Московский Государственный Технический Университет им. Н.Э. Баумана

На правах рукописи

Чучуева Ирина Александровна

МОДЕЛЬ ПРОГНОЗИРОВАНИЯ ВРЕМЕННЫХ РЯДОВ ПО ВЫБОРКЕ МАКСИМАЛЬНОГО ПОДОБИЯ

Специальность 05.13.18 – Математическое моделирование, численные методы и комплексные программы

Диссертация на соискание ученой степени
кандидата технических наук

Научный руководитель:
доктор технических наук, профессор
Павлов Юрий Николаевич

Москва – 2012

Посвящается бабушке Л.В. Чучуевой и дедушке А.А. Чучуеву (1933 – 2003)

ОГЛАВЛЕНИЕ

Стр.

Стр.

ВВЕДЕНИЕ

Актуальность темы. Задача прогнозирования будущих значений временного ряда на основе его исторических значений является основой для финансового планирования в экономике и торговле, планирования, управления и оптимизации объемов производства, складского контроля [1,2].

В настоящее время компаниями осуществляется накопление исторических значений экономических и физических показателей в базах данных, что существенно увеличивает объемы входной информации для задачи прогнозирования. Вместе с тем, развитие аппаратных и программных средств предоставляет все более мощные вычислительные платформы, на которых возможна реализация сложных алгоритмов прогнозирования. Кроме того, современные подходы к экономическому и техническому управлению предъявляют все более жесткие требования к точности прогнозирования. Таким образом, задача прогнозирования временных рядов усложняется одновременно с развитием информационных технологий.

В настоящее время задача прогнозирования различных временных рядов актуальна и является неотъемлемой частью ежедневной работы многих компаний.

Задача прогнозирования временного ряда решается на основе создания модели прогнозирования, адекватно описывающей исследуемый процесс.

На сегодняшний день существует множество моделей прогнозирования временных рядов: регрессионные и авторегрессионные модели, нейросетевые модели, модели экспоненциального сглаживания, модели на базе цепей Маркова, классификационные модели и др. Наиболее популярными и широко используемыми являются классы авторегрессионных и нейросетевых моделей [3]. Существенным недостатком авторегрессионного класса является большое число свободных параметров, идентификация

которых неоднозначна и ресурсоемка [4]. Существенным недостатком класса нейросетевых моделей является недоступность промежуточных вычислений, выполняющихся в «черном ящике», и, как следствие, сложность интерпретации результатов моделирования. Кроме того, еще одним недостатком данного класса моделей является сложность выбора алгоритма обучения нейронной сети [5].

Диссертация посвящена разработке новой авторегрессионной модели прогнозирования, которая имеет сравнимую с другими моделями эффективность прогнозирования различных временных рядов и при этом устраняет основной и наиболее существенный недостаток авторегрессионого класса моделей — большое число свободных параметров.

Целью работы является разработка новой модели и соответствующего ей метода прогнозирования, относящейся к классу авторегрессионных моделей и устраняющей основной недостаток данного класса моделей — большое число свободных параметров. Новая модель и соответствующий ей метод должны иметь высокую скорость вычисления прогнозных значений и сравнимую с другими моделями точность прогнозирования различных временных рядов.

Для достижения этой цели были поставлены и решены следующие **задачи**.

1. Осуществить обзор моделей и методов прогнозирования временных рядов, выявить достоинства и недостатки каждого класса моделей. Выявить наиболее используемые классы моделей прогнозирования и их основные недостатки, определить перспективные подходы, позволяющие устранить недостатки авторегрессионного класса моделей.

2. Разработать новую модель прогнозирования временных рядов, устраняющую указанный недостаток авторегрессионного класса моделей.

3. Разработать новый метод прогнозирования на основании предложенной модели и выполнить программную реализацию алгоритмов.

4. Оценить эффективность предложенной модели прогнозирования при решении задачи прогнозирования различных временных рядов.

Методы исследования. При решении поставленных задач в работе использованы методы математического моделирования, анализ временных рядов, регрессионный анализ, методы объектно-ориентированного программирования.

Научная новизна. В диссертации получены следующие основные результаты, которые выносятся на защиту.

1. Модель экстраполяции временных рядов по выборке максимального подобия, относящаяся к классу авторегрессионных моделей и имеющая единственный параметр.

2. Метод прогнозирования временных рядов на основании разработанной модели, содержащий набор алгоритмов для экстраполяции временных рядов, идентификации модели и построения доверительного интервала прогнозных значений.

3. Результаты прогнозирования временных рядов показателей энергорынка РФ, а также временных рядов из других предметных областей, подтверждающие эффективность разработанной модели.

Достоверность и обоснованность выносимых на защиту результатов прогнозирования показателей энергорынка РФ документально подтверждается ЗАО «РусПауэр», использующего разработанные алгоритмы на ежедневной основе. Достоверность результатов прогнозирования временного ряда уровня сахара крови человека, больного диабетом первого типа, обеспечивается строгостью применяемого математического аппарата и подтверждается приведенным сравнительным анализом. Достоверность

результатов прогнозирования скорости движения транспорта по г. Москва обеспечивается условиями открытого конкурса, проводимого компанией «Яндекс». Результаты конкурса опубликованы в открытом доступе по адресу http://imat2010.yandex.ru/results.

Практическая ценность работы. Разработанная модель и метод прогнозирования по выборке максимального подобия могут применяться для прогнозирования временных рядов различных предметных областей. Разработанные алгоритмы экстраполяции временных рядов с учетом и без учета внешних факторов наглядны для программной реализации. Скорость вычисления прогнозных значений при использовании модели высока. Задача идентификации модели упрощена в сравнении с другими моделями авторегрессионного класса.

Реализация и внедрение результатов работы. Результаты работы реализованы по заказу Закрытого акционерного общества «РусПауэр» в виде серверного приложения для прогнозирования показателей энергорынка РФ на ежедневной основе. Приложение работает в автоматическом режиме и предоставляет прогнозные значения показателей без вмешательства эксперта.

Апробация работы. Основные результаты диссертационной работы докладывались на I Международной научно-практической конференции ученых, аспирантов и студентов «Наука и современность 2010» (Новосибирск, 2010); на научно-технической конференции «Студенческая научная весна» (Москва, 2010); на III Международной конференции «Математическое моделирование социальной и экономической динамики (MMSED-2010)» (Москва, 2010).

Публикации. Основные результаты диссертации опубликованы в 8 научных статьях, в том числе в 5 статьях, опубликованных в журналах из Перечня рецензируемых ведущих научных журналов и изданий, и 2 тезисов

докладов.

Личный вклад соискателя. Все исследования, результаты которых изложены в диссертационной работе, получены лично соискателем в процессе научных исследований. Из совместных публикаций в диссертацию включен лишь тот материал, который непосредственно принадлежит соискателю.

Структура и объем работы.

Диссертационная работа состоит из введения, четырёх глав, заключения, списка литературы и приложения, занимающих 154 страниц текста, в том числе 33 рисунка на 29 страницах, 37 таблиц на 29 страницах, список использованной литературы из 75 наименования на 10 страницах.

В первой главе сформулирована постановка задачи прогнозирования временного ряда. Рассмотрены существующие классы моделей прогнозирования, установлены достоинства и недостатки каждого класса. В результате обзора моделей прогнозирования выявлен основной недостаток авторегрессионного класса моделей и определены перспективные подходы, позволяющие его устранение.

Во второй главе диссертации описаны две модели экстраполяции по выборке максимального подобия для двух видов постановок задачи. Новая модель экстраполяции имеет единственный параметр и устраняет основной недостаток авторегрессионного класса моделей.

В третьей главе сформулирован метод прогнозирования временных рядов на основании предложенной модели экстраполяции, содержащий набор алгоритмов для экстраполяции временных рядов, идентификации модели и построения доверительного интервала прогнозных значений.

В четвертой главе диссертации описана программная реализация предложенной модели экстраполяции для решения задач прогнозирования

показателей энергорынка РФ. В главе приведены результаты прогнозирования различных временных рядов. Проведен сравнительный анализ достигнутых оценок точности и доказана высокая эффективность разработанной модели для прогнозирования различных процессов.

ГЛАВА 1. ПОСТАНОВКА ЗАДАЧИ И ОБЗОР МОДЕЛЕЙ ПРОГНОЗИРОВАНИЯ ВРЕМЕННЫХ РЯДОВ

1.1. Содержательная постановка задачи

Слово прогноз возникло от греческого πρόγνωσις, что означает предвидение, предсказание. Под прогнозированием понимают предсказание будущего с помощью научных методов. Процессом прогнозирования называется специальное научное исследование конкретных перспектив развития какого-либо процесса. Согласно работе [1] процессы, перспективы которых необходимо предсказывать, чаще всего описываются временными рядами, то есть последовательностью значений некоторых величин, полученных в определенные моменты времени. Временной ряд включает в себя два обязательных элемента — отметку времени и значение показателя ряда, полученное тем или иным способом и соответствующее указанной отметке времени. Каждый временной ряд рассматривается как выборочная реализация из бесконечной популяции, генерируемой стохастическим процессом, на который оказывают влияние множество факторов [1]. На рисунке 1.1 представлен пример временного ряда цен на электроэнергию европейской территории РФ.

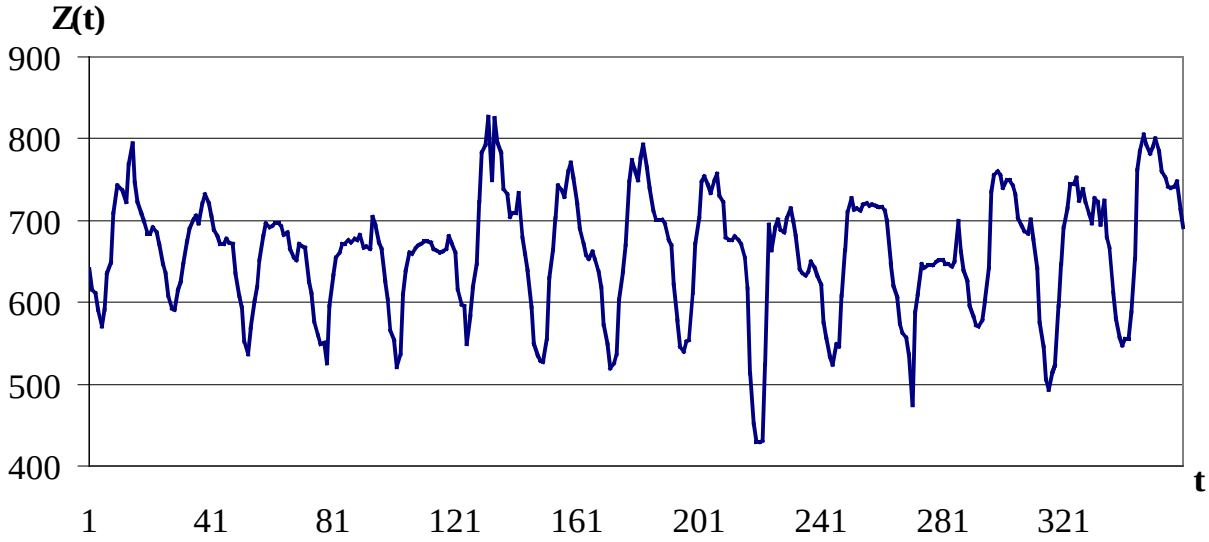


Рис. 1.1 Временной ряд цен на электроэнергию

Одна из классификаций временных рядов приведена в работе [6]. Согласно этой работе, временные ряды различаются способом определения значения, временным шагом, памятью и стационарностью.

В зависимости от способа определения значений временного ряда они делятся на

— интервальные временные ряды,

— моментные временные ряды.

