее, поскольку он имеет свойство выбирать истинную модель «почти наверняка».

2.2.4 Задания

Для анализа поведения временного ряда:

1. Преобразовать ряд к стационарному виду.
2. Идентифицировать модель, т.е. подобрать порядок p и q модели.
3. Оценить параметры модели.
4. Проверить адекватность модели.

3 СОДЕРЖАНИЕ И ОФОРМЛЕНИЕ ПОЯСНИТЕЛЬНОЙ ЗАПИСКИ

Заключительным этапом выполнения работы являются оформление пояснительной записки, графического материала и защита студентом курсовой работы в соответствии с утвержденным индивидуальным заданием.

3.1 Структура пояснительной записки

Пояснительная записка к курсовому проекту выполняется на стандартных листах формата А4 с одной стороны листа и должна иметь следующую структуру:

1	Титульный лист
2	Индивидуальное задание на выполнение работы
3	Содержание
4	Введение
5	Постановка задачи
6	Содержательная часть работы по отдельным разделам индивидуального задания
7	
8	
9	
10	
11	Заключение
12	Библиографический список
13	Приложения

3.2 Содержание разделов пояснительной записки

Титульный лист является первым листом пояснительной записки, с него начинается нумерация страниц, но номер 1 на титульном листе не ставится. Образец титульного листа приведен в приложении А.

Индивидуальное задание на выполнение работы представляет специальный бланк, который необходимо заполнить. Образец задания приведен в приложении Б.

В содержании перечисляют наименования разделов пояснительной записки и номера соответствующих страниц.

Во введении указывают цели и задачи курсового проектирования.

В разделе «Постановка задачи» приводят постановку задачи на курсовое проектирование.

В содержательной части работы по отдельным разделам, соответствующим индивидуальному заданию, описывают алгоритмы решения задач (отмечая их достоинства и недостатки) с основными и промежуточными преобразованиями и подстановкой числовых значений, а также приводят результаты проверки правильности решения.

В разделе “Заключение” должны быть сделаны основные выводы по результатам курсового проектирования.

В разделе "Библиографический список" указывают перечень научно-технических публикаций, нормативных технических документов и других научно-технических материалов, на которые есть ссылки в основном тексте. Если на источник нет ссылки из основного текста, его в список не помещают. Список должен быть приведен в конце текста пояснительной записки, начиная с новой страницы. Библиографические описания в перечне ссылок приводят в порядке, в котором они впервые упоминаются в тексте. Порядковые номера описаний в перечне являются ссылками в тексте (номерные ссылки). Библиографические описания ссылок в перечне приводят в соответствии с действующими стандартами по библиотечному и издательскому делу.

В приложение к пояснительной записке могут быть включены таблицы, обоснования, методики, расчеты, блок-схемы алгоритмов программ, графические материалы и другие документы, использованные при выполнении курсовой работы.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1.Box G.E. Time Series Analysis, Forecasting and Control / G.E. Box., G.M. Jenkins, G.C. Reinsel. — New Jersey: Prentice Hall, 1994. — 598 p.
2. W. Enders. Applied Econometric Time Series / W. Enders. — New York: John Wiley & Sons, 1995 — 433 p.
3. Суслов В.И. Эконометрия / В.И. Суслов, Н.М. Ибрагимов, Л.П. Талышева, А.А. Цыплаков. — Новосибирск: Издательство СО РАН, 2005. — 744с.
4. Бидюк П.И. Системный подход к построению регрессионной модели по временным рядам / П.И. Бидюк, И.В. Баклан, В.Н. Рифа // Системні дослідження та інформаційні технології, №3, 2002. — С.114 — 131.
5. Первухина Е. Л. Анализ нестационарных случайных процессов в задачах автоматизации производственных испытаний машиностроительных изделий / Е.Л. Первухина, В.В. Голикова // Сборка в машиностроении, приборостроении – Москва, 2007. – №8— С.29 – 35.