МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
Национальный исследовательский ядерный университет «МИФИ»

Институт

интеллектуальных кибернетических систем

Кафедра кибернетики (№ 22)

Направление подготовки 09.03.04 Программная инженерия

Расширенное содержание пояснительной записки

к учебно-исследовательской работе студента на тему:

Разработка программной системы для исследования авторегрессионных моделей временных рядов

Группа	Б17-514		
Студент		____________ (подпись)	Фирсов Г.В. (ФИО)
Руководитель	____________ (0-5 баллов)	____________ (подпись)	Трофимов А.Г. (ФИО)
Научный консультант	____________ (0-5 баллов)	____________ (подпись)	____________ (ФИО)

Москва 2020

Реферат

Пояснительная записка содержит — __ страницы, __ рисунков, __ таблиц, __ ссылок на источники.

Ключевые слова: временной ряд, авторегрессионная модель, прогнозирование.

Целью данной учебно-исследовательской работы является разработка программного комплекса для исследования, моделирования и прогнозирования временных рядов с помощью авторегрессионных моделей.

Содержание

Введение

Временными рядами можно описать множество процессов в различных сферах человеческой жизни, например, это могут быть курсы валют, цен на акции компаний, число случаев заражения какими-либо заболеваниями, объемы потребления продукции, объемы производства продукции, какие-ибо метеорологические данные, такие, как погода, и много другое. Зачастую при работе с данными, представленными временными рядами, возникает задача прогнозирования наблюдаемого процесса на некоторое количество измерений вперед. Такая задача, например, важна при игре на бирже, ведь если клиент будет знать курс определенных акций наперед с определенной степенью уверенности, то у него будет больше преимуществ перед конкурентами, так как он сможет построить свои действия в соответствии с прогнозом и в большем числе случаев это сыграет на руку.

Одним из классов моделей, применяемых в задаче прогнозирования временных рядов, является класс авторегрессионных моделей. Данные модели достаточно широко распространены в практической деятельности ввиду их относительной простоты и в то же время их возможности моделировать всевозможные ряды, которые могут содержать сезонности и тренды. Помимо этого, ещё одним достоинством данных моделей является относительно низкая вычислительная сложность их построения.

Целью данной работы является изучение авторегрессионных моделей временных рядов, а также разработка программного комплекса для исследования временных рядов на применимость авторегрессионных моделей, построения таких моделей и прогнозирования значений временных рядов.

1. Обзор методов моделирования и прогнозирования временных рядов

В данном разделе приводится обзор подходов к моделированию и прогнозированию временных рядов, даётся описание авторегрессионных моделей и анализируются результаты и эффективность применения последних на практике.

1.1. Анализ подходов к моделированию временных рядов

Существует несколько подходов к моделированию временных рядов. Примерами могут послужить авторегрессионные модели и модели скользящего среднего, рассматриваемые в пункте 1.1.1, нейросетевые модели из пункта 1.1.2 и модели на основе экспоненциального сглаживания, описанная в пункте 1.1.3. Конечно, набор моделей, применяемых в анализе и прогнозировании временных рядов, не ограничивается только этими тремя классами. Их круг много шире, так, например, существуют модели на основе опорных векторов (SVM-модели) [12], т.е. адаптированные к работе с временными рядами алгоритмы SVM-регрессии; временные ряды можно моделировать и с помощью цепей Маркова [13], которая применима к моделированию временных рядов ввиду того свойства, что она представляет собой последовательность случайных величин, распределение последней из которых зависит лишь от нескольких предыдущих значений (в зависимости от порядка цепи Маркова). Используются также и регрессионные деревья [14]. Помимо указанных существуют также и другие модели, но целью данной работы не является их полное перечисление, сосредоточимся в дальнейшем на трёх наиболее популярных и хорошо изученных моделях.

1.1.1 Авторегрессионные модели

Суть авторегрессионной модели состоит в том, что значение в момент времени t является линейной комбинацией значений в моменты времени t-1, t-2, …, t-p. Такая модель называется авторегрессионной моделью порядка p и обычно обозначается как AR(p). Математически это можно выразить следующим образом: $X_t = \phi_1 X_{t-1} + \dots + \phi_p X_{t-p} + \epsilon_t$, где $M[X_t] = 0$, а все коэффициенты $\phi_1, \dots \phi_p$ – константы, а ϵ_t – белый шум с нулевым математическим ожиданием и конечной дисперсией. Стоит отметить, что на коэффициенты модели для обеспечения стационарности необходимо наложить некоторые ограничения, причем с ростом параметра p вид этих ограничений усложняется.