Интервальный временной ряд представляет собой последовательность, в которой уровень явления (значение временного ряда) относят к результату, накопленному или вновь произведенному за определенный интервал времени. Интервальным, например, является временной ряд показателя выпуска продукции предприятием за неделю, месяц или год; объем воды, сброшенной гидроэлектростанцией за час, день, месяц; объем электроэнергии, произведенной за час, день, месяц и другие.

Если же значение временного ряда характеризует изучаемое явление в конкретный момент времени, то совокупность таких значений образует моментный временной ряд. Примерами моментных рядов являются

последовательности финансовых индексов, рыночных цен; физические показатели, такие как температура окружающего воздуха, влажность, давление, измеренные в конкретные моменты времени, и другие.

В зависимости от частоты определения значений временного ряда, они делятся на

— равноотстоящие временные ряды,

— неравноотстоящие временные ряды.

Равноотстоящие временные ряды формируются при исследовании и фиксации значений процесса в следующие друг за другом равные интервалы времени. Большинство физических процессов описываются при помощи равноотстоящих временных рядов. Неравноотстоящими временными рядами называются те ряды, для которых принцип равенства интервалов фиксации значений не выполняется. К таким рядам относятся, например, все биржевые индексы в связи с тем, что их значения определяются лишь в рабочие дни недели.

В зависимости от характера описываемого процесса временные ряды разделяются на

— временные ряды длинной памяти,

— временные ряды короткой памяти.

Задача отнесения временного ряда к рядам с короткой или длинной памятью описана в статье [7]. В целом, говоря о временных рядах с длинной памятью, подразумеваются временные ряды, для которых автокорреляционная функция, введенная в книге [1], убывает медленно. К временным рядам с короткой памятью относят временные ряды, автокорреляционная функция которых убывает быстро. Скорость потока транспорта по дорогам, а также многие физические процессы, такие как потребление электроэнергии, температура воздуха, относятся к временным

рядам с длинной памятью [7]. К временным рядам с короткой памятью относятся, например, временные ряды биржевых индексов.

Дополнительно временные ряды принято разделять на

— стационарные временные ряды,

— нестационарные временные ряды.

Стационарным временным рядом называется такой ряд, который остается в равновесии относительно постоянного среднего уровня. Остальные временные ряды являются нестационарными. В книге [1] указано, что и в промышленности, и в торговле, и в экономике, где прогнозирование имеет важное значение, многие временные ряды являются нестационарными, то есть не имеющими естественного среднего значения. Нестационарные временные ряды для решения задачи прогнозирования часто приводятся к стационарным при помощи разностного оператора [1].

Горизонт времени, на который необходимо определить значения временного ряда, называется временем упреждения [1]. В зависимости от времени упреждения задачи прогнозирования, как правило, делятся на следующие категории срочности:

— долгосрочное прогнозирование;

— среднесрочное прогнозирование;

— краткосрочное прогнозирование.

Важно отметить, что для каждого временного ряда приведенная классификация имеет собственные диапазоны. Например, для временного ряда уровня сахара крови классификация срочности задачи прогнозирования обуславливается типами инсулина [8]:

— ультракраткосрочное прогнозирование: до 3 – 4 часа;

— краткосрочное прогнозирование: до 5 – 8 часов;

— среднесрочное прогнозирование: до 16 – 24 часов.

Для задачи прогнозирования энергопотребления классификация задач предложена в работе [9]:

— ультракраткосрочное прогнозирование: до одного дня;

— краткосрочное прогнозирование: от одного дня до недели;

— среднесрочное прогнозирование: от одной недели до года;

— долгосрочное прогнозирование: более чем на год вперед.

То есть для различных временных рядов, с различным временным разрешением классификация срочности задач прогнозирования индивидуальна.

Говоря о прогнозировании временных рядов, необходимо различить два взаимосвязанных понятия — метод прогнозирования и модель прогнозирования.

Метод прогнозирования представляет собой последовательность действий, которые нужно совершить для получения модели прогнозирования временного ряда.

Модель прогнозирования есть функциональное представление, адекватно описывающее временной ряд и являющееся основой для получения будущих значений процесса. Часто, говоря о моделях прогнозирования, используется термин модель экстраполяции [10].

Метод прогнозирования содержит последовательность действий, в результате выполнения которой определяется модель прогнозирования конкретного временного ряда. Кроме того, метод прогнозирования содержит действия по оценке качества прогнозных значений. Общий итеративный подход к построению модели прогнозирования состоит из следующий шагов [1].

Шаг 1. На первом шаге на основании предыдущего собственного или стороннего опыта выбирается общий класс моделей для прогнозирования

временного ряда на заданный горизонт.

Шаг 2. Определенный общий класс моделей обширен. Для непосредственной подгонки к исходному временному ряду, развиваются грубые методы идентификации подклассов моделей. Такие методы идентификации используют качественные оценки временного ряда.

Шаг 3. После определения подкласса модели, необходимо оценить ее параметры, если модель содержит параметры, или структуру, если модель относится к категории структурных моделей (раздел 1.3.). На данном этапе обычно используется итеративные способы, когда производится оценка участка (или всего) временного ряда при различных значениях изменяемых величин. Как правило, данный шаг является наиболее трудоемким в связи с тем, что часто в расчет принимаются все доступные исторические значения временного ряда.

Шаг 4. Далее производится диагностическая проверка полученной модели прогнозирования. Чаще всего выбирается участок или несколько участков временного ряда, достаточных по длине для проверочного прогнозирования и последующей оценки точности прогноза. Выбранные для диагностики модели прогнозирования участки временного ряда называются контрольными участками (периодами).

Шаг 5. В случае если точность диагностического прогнозирования оказалась приемлемой для задач, в которых используются прогнозные значения, то модель готова к использованию. В случае если точность прогнозирования оказалось недостаточной для последующего использования прогнозных значений, то возможно итеративное повторение всех описанных выше шагов, начиная с первого.

Моделью прогнозирования временного ряда является функциональное представление, адекватно описывающее временной ряд.

При прогнозировании временных рядов возможны два варианта постановки задачи. В первом варианте для получения будущих значений исследуемого временного ряда используются доступные значения только этого ряда. Во втором варианте для получения прогнозных значений возможно использование не только фактических значений искомого ряда, но и значений набора внешних факторов, представленных в виде временных рядов. В общем случае временные ряды внешних факторов могут иметь разрешение по времени отличное от разрешения искомого временного ряда. Например, в работе [9] подробно обсуждаются внешние факторы, оказывающие влияние на временной ряд энергопотребления. К таким внешним факторам относят температуру окружающей среды, влажность воздуха, а также сезонность, т. е. час суток, день недели, месяц года. В общем случае внешние факторы могут быть дискретными, т. е. представленными временными рядами, например, температура воздуха; или категориальными, т. е. состоящими из подмножеств, например, в зависимости от веса тела человека можно отнести к трем категориям: «легкий», «средний», «тяжелый». Лишь некоторые модели прогнозирования позволяют учитывать категориальные внешние факторы, большинство моделей позволяют учитывать только дискретных (раздел 1.3.).

При прогнозировании временного ряда требуется определить функциональную зависимость, адекватно описывающую временной ряд, которая называется модель прогнозирования. Цель создания модели прогнозирования состоит в получении такой модели, для которой среднее абсолютное отклонение истинного значения от прогнозируемого стремится к минимальному для заданного горизонта, который называется временем упреждения. После того, как модель прогнозирования временного ряда определена, требуется вычислить будущие значения временного ряда, а также

их доверительный интервал.

1.2. Формальная постановка задачи

Прогнозирование без учета внешних факторов. Пусть значения временного ряда доступны в дискретные моменты времени $t=1,2,\dots,T$. Обозначим временной ряд $Z(t)=Z(1),Z(2),\dots,Z(T)$. В момент времени T необходимо определить значения процесса $Z(t)$ в моменты времени $T+1,\dots,T+P$. Момент времени T называется моментом прогноза, а величина P — временем упреждения [1].

1) Для вычисления значений временного ряда в будущие моменты времени требуется определить функциональную зависимость, отражающую связь между прошлыми и будущими значениями этого ряда

$$Z(t)=F(Z(t-1),Z(t-2),Z(t-3),\dots)+\varepsilon_t. \qquad (1.1)$$

Зависимость (1.1) называется моделью прогнозирования. Требуется создать такую модель прогнозирования, для которой среднее абсолютное отклонение истинного значения от прогнозируемого стремится к минимальному для заданного P

$$\bar{E}=\frac{1}{P}\sum_{t=T+1}^{T+P}|\varepsilon_t|\to min. \qquad (1.2)$$

Выражение (1.1) можно переписать в виде

$$\hat{Z}(t)=F(Z(t-1),Z(t-2),Z(t-3),\dots), \qquad (1.3)$$

где $\hat{Z}(t)$ прогнозные (расчетные) значения временного ряда $Z(t)$. Здесь и далее будем использовать «крышечку» для обозначения вычисляемых значений временного ряда.

2) Кроме получения будущих значений $\hat{Z}(T+1),\dots,\hat{Z}(T+P)$

требуется определить доверительный интервал возможных отклонений этих значений.

Задача прогнозирования временного ряда проиллюстрирована на рисунке 1.2.

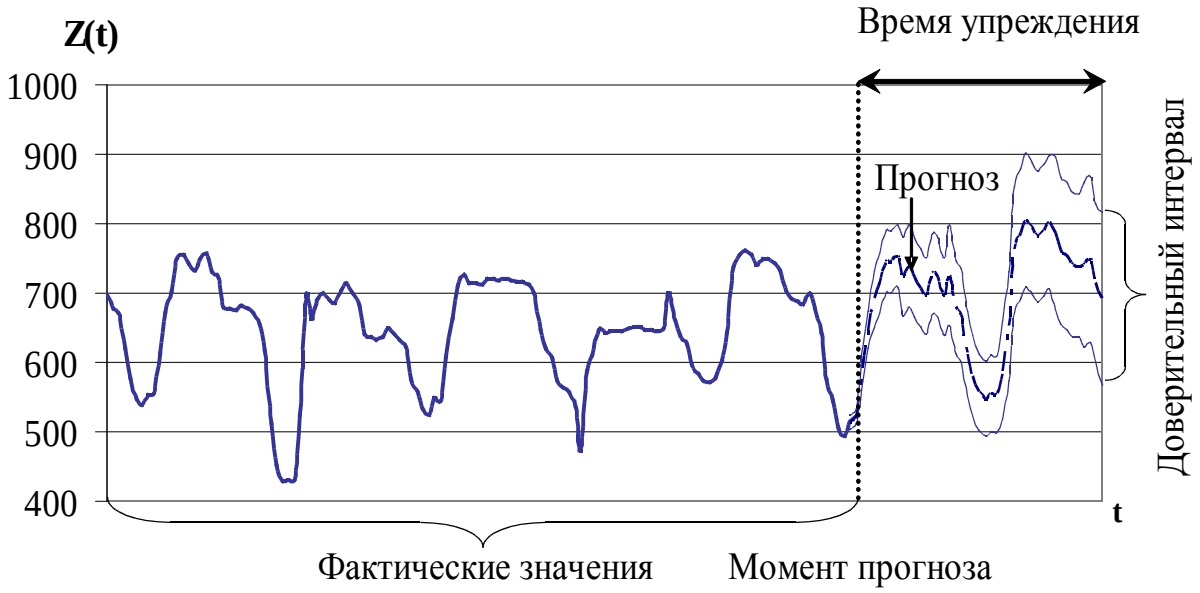


Рис. 1.2. Иллюстрация задачи прогнозирования временного ряда без учета внешних факторов

Прогнозирование с учетом внешних факторов. Пусть значения исходного временного ряда $Z(t)$ доступны в дискретные моменты времени $t=1,2,\ldots,T$. Предполагается, что на значения $Z(t)$ оказывает влияние набор внешних факторов. Пусть первый внешний фактор $X_1(t_1)$ доступен в дискретные моменты времени $t_1=1,2,\ldots,T_1$, второй внешний фактор $X_2(t_2)$ доступен в моменты времени $t_2=1,2,\ldots,T_2$ и т. д. В случае, если дискретность исходного временного ряда и внешних факторов, а также значения $T,T_1,\ldots,T_S$ различны, то временные ряды внешних факторов $X_1(t_1),\ldots, X_S(t_S)$ необходимо привести к единой шкале времени t.

В момент прогноза T необходимо определить будущие значения

исходного процесса $Z(t)$ в моменты времени $T+1,\dots,T+P$, учитывая влияние внешних факторов $X_1(t),\dots, X_S(t)$. При этом считаем, что значения внешних факторов в моменты времени $X_1(T+1),\dots, X_1(T+P),\dots, X_S(T+1)$, $\dots$, $X_S(T+P)$ являются доступными.

1) Для вычисления будущих значений процесса $Z(t)$ в указанные моменты времени требуется определить функциональную зависимость, отражающую связь между прошлыми значениями $Z(t)$ и будущими, а также принимающую во внимание влияние внешних факторов $X_1(t),\dots, X_S(t)$ на исходный временной ряд

$$Z(t)=F(Z(t-1),Z(t-2),\dots,X_1(t),X_1(t-1),\dots,X_S(t),X_S(t-1),\dots)+\varepsilon_t. \qquad (1.4)$$

Зависимость (1.4) называется моделью прогнозирования с учетом внешних факторов $X_1(t),\dots, X_S(t)$. Требуется создать такую модель прогнозирования, для которой среднее абсолютное отклонение истинного значения от прогнозируемого стремится к минимальному для заданного P (1.2).

2) Кроме получения будущих значений $\hat{Z}(T+1),\dots,\hat{Z}(T+P)$ требуется определить доверительный интервал возможных отклонений этих значений.

Задача прогнозирования временного ряда с учетом одного внешнего фактора представлена на рисунке 1.3.

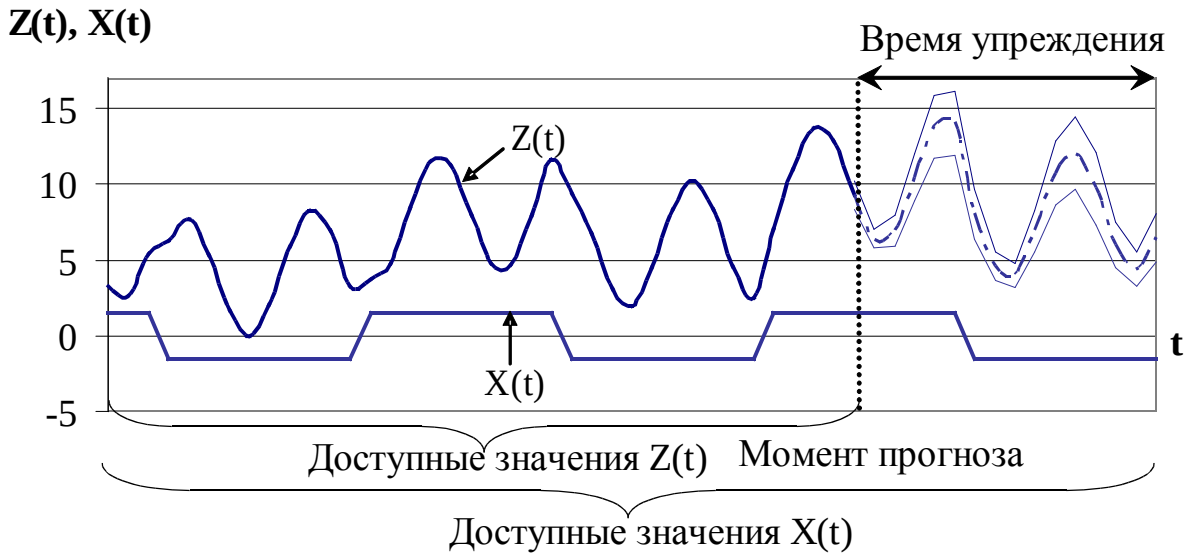


Рис. 1.3. Иллюстрация задачи прогнозирования временного ряда с учетом внешнего фактора

1.3. Обзор моделей прогнозирования

Перед тем как перейти к обзору моделей, необходимо отметить, что названия моделей и соответствующих методов как правило совпадают. Например, работы [1,11-13] посвящены одной из самых распространенных моделей прогнозирования авторегрессия проинтегрированного скользящего среднего с учетом внешнего фактора (auto regression moving average external, ARIMAX). Эту модель и соответствующий ей метод обычно называют ARIMAX. В настоящее время принято использовать английские аббревиатуры названий как моделей, так и методов.

Согласно работе [14], в настоящее время насчитывается свыше 100 классов моделей. Число общих классов моделей, которые в тех или иных вариациях повторяются в других, гораздо меньше. Часть моделей и соответствующих методов относится к отдельным процедурам

прогнозирования. Часть методов представляет набор отдельных приемов, отличающихся от базовых или друг от друга количеством частных приемов и последовательностью их применения.

В аналитическом обзоре [14] все методы прогнозирования делятся на две группы: интуитивные и формализованные.

Интуитивное прогнозирование применяется тогда, когда объект прогнозирования либо слишком прост, либо, напротив, настолько сложен, что аналитически учесть влияние внешних факторов невозможно. Интуитивные методы прогнозирования не предполагают разработку моделей прогнозирования и отражают индивидуальные суждения специалистов (экспертов) относительно перспектив развития процесса. Интуитивные методы основаны на мобилизации профессионального опыта и интуиции. Такие методы используются для анализа процессов, развитие которых либо полностью, либо частично не поддается математической формализации, то есть для которых трудно разработать адекватную модель. В статье [6] указано, что к таким методам относятся методы экспертных оценок, исторических аналогий, предвидения по образцу. Кроме того, в настоящее время широко распространено применение экспертных систем, в том числе с использованием нечеткой логики [15]. В статье [16] подробно описаны интуитивные методы прогнозирования.

Формализованные методы рассматривают модели прогнозирования. В обзоре [9] модели прогнозирования разделяются на статистические модели и структурные модели.

В статистических моделях функциональная зависимость между будущими и фактическими значениями временного ряда, а также внешними факторами задана аналитически. К статистическим моделям относятся следующие группы:

— регрессионные модели;

— авторегрессионные модели;

— модели экспоненциального сглаживания.

В структурных моделях функциональная зависимость между будущими и фактическими значениями временного ряда, а также внешними факторами задана структурно. К структурным моделям относятся следующие группы:

— нейросетевые модели;

— модели на базе цепей Маркова;

— модели на базе классификационно-регрессионных деревьев.

Кроме того, необходимо отметить, что для узкоспециализированных задач иногда применяются особые модели прогнозирования. Так, например, для задачи прогнозирования уровня сахара крови человека применяются модели на основе дифференциальных уравнений [8]. Для задачи прогнозирования транспортного потока, которая в последние несколько лет актуальна для мегаполисов, применяются гидродинамические модели [17]. Для прогнозирования природных явлений, таких как землетрясения, применяется, например, модель, в основу которой положены нелинейные клетки (или соты), находящиеся под воздействием внешнего поля, и у которых есть внутреннее состояние, изменяющееся во времени под воздействием этого поля [18]. Аналогичные модели разрабатываются и применяются для специальных процессов и систем. В рамках настоящей работы данный класс формализованных моделей не рассматривается.

1.3.1. Регрессионные модели

Существует много задач, требующих изучения отношения между двумя и более переменными. Для решения таких задач используется регрессионный анализ [19]. В настоящее время регрессия получила широкое применение,

включая задачи прогнозирования и управления. Целью регрессионного анализа является определение зависимости между исходной переменной и множеством внешних факторов (регрессоров). При этом коэффициенты регрессии могут определяться по методу наименьших квадратов [19] или методу максимального правдоподобия [20].

Линейная регрессионная модель. Самым простым вариантом регрессионной модели является линейная регрессия. В основу модели положено предположение, что существует дискретный внешний фактор $X(t)$, оказывающий влияние на исследуемый процесс $Z(t)$, при этом связь между процессом и внешним фактором линейна. Модель прогнозирования на основании линейной регрессии описывается уравнением

$$Z(t)=\alpha_0+\alpha_1 X(t)+\varepsilon_t, \qquad (1.5)$$

где α_0 и α_1 — коэффициенты регрессии; ε_t — ошибка модели. Для получения прогнозных значений $Z(t)$ в момент времени t необходимо иметь значение $X(t)$ в тот же момент времени t, что редко выполнимо на практике.

Множественная регрессионная модель. На практике на процесс $Z(t)$ оказывают влияние целый ряд дискретных внешних факторов $X_1(t)$,…, $X_S(t)$. Тогда модель прогнозирования имеет вид

$$Z(t)=\alpha_0+\alpha_1 X_1(t)+\alpha_2 X_2(t)+...+\alpha_s X_S(t)+\varepsilon_t. \qquad (1.6)$$

Недостатком данной модели является то, что для вычисления будущего значения процесса $Z(t)$ необходимо знать будущие значения всех факторов $X_1(t)$, …, $X_S(t)$, что почти невыполнимо на практике.

В основу **нелинейной регрессионной модели** положено предположение о том, что существует известная функция, описывающая зависимость между исходным процессом $Z(t)$ и внешним фактором $X(t)$

$$Z(t)=F(X(t),A). \tag{1.7}$$

В рамках построения модели прогнозирования необходимо определить параметры функции A. Например, можно предположить, что

$$Z(t)=\alpha_1\cos(X(t))+\alpha_0. \tag{1.8}$$

Для построения модели достаточно определить параметры $A=[\alpha_1,\alpha_0]$. Однако на практике редко встречаются процессы, для которых вид функциональной зависимости между процессом $Z(t)$ и внешним фактором $X(t)$ заранее известен. В связи с этим нелинейные регрессионные модели применяются редко.

Модель группового учета аргументов (МГУА) была разработана Ивахтенко А.Г. [21]. Модель имеет вид

$$\begin{aligned} Z(t)=\alpha_0+\sum_{i=1}^{s}\alpha_i X_i(t)+\sum_{i=1}^{s}\sum_{j=1}^{s}\alpha_{i,j}X_i(t)X_j(t) \\ +\sum_{i=1}^{s}\sum_{j=1}^{s}\sum_{k=1}^{s}\alpha_{i,j,k}X_i(t)X_j(t)X_k(t)+... \end{aligned} \tag{1.9}$$

Уравнение (1.9) называется опорной функцией. Используя опорную функцию, строят различные варианты моделей для некоторых или всех аргументов. Например, строятся полиномы с одной переменной, полиномы со всевозможными парами переменных, полиномы со всевозможными тройками переменных и т. д. Для каждой модели определяются её линейные коэффициенты $\alpha_{i,j,k,...}$ методом регрессионного анализа. Среди всех моделей выбираются несколько (от 2 до 10) наилучших. При этом качество моделей определяется, например, среднеквадратичным отклонением или иным критерием. Если среди выбранных имеется модель, качество которой достаточно для использования полученных прогнозных значений, то процесс перебора моделей прекращается. Иначе отобранные модели используются в качестве аргументов $X_1(t)$,..., $X_s(t)$ для опорных функций следующего

этапа итерации. То есть уже найденные модели участвуют в формировании более сложных.

1.3.2. Авторегрессионные модели

В основу авторегрессионных моделей заложено предположение о том, что значение процесса $Z(t)$ линейно зависит от некоторого количества предыдущих значений того же процесса $Z(t-1)$, …, $Z(t-p)$.

Авторегрессионная модель скользящего среднего. В области анализа временных рядов модель авторегрессии (autoregressive, AR) и модель скользящего среднего (moving average, MA) является одной из наиболее используемых [1,5].