Модель скользящего среднего предполагает, что значение временного ряда в момент времени t является линейной комбинацией значений белого шума в моменты времени t, t-1, …, t-q, где q называется порядком модели. Обозначение: MA(q). В математической записи это выглядит следующим образом: $X_t = \epsilon_t + \theta_1 \epsilon_{t-1} + \dots + \theta_q \epsilon_{t-q}$, где $\epsilon_t, \epsilon_{t-1}, \dots, \epsilon_{t-q}$ – независимые одинаково распределенные случайные величины с нулевым математическим ожиданием и конечной дисперсией (белый шум), а коэффициенты $\theta_1, \dots, \theta_q$ – константы. На эти коэффициенты также накладываются усложняющиеся с ростом порядка модели ограничения, которые необходимы для обеспечения обратимости, т.е. свойства, позволяющего настроить модель по имеющимся данным.

На основе вышеуказанных простых моделей строятся более сложные, такие, как, например, модель ARMA(p, q), которая представляет значение X_t в виде суммы выражений для X_t в AR(p) и MA(q) моделях, т.е.: $\mathrm{X_t} = \sum_{\mathrm{i=1}}^{\mathrm{p}} \phi_{\mathrm{i}}\, \mathrm{X_{t-i}} + \sum_{\mathrm{j=0}}^{\mathrm{q}} \theta_{\mathrm{j}}\, \epsilon_{\mathrm{t-j}}$. Модель ARMA может прогнозировать только стационарные временные ряды. Кроме того, существует модель ARIMA(p, d, q), которая говорит о том, что разность ряда порядка d подчиняется модели ARMA(p, q). Эта модель является расширением модели ARMA(p, q) на нестационарные временные ряды.

1.1.2 Нейросетевые модели

Целевые функции моделей AR, MA, ARMA, ARIMA и т.д. можно воспроизвести на нейронных сетях: как рекуррентных, так и прямого распространения. Возьмём, к примеру, такую функцию: $\mathrm{X}_t = h(X_{t-1}, \dots, X_{t-p}) + \epsilon_t$, которая представляет собой прямое обобщение целевой функции модели AR(p) на класс нелинейных функций (такая модель называется NAR(p) – nonlinear autoregressive). Тогда нейросетевой аппроксимацией данной функции будет следующее выражение: $\hat{\mathrm{X}}_{\mathrm{t}} = \hat{\mathrm{h}}(\mathrm{X_{t-1}}, \dots, \mathrm{X_{t-p}}) = \beta_0 + \sum_{\mathrm{i=1}}^{\mathrm{l}} \beta_{\mathrm{i}} \mathrm{f}\left(\alpha_{\mathrm{i}} + \sum_{\mathrm{j=1}}^{\mathrm{p}} \omega_{\mathrm{ij}}\, \mathrm{X_{t-j}}\right)$, где f – функция активации, а $\Theta = (\beta_0, \dots, \beta_l, \alpha_1, \dots, \alpha_p, \omega_{11}, \dots, \omega_{lp})^T$ – вектор гиперпараметров нейросетевой модели. Такая модель обозначается как ARNN(p). Обобщение модели ARMA(p, q) на класс нелинейных функций также можно воспроизвести на нейронной сети, но уже на рекуррентной: $\hat{X}_t = \alpha_0 + \sum_{j=1}^{h} \alpha_j g\left(\beta_{0j} + \sum_{i=1}^{p} \beta_{ij} X_{t-i} + \sum_{i=p+1}^{p+q} \beta_{ij} \hat{\epsilon}_{t+p-i}\right)$.