Согласно работе [1], модель авторегрессии является исключительно полезной для описания некоторых встречающихся на практике временных рядов. В этой модели текущее значение процесса выражается как конечная линейная совокупность предыдущих значений процесса и импульса, который называется «белым шумом»,

$$Z(t)=C+\phi_1 Z(t-1)+\phi_2 Z(t-2)+...\phi_p Z(t-p)+\varepsilon_t. \qquad (1.10)$$

Формула (1.10) описывает процесс авторегрессии порядка p, который в литературе часто обозначается AR(p), здесь C — вещественная константа, $\phi_1,..,\phi_p$ — коэффициенты, ε_t — ошибка модели. Для определения ϕ_i и C используют метод наименьших квадратов [19] или метод максимального правдоподобия [20].

Другой тип модели имеет большое значение в описании временных рядов и часто используется совместно с авторегрессией называется моделью скользящего среднего порядка q и описывается уравнением

$$Z(t)=\frac{1}{q}(Z(t-1)+Z(t-2)+...+Z(t-q))+\varepsilon_t. \qquad (1.11)$$

В литературе процесс (1.11) часто обозначается MA(q); здесь q — порядок

скользящего среднего, ε_t — ошибка прогнозирования. Модель скользящего среднего является по сути дела фильтром низких частот. Нужно отметить, что существуют простые, взвешенные, кумулятивные, экспоненциальные модели скользящего среднего.

Согласно работе [1], для достижения большей гибкости в подгонке модели часто целесообразно объединить в одной модели авторегрессию и скользящее среднее. Общая модель обозначается ARMA(p,q) соединяет в себе фильтр в виде скользящего среднего порядка q и авторегрессию фильтрованных значений процесса порядка p.

Если в качестве входных данных используются не сами значения временного ряда, а их разность d-того порядка (на практике d необходимо определять, однако в большинстве случаев $d \leqslant 2$), то модель носит название авторгерессии проинтегрированного скользящего среднего. В литературе данную модель называют ARIMA(p,d,q) (autoregression integrated moving average).

Развитием модели ARIMA(p,d,q) является модель ARIMAX(p,d,q), которая описывается уравнением [1]

$$Z(t) = AR(p) + \alpha_1 X_1(t) + ... + \alpha_s X_S(t) \quad (1.12)$$

Здесь α_1,..., α_s — коэффициенты внешних факторов $X_1(t)$,..., $X_S(t)$. В данной модели чаще всего процесс $Z(t)$ является результатом модели MA(q), то есть отфильтрованными значениями исходного процесса. Далее для прогнозирования $Z(t)$ используется модель авторегрессии, в которой введены дополнительные регрессоры внешних факторов $X_1(t)$,..., $X_S(t)$.

Авторегрессионная модель с условной гетероскедастичностью (autoregressive conditional heteroskedasticity, GARCH) была разработана в 1986 году Тимом Петером Борреслевом и является моделью остатков для модели AR(p) [22]. На первом этапе для исходного временного ряда определяется

модель AR(p) (1.10). Далее предполагается, что ошибка модели (1.10) ε_t имеет две составляющие

$$\varepsilon_t = \sigma_t \cdot \varsigma_t , \tag{1.13}$$

где σ_t — зависимое от времени стандартное отклонение; ς_t — случайная величина, имеющая нормальное распределение, среднее значение, равное 0, и стандартное отклонение, равное 1. При этом зависимое от времени стандартное отклонение описывается уравнением

$$\sigma_t^2 = \beta_0 + \beta_1 \varepsilon_{t-1}^2 + ... + \beta_q \varepsilon_{t-q}^2 + \gamma_1 \sigma_{t-1} + ... + \gamma_p \sigma_{t-p} . \tag{1.14}$$

Здесь β_0,..., β_q и γ_0,..., γ_p — коэффициенты. Уравнение (1.14) называется моделью GARCH(p,q) и имеет два параметра: p характеризует порядок авторегрессии квадратов остатков; q — количество предшествующих оценок остатков.

Наиболее частое применение данная модель получила в финансовом секторе, где с помощью нее моделируется волатильность. На сегодняшний день существует ряд модификаций модели под названиями NGARCH, IGARCH, EGARCH, GARCH-M и другие [22].

Авторегрессионнная модель с распределенным лагом (autoregressive distributed lag models, ARDLM) недостаточно подробно описана в литературе. Основное внимание данной модели уделяется в книгах по эконометрике [23].

Часто при моделировании процессов на изучаемую переменную влияют не только текущие значения процесса, но и его лаги, то есть значения временного ряда, предшествующие изучаемому моменту времени. Модель авторегрессии распределенного лага описывается уравнением

$$Z(t) = \phi_0 + \phi_1 Z(t-l-1) + ... + \phi_p Z(t-l-p) + \varepsilon_t . \tag{1.15}$$

Здесь ϕ_0,..., ϕ_p — коэффициенты, l — величина лага. Модель (1.15) называется ARDLM(p,l) и чаще всего применяется для моделирования экономических процессов [23].

1.3.3. Модели экспоненциального сглаживания

Модели экспоненциального сглаживания разработаны в середине XX века и до сегодняшнего дня являются широко распространенными в силу их простоты и наглядности.

Модель экспоненциального сглаживания (exponential smoothing, ES) применяется для моделирования финансовых и экономических процессов [24]. В основу экспоненциального сглаживания заложена идея постоянного пересмотра прогнозных значений по мере поступления фактических. Модель ES присваивает экспоненциально убывающие веса наблюдениям по мере их старения. Таким образом, последние доступные наблюдения имеют б*о*льшее влияние на прогнозное значение, чем старшие наблюдения.

Функция модели ES имеет вид

$$\begin{gathered} Z(t)=S(t)+\varepsilon_t, \\ S(t)=\alpha\cdot Z(t-1)+(1-\alpha)\cdot S(t-1), \end{gathered} \tag{1.16}$$

где α — коэффициент сглаживания, $0<\alpha<1$; начальные условия определяются как $S(1)=Z(0)$. В данной модели каждое последующее сглаженное значение $S(t)$ является взвешенным средним между предыдущим значением временного ряда $Z(t)$ и предыдущего сглаженного значения $S(t-1)$.

Модель Хольта или двойное экспоненциальное сглаживание применяется для моделирования процессов, имеющих тренд. В этом случае в модели необходимо рассматривать две составляющие: уровень и тренд [24]. Уровень и тренд сглаживаются отдельно

$$\begin{gathered} Z(t)=S(t)+\varepsilon_t; \\ S(t)=\alpha\cdot Z(t-1)+(1-\alpha)\cdot(S(t-1)-B(t-1)); \\ B(t)=\gamma\cdot(S(t-1)-S(t-2))+(1+\gamma)\cdot B(t-1). \end{gathered} \tag{1.17}$$

Здесь α — коэффициент сглаживания уровня, как и в модели (1.16), γ —

коэффициент сглаживания тренда.

Модель Хольта-Винтерса или тройное экспоненциальное сглаживание применяется для процессов, которые имеют тренд и сезонную составляющую

$$Z(t)=(R(t)+G(t))\cdot S(t)\,. \tag{1.18}$$

Здесь $R(t)$ — сглаженный уровень без учета сезонной составляющей

$$R(t)=\frac{\alpha\cdot Z(t-1)}{S(t-L)}+(1+\alpha)\cdot(R(t-1)+G(t-1))\,, \tag{1.19}$$

$G(t)$ — сглаженный тренд

$$G(t)=\beta\cdot(S(t-1)-S(t-2))+(1-\beta)\cdot G(t-1)\,, \tag{1.20}$$

а $S(t)$ — сезонная составляющая

$$S(t)=\frac{\gamma\cdot Z(t-1)}{S(t-L)}+(1-\gamma)\cdot S(t-L)\,. \tag{1.21}$$

Величина L определяется длиной сезона исследуемого процесса. Модели экспоненциального сглаживания наиболее популярны для долгосрочного прогнозирования.

1.3.4. Нейросетевые модели

В настоящее время самой популярной среди структурных моделей является модель на основе искусственных нейронных сетей (artificial neural network, ANN) [5]. Нейронные сети состоят из нейронов (рис. 1.4).

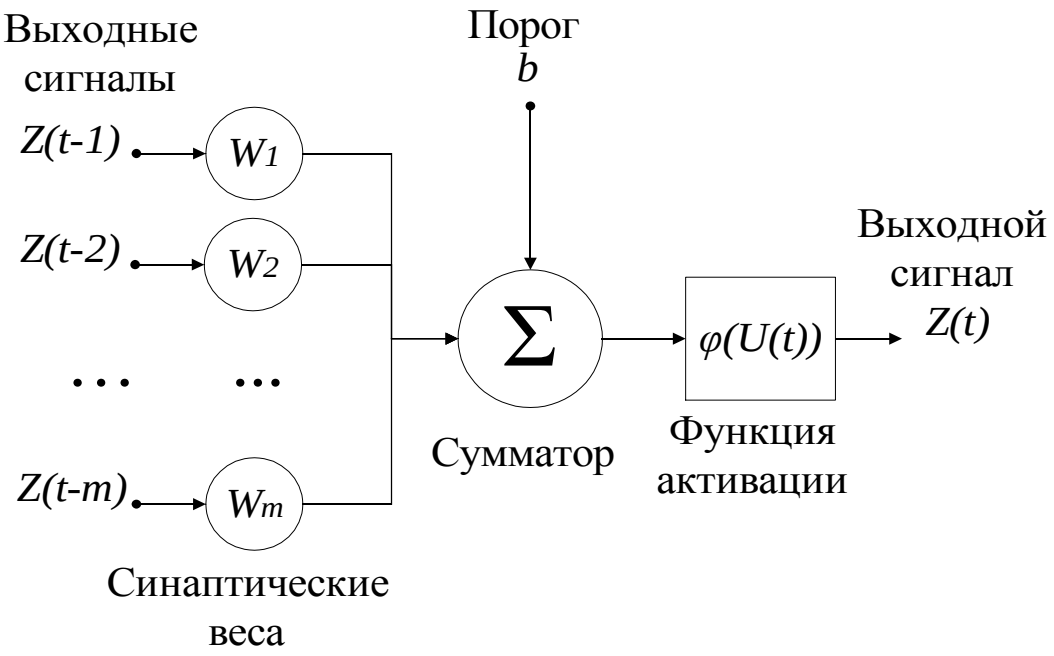


Рис. 1.4. Нелинейная модель нейрона

Модель нейрона можно описать парой уравнений

$$U(t)=\sum_{i=1}^{m}\omega_i\cdot Z(t-i)+b,$$
$$Z(t)=\varphi(U(t)), \qquad (1.22)$$

где $Z(t-1),...,Z(t-m)$ — входные сигналы; $\omega_1,...,\omega_m$ — синаптические веса нейрона; b — порог; $\varphi(U(t))$ — функция активации.

Функция активации бывают трех основных типов [25]:

— функция единичного скачка;

— кусочно-линейная функция;

— сигмоидальная функция.

Способ связи нейронов определяет архитектуру нейронной сети. Согласно работе [25], в зависимости от способа связи нейронов сети делятся на

— однослойные сети прямого распространения,

— многослойные сети прямого распространения,

— рекуррентные сети.

На рисунке 1.5 представлена структура трехслойной нейронной сети прямого распространения, применяемая для прогнозирования в работах [26-29].

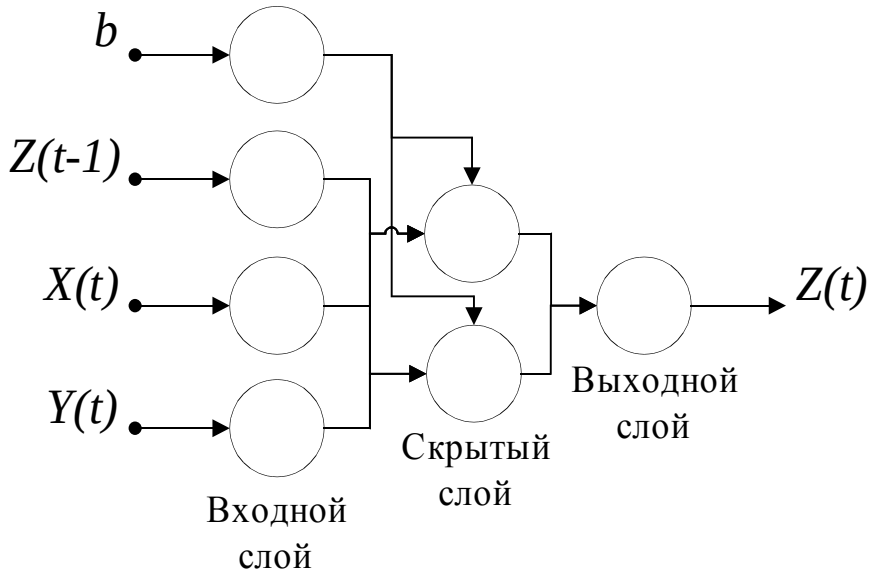


Рис. 1.5. Трехслойная нейронная сеть прямого распространения

Таким образом, при помощи нейронных сетей возможно моделирование нелинейной зависимости будущего значения временного ряда от его фактических значений и от значений внешних факторов. Нелинейная зависимость определяется структурой сети и функцией активации.

1.3.5. Модели на базе цепей Маркова

Модели прогнозирования на основе цепей Маркова (Markov chain model) предполагают, что будущее состояние процесса зависит только от его текущего состояния и не зависит от предыдущих [30]. В связи с этим процессы, моделируемые цепями Маркова, должны относиться к процессами с короткой памятью.

Пример цепи Маркова для процесса, имеющего три состояния, представлен на рис. 1.6.

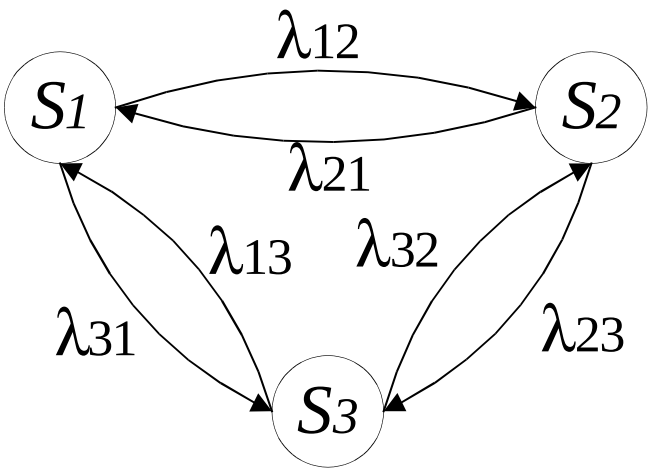


Рис. 1.6. Цепь Маркова с тремя состояниями

Здесь $S_1,...,S_3$ — состояния процесса $Z(t)$; λ_{12} — вероятность перехода из состояния S_1 в состояние S_2, λ_{23} — вероятность перехода из состояния S_2 в состояние S_3 и т. д. При построении цепи Маркова определяется множество состояний и вероятности переходов. Есть текущее состояние процесса S_i, то качестве будущего состояния процесса выбирается такое состояние S_j, вероятность перехода в которое (значение λ_{ij}) максимальна.

Таким образом, структура цепи Маркова и вероятности перехода состояний определяют зависимость между будущим значением процесса и его текущим значением.

1.3.6. Модели на базе классификационно-регрессионных деревьев

Классификационно-регрессионные деревья (classification and regression trees, CART) являются еще одной популярной структурной моделью прогнозирования временных рядов [31]. Структурные модели CART разработаны для моделирования процессов, на которые оказывают влияние как непрерывные внешние факторы, так и категориальные. Если внешние факторы, влияющие на процесс $Z(t)$, непрерывны, то используются

регрессионные деревья; если факторы категориальные, то — классификационные деревья. В случае, если необходимо учитывать факторы обоих типов, то используются смешанные классификационно-регрессионные деревья.

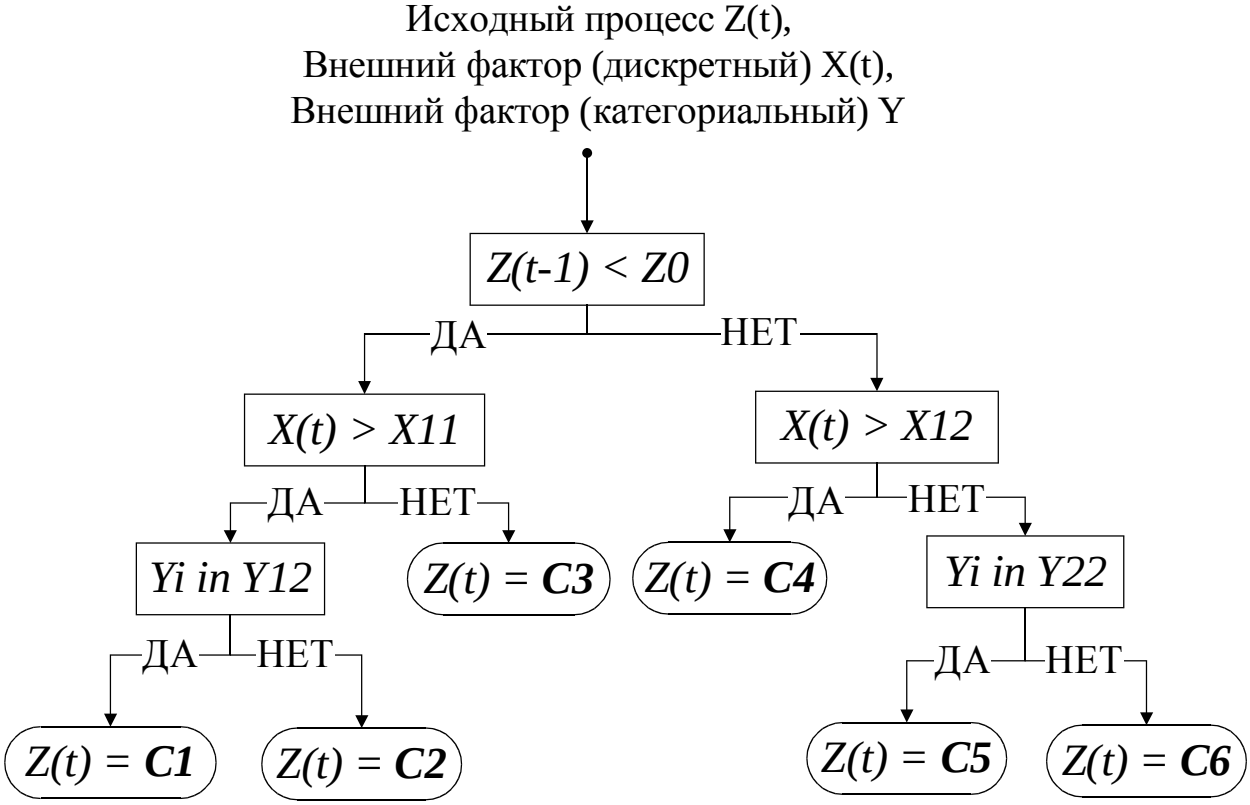


Рис. 1.7. Бинарное классификационно-регрессионное дерево

Согласно модели CART, прогнозное значение временного ряда зависит от предыдущих значений, а также некоторых независимых переменных. На приведенном на рисунке 1.7 примере сначала предыдущее значение процесса сравнивается с константой Z_0. Если значение $Z(t-1)$ меньше Z_0, то выполняется следующая проверка: $X(t) > X_{11}$. Если неравенство не выполняется, то $Z(t) = C_3$, иначе проверки продолжаются до того момента, пока не будет найден лист дерева, в котором происходит определение будущего значения процесса $Z(t)$. Важно, что при определении значения в расчет принимаются как непрерывные переменные, например, $X(t)$, так

и категориальные Y, для которых выполняется проверка присутствия значения в одном из заранее определенных подмножеств. Значения пороговых констант, например, Z_0, X_{11}, а также подмножеств Y_{12}, Y_{22} выполняется на этапе обучения дерева [31].

Таким образом, CART моделирует зависимость будущей величины процесса $Z(t)$ при помощи структуры дерева, а также пороговых констант и подмножеств.

1.3.7. Другие модели и методы прогнозирования

Кроме классов моделей прогнозирования, рассмотренных выше, существуют менее распространенные модели и методы прогнозирования. Главным недостатком моделей и методов, упомянутых в настоящем разделе, является недостаточная методологическая база, т. е. недостаточно подробное описание возможностей как моделей, так и путей определения их параметров. Кроме того, в открытом доступе можно найти лишь небольшое количество статей, посвященных применению данных методов.

Метод опорных векторов (support vector machine, SVM) применяется, например, для прогнозирования движения рынков [32] и цен на электроэнергию [5]. В основу метода положена классификация, производимая за счет перевода исходных временных рядов, представленных в виде векторов, в пространство более высокой размерности и поиска разделяющей гиперплоскости с максимальным зазором в этом пространстве. Алгоритм SVM работает в предположении, что чем больше разница или расстояние между этими параллельными гиперплоскостями, тем меньше будет средняя ошибка классификатора [33]. При этом задача прогнозирования решается таким образом, что на этапе обучения классификатора выявляются независимые переменные (внешние факторы), будущие значения которых определяют в какой из определенных ранее подклассов попадет прогноз $Z(t)$

.

Генетический алгоритм (genetic algorithm, GA) был разработан и часто применяется для решения задач оптимизации, а также поисковых задач. Однако некоторые модификации GA позволяют решать задачи прогнозирования.

В статье [34] указано, что алгоритм прогнозирования на основе GA позволяет принимать в расчет более 15 внешних факторов, используя базовый GA. Принцип работы основан на том, что исходные значения процесса $Z(t)$ и внешних факторов $X_1(t),\ldots, X_s(t)$ раскладывают в наборы, состоящие из 0 и 1, которые называют генотипами. Далее применяют ряд преобразований: скрещивание и мутирование для формирования преобразованных наборов, которые называются фенотипами. Исходные и полученные наборы исследуются с использованием функции приспособленности. Если решение получилось неудовлетворительным, то снова производится скрещивание и мутирование, в результате чего получается еще более новые наборы (новое поколение), которые снова оцениваются. Итеративный процесс продолжается до тех пор, пока решение не будет удовлетворительным.

Модель на основе передаточных функций (transfer function, TF) применяется для прогнозирования процесса $Z(t)$ с учетом внешнего фактора $X(t)$ [35]. Уравнение, отражающее зависимость будущего значения имеет вид

$$Z(t)=\nu(B)X(t)+\eta(t), \qquad (1.23)$$

где B — оператор сдвига $BZ(t)=Z(t-1),\ldots,\ B^kZ(t)=Z(t-k)$. Временной ряд $\eta(t)$ характеризует внешнее возмущение. При этом функция $\nu(B)$ имеет вид

$$\nu(B)=\nu_0+\nu_1 B+\nu_2 B^k+... \quad (1.24)$$

Коэффициенты функции (1.24) ν_i описывают динамические отношения между процессами $Z(t)$ и $X(t)$.

1.4. Сравнение моделей прогнозирования

В предыдущем разделе представлен обзор существующих моделей прогнозирования. В настоящем разделе рассмотрены преимущества и недостатки не только описанных выше моделей, но и методов. Говоря о достоинствах и недостатках моделей прогнозирования, необходимо принимать во внимание и соответствующие им методы.