Нейросетевые модели не предполагают линейности целевой функции, поэтому разумно будет сравнивать их как раз с обобщениями линейных авторегрессионных моделей на класс нелинейных функций (например, с уже упомянутой NAR или NARMA(p, q) – обобщение для ARMA(p, q)). Так, например, в статье [4] приводится сравнение рекуррентной нейронной сети для модели NARMA(p, q) и нейросетевой модели ARNN(p) для моделирования и прогнозирования нелинейного временного ряда. В заключении указывается, что обе модели не способны полностью отразить поведение временного ряда с компонентой скользящего среднего (MA-компонента), хотя там же говорится, что нейросетевые модели являются более подходящими для моделирования нелинейных временных рядов с MA-компонентой.

1.1.3. Модели экспоненциального сглаживания

Модели экспоненциального сглаживания имеют характерную особенность – при вычислении значения S_t наибольший вес имеет значение X_t, в то время как веса предыдущих значений экспоненциально уменьшаются. Рассмотрим, например, так называемое простое экспоненциальное сглаживание, которое задается соотношением: $S_t = \alpha X_t + (1 - \alpha) S_{t-1}$. Стоит отметить, что значение X_t не равно значению S_t (за исключением вырожденного случая, когда параметр α равен единице), так как каждое новое рассчитанное значение является взвешенным средним между текущим наблюдением и предыдущим уже посчитанным значением, но это и кроется за словом «сглаживание», т.е. модель даёт более плавный, гладкий временной ряд на выходе. Параметр α принимает значения в промежутке [0; 1] и чем ближе он находится к единице, тем меньше вклад предыдущего значения, и тем менее сглаженным получается ряд на выходе модели. Данная модель применима к стационарным временным рядам. Если в ряде присутствует тренд, то необходимо применять двойное экспоненциальное сглаживание, отличие которого от простого, состоит в том, что каждое новое значение складывается из значений уровня и трендовой составляющей. Такая модель задаётся двумя соотношениями: $S_t = \alpha X_t + (1-\alpha)(S_{t-1} + tr_{t-1})$, $tr_t = \beta(S_t - S_{t-1}) + (1-\beta)tr_{t-1}$, где S_t – вычисляемое значение, а tr_t – трендовая составляющая. Параметры α и β так же принадлежат промежутку $[0;\ 1]$. Для рядов с сезонностью существует также тройное экспоненциальное сглаживание, которое эту сезонность учитывает и добавляет, в свою очередь, третий сглаживающий параметр γ.

Данные модели могут применяться для моделирования временных рядов, тренд и сезонность которых могут меняться со временем, в то время как авторегрессионные модели (ARMA) требуют определённых преобразований ряда для того, чтобы быть применимыми. Дополнительным преимуществом может служить простота данной модели, что может дать некоторый прирост в качестве по сравнению с более сложными авторегрессионными моделями на практике.

1.2. Обзор методов прогнозирования временных рядов

На основе моделей, приведенных выше, а также некоторых схожих им, но не упомянутых явно, можно прогнозировать значения временного ряда в будущие моменты времени. Но для того, чтобы прогнозировать будущие значения временного ряда, сначала надо построить саму модель, то есть подобрать её коэффициенты (или гиперпараметры). В данном разделе пойдёт речь о том, как строятся модели временных рядов и какие особенности возникают при прогнозировании с использованием данных моделей.

1.2.1. Прогнозирование на основе авторегрессионных моделей

В данном контексте представляется важной теорема, которая гласит о том, что любой стационарный временной ряд может быть представлен моделью ARMA(p, q) с любой наперед заданной точностью. Для построения модели необходимо подобрать ее гиперпараметры – в нашем случае это коэффициенты при значениях ошибок и предыдущих значениях временного ряда, а также параметры p, d и q. Все коэффициенты в линейных комбинациях подираются по методу наименьших квадратов при фиксированных p, d и q. Параметр d подбирается как наименьшее количество дифференцирований ряда, необходимое для обеспечения его стационарности, которую можно проверить по критерию Дики-Фуллера. Параметры p и q для модели ARMA можно подобрать по критерию Акаике, Байесовскому критерию или критерию Ханнана-Куинна, которые, в сущности, представляют собой минимизацию логарифма функции правдоподобия с добавленным слагаемым специального вида. Кроме того, параметры p и q по отдельности (т.е. для модели AR или MA) можно оценить по графику функций автокорреляции (ACF) и частичной автокорреляции (PACF), так, оценка q – последний значимый несезонный лаг на графике ACF, а оценка p – тоже последний значимый несезонный лаг, но уже на графике PACF.