1.4.1. Достоинства и недостатки моделей

Регрессионные модели и методы. К достоинствам данных моделей относят простоту, гибкость, а также единообразие их анализа и проектирования [19]. При использовании линейных регрессионных моделей результат прогнозирования может быть получен быстрее, чем при использовании остальных моделей. Кроме того, достоинством является прозрачность моделирования [5], т. е. доступность для анализа всех промежуточных вычислений.

Основным недостатком нелинейных регрессионных моделей является сложность определения вида функциональной зависимости [14], а также трудоемкость определение параметров модели. Недостатками линейных регрессионных моделей являются низкая адаптивность и отсутствие способности моделирования нелинейных процессов [28].

Авторегрессионные модели и методы. Важными достоинствами данного класса моделей являются их простота и прозрачность моделирования. Еще одним достоинством является единообразие анализа

и проектирования, заложенное в работе [1]. На сегодняшний день данный класс моделей является одним из наиболее популярных [3], а потому в открытом доступе легко найти примеры применения авторегрессионных моделей для решения задач прогнозирования временных рядов различных предметных областей.

Недостатками данного класса моделей являются: большое число параметров модели, идентификация которых неоднозначна и ресурсоемка [4]; низкая адаптивность моделей, а также линейность и, как следствие, отсутствие способности моделирования нелинейных процессов, часто встречающихся на практике [26].

Модели и методы экспоненциального сглаживания. Достоинствами данного класса моделей являются простота и единообразие их анализа и проектирования. Данный класс моделей чаще других используется для долгосрочного прогнозирования [24].

Недостатком данного класса моделей является отсутствие гибкости [36].

Нейросетевые модели и методы. Основным достоинством нейросетевых моделей является нелинейность, т. е. способность устанавливать нелинейные зависимости между будущими и фактическими значениями процессов. Другими важными достоинствами являются: адаптивность, масштабируемость (параллельная структура ANN ускоряет вычисления) и единообразие их анализа и проектирования [25].

При этом недостатками ANN являются отсутствие прозрачности моделирования; сложность выбора архитектуры, высокие требования к непротиворечивости обучающей выборки; сложность выбора алгоритма обучения и ресурсоемкость процесса их обучения [5].

Модели и методы на базе цепей Маркова. Простота и единообразие

анализа и проектирования являются достоинствами моделей на базе цепей Маркова.

Недостатком данных моделей является отсутствие возможности моделирования процессов с длинной памятью [30].

Модели на базе классификационно-регрессионных деревьев. Достоинствами данного класса моделей являются: масштабируемость, за счет которой возможна быстрая обработка сверхбольших объемов данных; быстрота и однозначность процесса обучения дерева (в отличие от ANN) [9], а также возможность использовать категориальные внешние факторы.

Недостатками данных моделей являются неоднозначность алгоритма построения структуры дерева; сложность вопроса останова т. е. вопроса о том, когда стоит прекратить дальнейшие ветвления; отсутствие единообразия их анализа и проектирования [31].

Достоинства и недостатки моделей и методов систематизированы в таблице 1.

Таблица 1.

Сравнение моделей и методов прогнозирования

Модель и метод	**Достоинства**	**Недостатки**
Регрессионные модели и методы	простота, гибкость, прозрачность моделирования; единообразие анализа и проектирования	сложность определения функциональной зависимости; трудоемкость нахождения коэффициентов зависимости; отсутствие возможности моделирования нелинейных процессов (для нелинейной регрессии)

Таблица 1. – окончание

Модель и метод	**Достоинства**	**Недостатки**
Авторегрессионные модели и методы	простота, прозрачность моделирования; единообразие анализа и проектирования; множество примеров применения	трудоемкость и ресурсоемкость идентификации моделей; невозможность моделирования нелинейностей; низкая адаптивность
Модели и методы экспоненциального сглаживания	простота моделирования; единообразие анализа и проектирования	недостаточная гибкость; узкая применимость моделей
Нейросетевые модели и методы	нелинейность моделей; масштабируемость, высокая адаптивность; единообразие анализа и проектирования; множество примеров применения	отсутствие прозрачности; сложность выбора архитектуры; жесткие требования к обучающей выборке; сложность выбора алгоритма обучения; ресурсоемкость процесса обучения
Модели и методы на базе цепей Маркова	простота моделирования; единообразие анализа и проектирования	невозможность моделирования процессов с длинной памятью; узкая применимость моделей
Модели и методы на базе классификационно-регрессионных деревьев	масштабируемость; быстрота и простота процесса обучения; возможность учитывать категориальные переменные	неоднозначность алгоритма построения дерева; сложность вопроса останова

Нужно дополнительно отметить, что ни для одной из рассмотренных групп моделей (и методов) в достоинствах не указана точность прогнозирования. Это сделано в связи с тем, что точность прогнозирования того или иного процесса зависит не только от модели, но и от опыта исследователя, от доступности данных, от располагаемой аппаратной мощности и многих других факторов. Точность прогнозирования будет оцениваться для конкретных задач, решаемых в рамках данной работы.

В ряде работ [12,36,37] указано, что на сегодняшний день наиболее распространенными моделями прогнозирования являются авторегрессионные модели (ARIMAX), а также нейросетевые модели (ANN). В статье [3], в частности, утверждается: «Without a doubt ARIMA(X) and GRACH modeling methodologies are the most popular methodologies for forecasting time series. Neural networks are now the biggest challengers to conventional time series forecasting methods». (Без сомнений модели ARIMA(X) и GARCH являются самыми популярными для прогнозирования временных рядов. В настоящее время главную конкуренцию данным моделям составляют модели на основе ANN.)

1.4.2. Комбинированные модели

Одной из популярных современных тенденций в области создания моделей прогнозирования является создание комбинированных моделей и методов. Подобный подход дает возможность компенсировать недостатки одних моделей при помощи других и направлен на повышение точности прогнозирования, как одного из главных критериев эффективности модели.

Одной из первых работ в этой области является статья [38]. В ней предлагается подход, в котором прогнозирование временного ряда осуществляется в два этапа. На первом этапе на основании моделей распознавания образов (pattern recognition) выделяются гомогенные группы

(patterns) временного ряда. На следующем этапе для каждой группы строится отдельная модель прогнозирования. В статье указывается, что при комбинированном подходе удается повысить точность прогнозирования временных рядов.

В работе [13] предлагается модель для прогнозирования цен на электроэнергию Испании. При помощи вейвлет преобразования (wavelet transform) доступные значения временного ряда разделяются на несколько последовательностей, для каждой из которых строится отдельная модель ARIMA.

В обзоре моделей прогнозирования энергопотребления [36] рассматривается следующие типы комбинаций:

— ANN + нечеткая логиа;
— ANN + ARIMA;
— ANN + регрессия;
— ANN + GA + нечеткая логика;
— регрессия + нечеткая логика.

В большинстве комбинаций модели на основе ANN применяются для решения задачи кластеризации, а далее для каждого кластера строиться отдельная модель прогнозирования на основе ARIMA, GA, нечеткой логики и др. В работе утверждается, что применение комбинированных моделей, выполняющих предварительную кластеризации и последующее прогнозирование внутри определенного кластера, является наиболее перспективным направлением развития моделей прогнозирования.

Работа [39] посвящена вопросам кластеризации временных рядов для того, чтобы на основании полученных кластеров выполнять прогнозирование. Для кластеризации предлагается два метода: метод K- средних (K-mean) и метод нечетких C-средних (fuzzy C-mean). Целью

обоих алгоритмов кластеризации является извлечение полезной информации из временного ряда для последующего прогнозирования. Авторы утверждают, что применение кластеризации дает возможность повысить точность прогнозирования.

Применение комбинированных моделей является направлением, которое при корректном подходе позволяет повысить точность прогнозирования. Главным недостатком комбинированных моделей является сложность и ресурсоемкость их разработки: нужно разработать модели таким образом, чтобы компенсировать недостатки каждой из них, не потеряв достоинств.

Ряд исследователей пошли по альтернативному пути и разработали авторегрессионные модели, в основе которых лежит предположение о том, что временной ряд есть последовательность повторяющихся кластеров (patterns). Однако при этом разработчики не создавали комбинированных моделей, а определяли кластеры и выполняли прогноз на основании одной модели. Рассмотрим эти модели подробнее.

В работе [40] предложена модель прогнозирования направления движения индексов рынка (index movement), учитывающая кластеры временного ряда. Пусть временной ряд $Z(t)$ содержит три значения -1, 0 и 1, которые характеризуют спад, стабильное состояние и подъем рынка соответственно. Кластером (pattern) называется последовательность $Z_i^M = Z(i), Z(i+1), \ldots, Z(i+M)$ для $i \in \{1,2,\ldots,N-M\}$, где N — число доступных отчетов временного ряда $Z(t)$. Для определения прогнозного значения рассмотрена последняя доступная информация, а именно последовательность $Z(N,M) = Z(N-M+1), Z(N-M+2), \ldots, Z(N)$ для которой определена ближайшая похожая (closet match) $Z(Q,M) = Z(Q+1), Z(Q+2), \ldots, Z(Q+M)$. При этом функция,

определяющая близость, имеет вид

$$F(N-M,Q)=\sum_{j=1}^{M}|Z(N-M+1)-Z(Q+1)|, \quad (1.25)$$

т. е. близость кластеров определяется простым сравнением. Далее вычисляется прогнозное значение

$$Z(N+1)=Z(Q+M+1). \quad (1.26)$$

Таким образом, в данной модели предполагается, что если в некоторый момент времени в прошлом рынок вел себя определенным образом, то в будущем его поведение повторится в связи с тем, что временной ряд является последовательностью кластеров.

Еще в двух работах [41,42] предложена модель прогнозирования, основанная на модели авторегрессии, но принимающая во внимание кусочки временного ряда. Здесь прогнозное значение временного ряда определено выражением

$$Z(t)=\alpha_0+\alpha_1\cdot Z(t-1)+\alpha_2\cdot Z(t-2)+...+\alpha_M\cdot Z(t-M), \quad (1.27)$$

которое является линейной авторегрессией порядка M. При этом коэффициенты авторегрессии $\alpha_0, \alpha_1, \ldots, \alpha_M$ определяются следующим образом. Предполагается, что существует K кусочков (векторов) длины M временного ряда, для которых выполняется выражение

$$\begin{bmatrix} Z(i_1) \\ Z(i_2) \\ ... \\ Z(i_K) \end{bmatrix} = \alpha_0 \cdot \begin{bmatrix} 1 \\ 1 \\ ... \\ 1 \end{bmatrix} + \alpha_1 \cdot \begin{bmatrix} Z(i_1-1) \\ Z(i_2-1) \\ ... \\ Z(i_K-1) \end{bmatrix} + ... + \alpha_M \cdot \begin{bmatrix} Z(i_1-M) \\ Z(i_2-M) \\ ... \\ Z(i_K-M) \end{bmatrix}. \quad (1.28)$$

При определении ближайших векторов (closest vectors) $Z(i_1-1), Z(i_1-2), ..., Z(i_1-M), \ldots, Z(i_K-1), Z(i_K-2), ..., Z(i_K-M)$ в статье [41] использовано значение линейной корреляции Пирсона между всеми возможными векторами и новейшим вектором (last available vector)

$Z(t-1), Z(t-2), \ldots, Z(t-M)$; а в статье [42] вместо линейной корреляции рассчитывается евклидово расстояние между векторами.

Отметим, что существует путаница в терминологии: в статье [41] использован термин pieces (кусочки), в статьях [4,42] — термин vector, set (вектор, выборка); в работе [40] для аналогичного понятия использован термин pattern (выборка, кластер). В настоящей работе используем термин выборка (pattern) [40]. Англицизм паттерн в русском языке чаще применяется для описания задач классификации, например, в работе [43], а также кластеризации и распознавания образов (pattern recognition) [44].

Разработчики рассмотренных выше моделей утверждают, что предложенные модели просты, прозрачны и эффективны для исследованных временных рядов. При этом очевидно, что главными недостатками данных моделей являются:

— невозможность учитывать внешние факторы;

— неоднозначность критерия определения похожей выборки;

— сложность определения эффективной комбинации двух параметров M (длина векторов) и K (число векторов, принимаемых в расчет) в работах [41,42].

В рамках диссертации установлено, что подход, предложенный авторами работ [40-42], является перспективным в области создания моделей прогнозирования временных рядов. Предложенная в диссертации модель прогнозирования развивает модели [40-42] и устраняет все перечисленные выше недостатки: модель позволяет учитывать влияния внешних факторов; формулируется критерий определения похожей выборки для двух видов постановок задачи прогнозирования (1.2.); количество параметром модели сокращается до одного, что существенно упрощает идентификацию модели.

1.5. Выводы

1) Задача прогнозирования временных рядов имеет высокую актуальность для многих предметных областей и является неотъемлемой частью повседневной работы многих компаний.

2) Установлено, что к настоящему времени разработано множество моделей для решения задачи прогнозирования временного ряда, среди которых наибольшую применимость имеют авторегрессионные и нейросетевые модели.

3) Выявлены достоинства и недостатки рассмотренных моделей. Установлено, что существенным недостатком авторегрессионных моделей является большое число свободных параметров, требующих идентификации; недостатками нейросетевых моделей является ее непрозрачность моделирования и сложность обучения сети.

4) Определено, что наиболее перспективным направлением развития моделей прогнозирования с целью повышения точности является создание комбинированных моделей, выполняющих на первом этапе кластеризацию, а затем прогнозирование временного ряда внутри установленного кластера.

ГЛАВА 2. МОДЕЛИ ЭКСТРАПОЛЯЦИИ ВРЕМЕННЫХ РЯДОВ ПО ВЫБОРКЕ МАКСИМАЛЬНОГО ПОДОБИЯ

2.1. Модель без учета внешних факторов

2.1.1. Выборки временного ряда

Пусть задан временной ряд $Z(t)=Z(1),Z(2),...,Z(T)$. Набор последовательных значений $Z_t^M=Z(t),Z(t+1),...,Z(t+M-1)$, лежащий внутри исходного временного ряда, назовем выборкой длины M с моментом начала отчета t; $M\in\{1,2,...,T\}$, $t\in\{1,2,..,T-M+1\}$. Фактически выборкой является кусочек временного ряда, имеющий точку начала отсчета и длину.

Две выборки одинаковой длины, принадлежащие одному временному ряду, обозначим через временную задержку k: $Z_t^M=Z(t),...,Z(t+M-1)$ и $Z_{t-k}^M=Z(t-k),...,Z(t-k+M-1)$; $k\in\{1,2,...,t-1\}$. На рисунках 2.1 — 2.4 показаны выборки различных временных рядов.

В работе [40] сформулирован подход к моделированию временных рядов по помощи выборок: «Pattern modelling refers to the process of describing the time series as a series of patterns». (Моделирование временных рядов при помощи выборок основано на предположении, что временной ряд представляет собой последовательность выборок.)

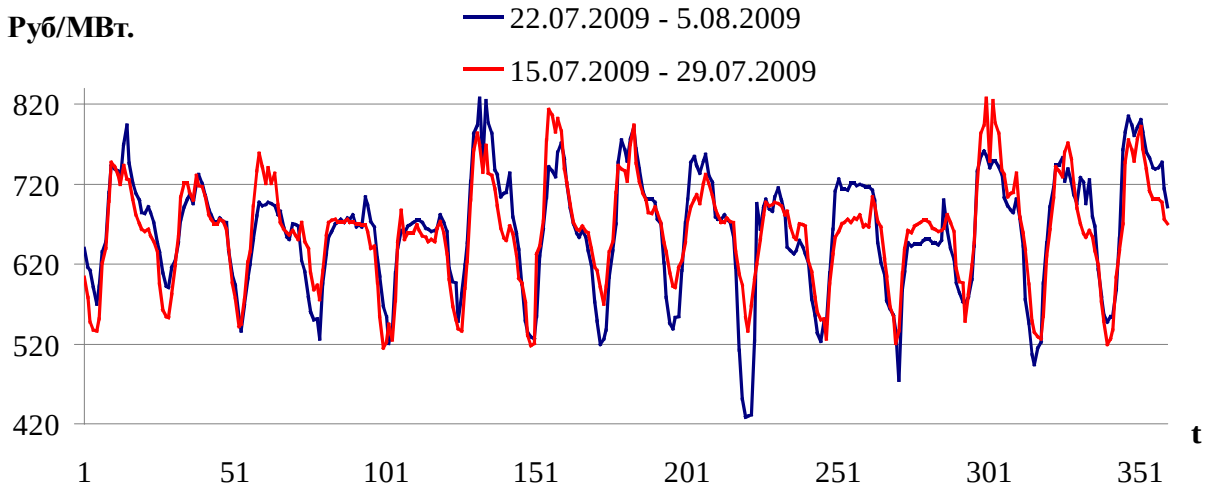


Рис. 2.1. Цена на электроэнергию европейской территории РФ

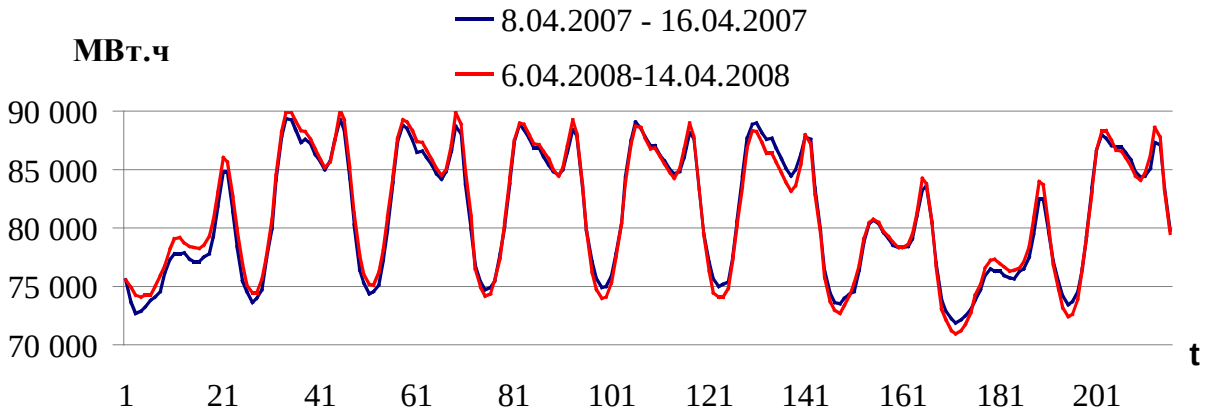


Рис. 2.2. Энергопотребление европейской территории РФ

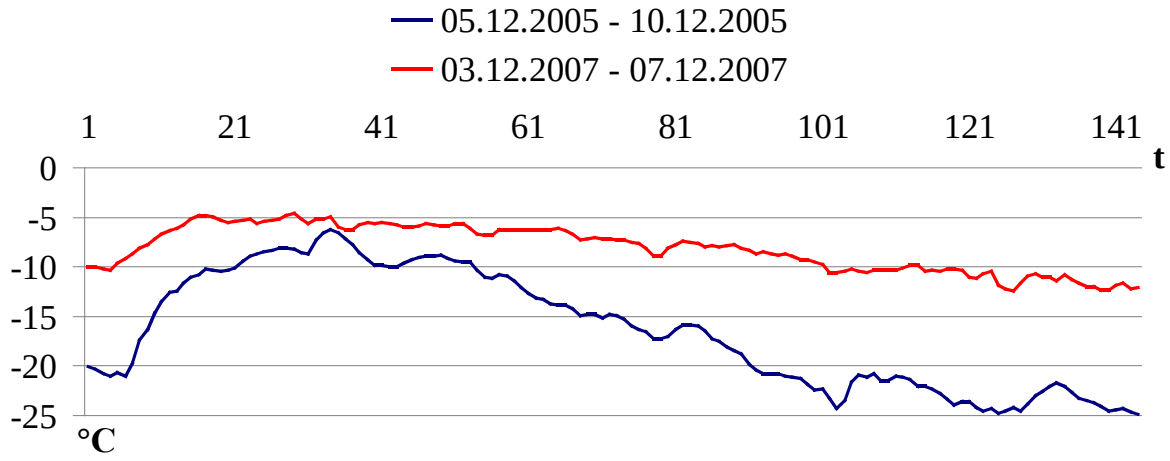


Рис. 2.3. Температура воздуха Новосибирской области

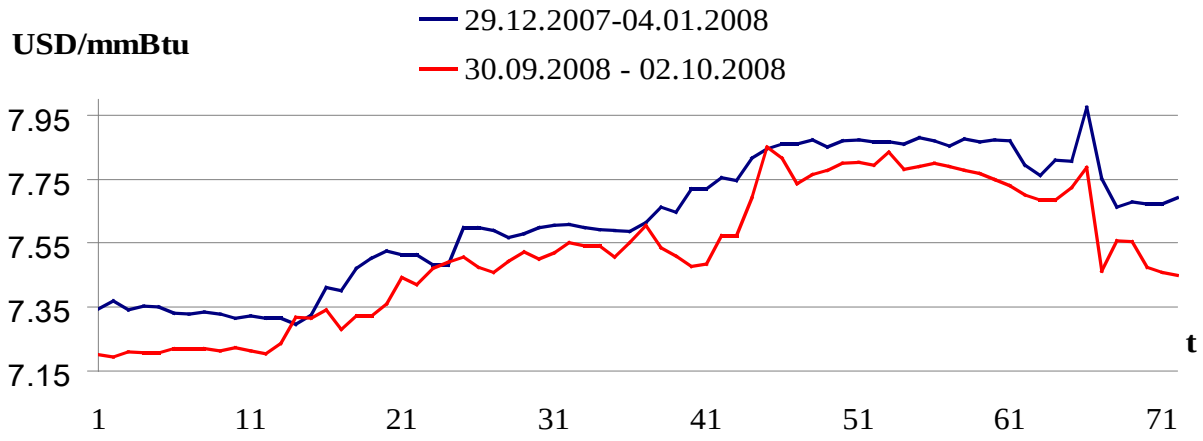


Рис. 2.4. Цена на природный газ на NYMEX

На рисунках 2.1 — 2.4 изображены выборки временных рядов, которые иллюстрируют свойство выборок, сформулированное в работе [42]: «Pieces of time series in the past might have a resemblance to pieces in the future». (Фактические выборки временного ряда могут иметь подобие с будущими выборками). Указанное свойство выборок будет использовано для построения модели прогнозирования.

В настоящем разделе сначала рассматривается задача аппроксимации одной выборки при помощи другой, а затем результаты аппроксимации применяются для построения модели прогнозирования временного ряда.

2.1.2. Аппроксимация выборки

Для расчетов перейдем к векторному обозначению выборки $\boldsymbol{Z}_t^{\boldsymbol{M}}=(Z(t),Z(t+1),\dots,Z(t+M-1))^T$ и временного ряда $\boldsymbol{Z}_1^{\boldsymbol{T}}=(Z(1),Z(2),\dots,Z(T))^T$. Здесь и далее, говоря о выборках временного ряда будем использовать обозначение Z_t^M, говоря о векторах, соответствующих указанным выборкам, будем писать жирную $\boldsymbol{Z}_t^{\boldsymbol{M}}$.