В данном разделе также стоит упомянуть наличие сезонных авторегрессионных моделей. Одной из наиболее общих сезонных является модель SARIMA(p, d, q)×(P, D, Q), в которой добавляется D сезонных дифференцирований, P сезонных авторегрессионных компонент и Q сезонных компонент скользящего среднего. Параметры P, D и Q определяются практически аналогично параметрам p, d и q, но с тем отличием, что производится сезонное дифференцирование (для D) и выбираются последние значимые сезонные лаги на PACF и ACF (для P и Q соответственно).

Способ же построения прогноза вполне понятен – необходимо посчитать значение X_{t+1} на основе имеющихся данных, а затем использовать этот прогноз для предсказания последующих значений в моменты времени t+2, t+3 и так далее. Есть важный момент: так как математическое ожидание $\epsilon_{t+\tau}$ равно нулю, то все ошибки, начиная с $\tau = 1$ заменяются нулями, а ошибки «из прошлого» меняются на остатки, т.е. разности фактического значения и смоделированного.

1.2.2. Прогнозирование на основе нейросетевых моделей

Искусственная нейронная сеть характеризуется набором гиперпараметров, в который входят синаптические коэффициенты нейронов и смещения. Подбор этих коэффициентов (или обучение нейронной сети) можно осуществить, например, методом градиентного спуска. Вводится так называемая функция потерь $L(\Theta)$ – фактически метрика качества построенной нейросетевой модели. Метод градиентного спуска служит для поиска минимума этой функции. Существует множество модификаций данного метода, как, к примеру, сопряженный градиентный спуск, градиентный спуск Нестерова и другие. Для вычисления частных производных применяется метод обратного распространения ошибки, но несмотря на это обучение нейронной сети всё равно обходится вычислительно дороже, нежели построение авторегрессионной модели или модели на основе экспоненциального сглаживания.

Ввиду того свойства нейронных сетей, что характер аппроксимируемой функции не играет значения, их можно применять для прогнозирования нелинейных временных рядов. В то же время такие модели дают большую точность при прогнозировании на длинной дистанции в сравнении с моделями экспоненциального сглаживания.

1.2.3. Прогнозирование на основе модели экспоненциального сглаживания

Подбор коэффициентов модели можно осуществить, например, при помощи методов оптимизации: на единичном k-мерном кубе $[0;1]^k$, где k – количество коэффициентов модели, минимизируется оценка MAPE, т.е. среднее абсолютное отклонение в процентах (такая оценка выбрана в работе [5]), хотя можно выбрать и иные метрики. Существуют и другие методы подбора параметров, например, перебор всевозможных комбинаций коэффициентов на «решетке», представляющей из себя k-мерную матрицу, в ячейках которой стоят как раз эти комбинации коэффициентов.

Особенностью прогноза, сделанного с помощью модели экспоненциального сглаживания является то, что он достаточно хорошо оценивает значение временного ряда на короткой дистанции, то есть данный метод стоит применять для прогноза на 4-6 наблюдений вперёд, а вот более длинные прогнозы будут терять свою точность. Сам же процесс прогнозирования вновь достаточно очевиден – в качестве нового значения берётся вычисленное по формулам, указанным в разделе 1.1.3., значение S_{t+1}, которое впоследствии используется для дальнейших прогнозов.

1.3. Анализ результатов применения авторегрессионных моделей для решения различных прикладных задач прогнозирования

Авторегрессионные модели получили широкое распространения при прогнозировании временных рядов во множестве сфер как научной деятельности, так и прикладной. Так, например, с помощью таких моделей очень хорошо описываются объемы продаж каких-либо товаров. Кроме того, с помощью упомянутых выше моделей AR(p) и ARIMA(p, d, q) хорошо прогнозируются курсы валют. Стоит отметить, что параметры p, d, и q на практике редко превышают 2, так, в частности, курс одного доллара США к рублю можно спрогнозировать при помощи модели ARIMA(1, 1, 0).

Из относительно свежих исследований, в который применялась авторегрессионная модель временного ряда, можно выделить изучение случаев заражения вирусом COVID-2019 в Италии в целом и в трех ее регионах – Марке, Венеции и Ломбардии – в отдельности [8]. Для каждого временного ряда выбиралась своя модель из множества моделей ARIMA, наиболее точно описывающая ряд. Для подбора гиперпараметров применялся критерий Акаике (подбор p и q), а также критерий Дики-Фуллера для выбора d. В итоге выбранными моделями оказались следующие: ARIMA(4, 2, 2) для Италии в целом, ARIMA(0, 2, 2) для региона Марке, ARIMA(6, 2, 1) для Ломбардии и ARIMA(0, 2, 2) для Венеции. Все эти модели показали снижение практически к нулю количества случаев заражения приблизительно через 2 месяца после начала эпидемии или спустя примерно две трети месяца после начала исследования. Факт правильности ещё потребует подтверждения, но, как указано в статье [8], графики автокорреляционных функций для остатков рядов показывают, что модель достаточно хорошо описывает поведение временного ряда, то есть остатки не являются коррелированными.

Иным примером использования авторегрессионных моделей временных рядов на практике (и тоже в медицине) является моделирование и прогнозирование случаев госпитализации в связи с заболеванием пневмонией в провинции Чангмай в Таиланде [6]. Указанная статья является примером некого расхождения терминологии: в исследовании используется сезонная модель SARIMA(p, d, q)×(P, D, Q), но в тексте статьи она упоминается как ARIMA(p, d, q)(P, D, Q). Здесь будет использоваться обозначение SARIMA. То, что используется сезонная модель вытекает из практически очевидной сезонности данных (из года в год шаблон повторяется), хотя в данных также присутствует тренд. Исходный временной ряд был прологарифмирован для стабилизации дисперсии, так как данные изначально обладали свойством гетероскедастичности. Для оценки гиперпараметров модели использовался критерий Акаике, а параметры d и D были подобраны по тому, как хорошо модель аппроксимирует данные, и в итоге оказались равными нулю – лучшей моделью оказалась модель SARIMA(1, 0, 2)×(2, 0, 0) с учетом также внешних факторов, таких, как уровень загрязнения воздуха, климатические изменения. В итоге построенная модель с достаточной точностью предсказывает количество случаев госпитализации.

Ещё одним примером применения авторегрессии для прогнозирования может служить исследование об объемах урожая и производства зерна в Индии [9]. Исходные данные имеют ярко выраженный линейный тренд, который устраняется дифференцированием ряда один раз, после этого стационарность ряда подтверждается критерием Дики-Фуллера. Для выбора значений p и q для модели ARIMA(p, d=1, q) в исследовании используются графики ACF и PACF, по которым начальные значения оцениваются как 1 и 1. Далее модели со значениями p и q, принимающими значения 0 или 1, были протестированы на продифференцированных рядах объема производства и урожая, в результате чего для прогнозирования лучше всего подошла модель ARIMA(0, 1, 1). В подтверждение того, что модель хорошо описывает данные, для остатков был выполнен Q-тест Льюнга-Бокса, который показал отсутствие каких-либо корреляций в них. Данное исследование фактически шире, нежели применение модели ARIMA для прогнозирования. Помимо неё были построены нейросетевые модели ARNN(3, 4) и ARNN(4, 3) для моделирования производства и урожая соответственно. В итоге для каждого ряда получилась своя гибридная модель, авторегрессионная часть которой хорошо воспроизводит линейную составляющую исходного ряда, а нелинейная составляющая прогнозируется нейросетевой частью. Данная модель позволяет предсказывать урожай и объем производства приблизительно на 9-10 лет вперед.

2. Моделирование и прогнозирование временных рядов с помощью авторегрессионных моделей

В данном разделе подробно описываются авторегрессионные модели для прогнозирования временных рядов, подробнее описываются методы построения моделей программно, а также даются некоторые методы оценивания качества прогноза модели, например, на основе визуального и статистического анализа остатков.

2.1. Методика моделирования и прогнозирования временных рядов

Данный подраздел включает в себя подробное описание модели ARIMA, моделей, являющихся «подмножествами» последней, а также схожих с ней. Приводятся методы подбора гиперпараметров моделей.

2.2. Методы декомпозиции временных рядов

В этом подразделе излагается теория об аддитивной и мультипликативной моделях декомпозиции временного ряда на составляющие и программные методы осуществления этой декомпозиции.

2.3. Статистические тесты на стационарность временных рядов

Подраздел целиком посвящен описанию статистических тестов, позволяющих определить, стационарен ли данный временной ряд или нет.

2.4. Методика структурной и параметрической идентификации моделей временных рядов

Подраздел включает описание методов структурной и параметрической идентификации моделей временных рядов.

2.5. Методика оценивания качества моделирования и прогнозирования временных рядов

В подразделе приводятся критерии оценивания качества построенных моделей временных рядов.

3. Разработка программной системы для исследования авторегрессионных моделей временных рядов

Данный раздел посвящен реализации программного комплекса для моделирования и прогнозирования временных рядов с помощью авторегрессионных моделей на языке программирования Python.

3.1. Выбор средств программной реализации

Подраздел представляет собой перечень инструментов, выбранных для реализации программной системы с кратким их описанием.

3.2. Архитектура и требования к программной системе

В данном подразделе приводится описание архитектурного решения и требований к проектируемой программной системе.

3.3. Программная реализация модулей системы

Подраздел посвящен результатам разработки программной системы и пояснениям о деталях реализации.

Заключение

В заключении будут изложены результаты проделанной работы – как аналитической и теоретической, так и практической с описанием особенностей изучаемых моделей, а также разработанной программной системы.

Список литературы

1. Robert H. Shumway. Time Series Analysis and Its Applications: With R Examples / Robert H. Shumway, David S. Stoffer, Springer, 2017. - 562 p.

2. Handbook on Seasonal Adjustment. – Luxembourg: Publications Office of the European Union, 2018. – 830 p.

3. Rob J Hyndman. Forecasting: Principles & Practice. – Perth: University of Western Australia, 2014. – 137 p.

4. J. Lee, “Univariate time series modeling and forecasting (Box-Jenkins Method)”, Econ 413, lecture 4

5. Ahmed Tealab. Forecasting of nonlinear time series using ANN / Ahmed Tealab, Hesham Hefny, Amr Badr, Future Computing and Informatics Journal, 2017, pp. 39-47.

6. J.M. Kihoro, R.O. Otieno, C. Wafula, “Seasonal Time Series Forecasting: A Comparative Study of ARIMA and ANN Models”, African Journal of Science and Technology (AJST) Science and Engineering Series Vol. 5, No. 2, pages: 41-49

7. Prajakta S. Kalekar. Time series Forecasting using Holt-Winters Exponential Smoothing / Kanwal Rekhi School of Information Technology Journal, 2004, p. 13.

8. Ruchiraset, A. Time series modeling of pneumonia admissions and its association with air pollution and climate variables in Chiang Mai Province, Thailand. / Ruchiraset, A., Tantrakarnapa, K., Environ Sci Pollut Res 25, 2018, 33277–33285

9. Хайкин С. Нейронные сети: полный курс, 2-е издание – М. Издательский дом «Вильямс», 2006. – 1104 с.

10. Gaetano Perone. An ARIMA model to forecast the spread of COVID-2019 epidemic in Italy / University of Bergamo, 2020

11. Puneet Dheer. Time series modelling for forecasting of food grain production and productivity of India / Journal of Pharmacognosy and Phytochemistry 8(3), 2019, pp. 476-482

12. Loris Foresti. Time Series Input Selection using Multiple Kernel Learning / Loris Foresti, Devis Tuia, Vadim Timonin, Mikhail Kanevski, ESANN 2010 proceedings, European Symposium on Artificial Neural Networks - Computational Intelligence and Machine Learning, 2010

13. Antoni Wilinski. Time series modeling and forecasting based on a Markov chain with changing transition matrices / Koszalin University of Technology, Faculty of Electronics and Computer Science, Sniadeckich 2, 2017, pp. 75-453

14. Yisehac Yohannes. Classification and regression trees, CART a user manual for identifying indicators of vulnerability to famine and chronic food insecurity / Yisehac Yohannes, Patrick Webb, International food policy research institute, 1999, - 50 p.