Используя свойство выборок повторяться, аппроксимируем выборку $\boldsymbol{Z}_t^{\boldsymbol{M}}$ при помощи выборки $\boldsymbol{Z}_{t-k}^{\boldsymbol{M}}$ следующим образом

$$\boldsymbol{Z}_{\boldsymbol{t}}^{\boldsymbol{M}}=\alpha_1 \boldsymbol{Z}_{\boldsymbol{t-k}}^{\boldsymbol{M}}+\alpha_0 \boldsymbol{I}^{\boldsymbol{M}}+\boldsymbol{E}^{\boldsymbol{M}}. \quad (2.1)$$

Здесь α_1 и α_0 — коэффициенты, $\boldsymbol{I}^{\boldsymbol{M}}$ — единичный вектор, $\boldsymbol{E}^{\boldsymbol{M}}$ — вектор значений ошибок аппроксимации. Выражение (2.1) можно переписать в виде

$$\hat{\boldsymbol{Z}}_{\boldsymbol{t}}^{\boldsymbol{M}}=\alpha_1 \boldsymbol{Z}_{\boldsymbol{t-k}}^{\boldsymbol{M}}+\alpha_0 \boldsymbol{I}^{\boldsymbol{M}}. \quad (2.2)$$

В формуле (2.2) $\hat{\boldsymbol{Z}}_{\boldsymbol{t}}^{\boldsymbol{M}}$ — аппроксимированные значения выборки $\boldsymbol{Z}_{\boldsymbol{t}}^{\boldsymbol{M}}$. Здесь и далее, говоря о вычисляемых (модельных) значениях выборки $\boldsymbol{Z}_{\boldsymbol{t}}^{\boldsymbol{M}}$ будем использовать обозначение $\hat{\boldsymbol{Z}}_{\boldsymbol{t}}^{\boldsymbol{M}}$ (с крышечкой).

Постановка задачи аппроксимации выборки. Пусть дана функциональная зависимость выборок (2.2). Необходимо определить такие значения α_1 и α_0, чтобы квадрат отклонений модельных значений выборки $\hat{\boldsymbol{Z}}_{\boldsymbol{t}}^{\boldsymbol{M}}$ от фактических $\boldsymbol{Z}_{\boldsymbol{t}}^{\boldsymbol{M}}$ был минимален

$$\sigma^2=\sum_{i=0}^{M-1}(Z(t+i)-\hat{Z}(t+i))^2\to min. \quad (2.3)$$

После нахождения коэффициентов α_1 и α_0, необходимо оценить вектор ошибок $\boldsymbol{E}^{\boldsymbol{M}}$.

Решение. Пусть дана линейная зависимость (2.1), тогда функция ошибки аппроксимации S_k^M для выборок $\boldsymbol{Z}_{\boldsymbol{t}}^{\boldsymbol{M}}$ и $\boldsymbol{Z}_{\boldsymbol{t-k}}^{\boldsymbol{M}}$ с задержкой k имеет вид

$$S_k^M(\alpha_1,\alpha_0)=\sum_{i=0}^{M-1}\sigma_i^2=\sum_{i=0}^{M-1}(Z(t+i)-\alpha_1 Z(t-k+i)-\alpha_0)^2. \quad (2.4)$$

Функция S_k^M называется функцией суммы квадратов (sum of squares function) [19]. Задача состоит в том, чтобы подобрать такие значения α_1 и α_0, чтобы при подстановке их в (2.4) было получено минимальное возможное значение $S_k^M(\alpha_1,\alpha_0)$. Рисунок 2.5 иллюстрирует определение функции суммы квадратов.

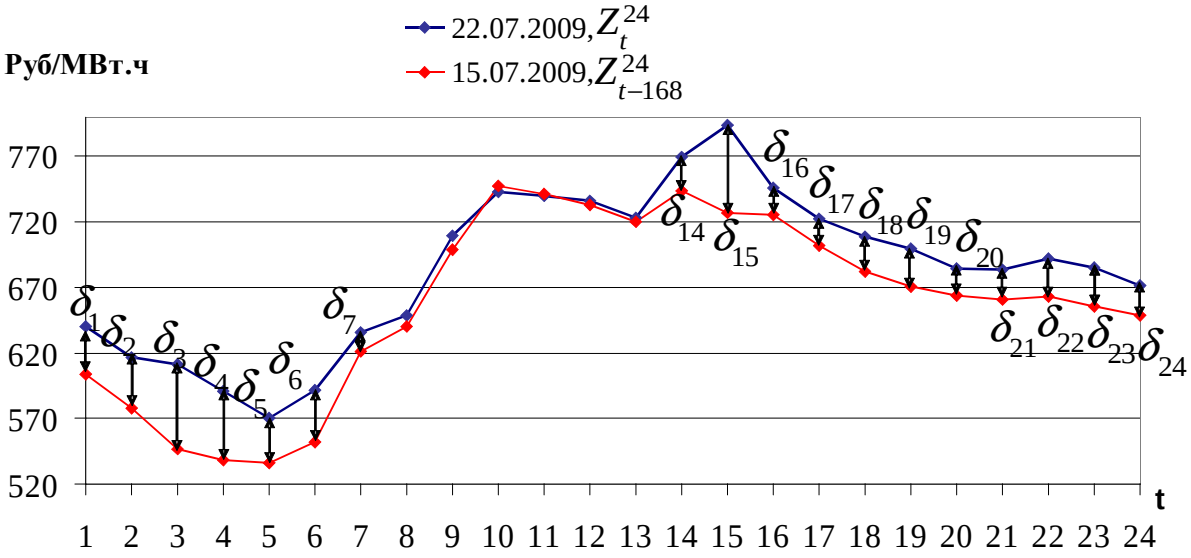


Рис. 2.5. Определение функции суммы квадратов S_k^M по выражению (2.4)

Решим задачу при помощи метода наименьших квадратов, подробно описанного в работе [19]. После ряда преобразований задача сводится к решению матричного уравнения

$$\boldsymbol{Z}_X \cdot \boldsymbol{A} = \boldsymbol{Z}_Y, \text{ где } \boldsymbol{A} = \begin{bmatrix} \alpha_1 \\ \alpha_0 \end{bmatrix}. \qquad (2.5)$$

Решая матричное уравнение (2.5), определяем значения коэффициентов α_1 и α_0, соответствующие минимуму функции $S_k^M(\alpha_1, \alpha_0)$. Уравнение (2.5) может быть решено любым известным методом. Исходные матрицы являются квадратными и решение может быть найдено, например, при помощи обратной матрицы

$$\boldsymbol{Z}_X^{-1} \cdot \boldsymbol{Z}_X \cdot \boldsymbol{A} = \boldsymbol{Z}_X^{-1} \cdot \boldsymbol{Z}_Y, \qquad (2.6)$$

$$\boldsymbol{A} = \boldsymbol{Z}_X^{-1} \cdot \boldsymbol{Z}_Y. \qquad (2.7)$$

Оценка ошибки аппроксимации. Ошибка определяется по формуле

$$\boldsymbol{E}^M = \boldsymbol{Z}_t^M - \hat{\boldsymbol{Z}}_t^M = \boldsymbol{Z}_t^M - (\alpha_1 \boldsymbol{Z}_{t-k}^M + \alpha_0 \boldsymbol{I}^M). \qquad (2.8)$$

В настоящее время точность моделирования временных рядов $\boldsymbol{E}^M$ принято

оценивать при помощи двух показателей [28]:

— средняя абсолютная ошибка (mean absolute error, MAE)

$$MAE = \frac{1}{M} \sum_{i=t}^{t+M-1} |Z(i) - \hat{Z}(i)| \,. \qquad (2.9)$$

— средняя абсолютная ошибка в процентах, (mean absolute percentage error, MAPE)

$$MAPE = \frac{1}{M} \sum_{i=t}^{t+M-1} \frac{|Z(i) - \hat{Z}(i)|}{Z(i)} \cdot 100\,\% \,. \qquad (2.10)$$

Здесь и далее, говоря о точности моделирования временных рядов (аппроксимации, прогнозирования) будем применять значения показателей MAE (2.9) и MAPE (2.10).

В настоящем разделе была рассмотрена аппроксимация одной выборки временного ряда при помощи другой, лежащей на оси времени на k отчетов левее, то есть раньше. Данное свойство представления новых выборок временного ряда при помощи старых будет использовано для определения модели экстраполяции.

2.1.3. Подобие выборок

Свойство двух выборок, заключенное в том, что одна выборка может быть выражена через другую с помощью линейной зависимости (2.1), назовем подобием двух выборок.

Покажем, что для общего случая линейной регрессии минимум ошибки регрессии соответствует максимуму линейной корреляции Пирсона. Пусть дана модель

$$\hat{\boldsymbol{Y}} = \alpha_1 \boldsymbol{X} + \alpha_0 \boldsymbol{I} \,. \qquad (2.11)$$

Тогда функция суммы квадратов определяется как разность модельных $\hat{Y}(i)$ и фактических значений $Y(i)$ некоторых наблюдений [19]

$$S_{reg}=\sum_{i=1}^{M}(\hat{Y}(i)-Y(i))^2 \tag{2.12}$$

Обозначим $\bar{Y}$ — среднее значение модельных и фактических наблюдений, которые по свойству линейной регрессии равны

$$\bar{Y}=\frac{1}{M}\sum_{i=1}^{M}Y(i)=\frac{1}{M}\sum_{i=1}^{M}\hat{Y}(i) \tag{2.13}$$

Из книги [19] известно, что сумма квадратов отклонений исследуемых наблюдений $Y(i)$ от среднего значения $\bar{Y}$ складывается из суммы квадратов отклонений модельных значений $\hat{Y}(i)$ от $\bar{Y}$ и суммы квадратов ошибок регрессии, определенной в выражении (2.12). Таким образом имеет место соотношение

$$\sum_{i=1}^{M}(Y(i)-\bar{Y})^2=\sum_{i=1}^{M}(\hat{Y}(i)-\bar{Y})^2+\sum_{i=1}^{M}(Y(i)-\hat{Y}(i))^2. \tag{2.14}$$

Выразим ошибку регрессии $\sum_{i=1}^{M}(Y(i)-\hat{Y}(i))^2$ и получим

$$\sum_{i=1}^{M}(Y(i)-\hat{Y}(i))^2=\sum_{i=1}^{M}(Y(i)-\bar{Y})^2-\sum_{i=1}^{M}(\hat{Y}(i)-\bar{Y})^2\rightarrow$$

$$S_{reg}=\sum_{i=1}^{M}(Y(i)-\hat{Y}(i))^2=\sum_{i=1}^{M}(Y(i)-\bar{Y})^2\cdot\left(1-\frac{\sum_{i=1}^{M}(\hat{Y}(i)-\bar{Y})^2}{\sum_{i=1}^{M}(Y(i)-\bar{Y})^2}\right). \tag{2.15}$$

Величина

$$R^2=\frac{\sum_{i=1}^{M}(\hat{Y}(i)-\bar{Y})^2}{\sum_{i=1}^{M}(Y(i)-\bar{Y})^2}\in[0,1] \tag{2.16}$$

называется квадратом множественного коэффициента корреляции. Иногда данный коэффициент называются коэффициентом детерминации.

Стоит обратить внимание на то, что при $Y(i)=const$ для $i\in\{1,2,3,...\}$ знаменатель R^2 обращается в ноль. Однако по свойству регрессии [19], для такого случая модельные значения также будут постоянными $\hat{Y}(i)=const$ для $i\in\{1,2,3,...\}$ и $R^2=1$. Известно, что на практике такой случай невозможен, в связи с тем, что значения $Y(i)$, как правило, являются результатами измерений.

Преобразовав (2.15) получим выражение

$$S_{reg}=\sum_{i=1}^{M}(Y(i)-\bar{Y})^2\cdot(1-R^2). \tag{2.17}$$

При этом сумма квадратов отклонений исследуемых наблюдений $Y(i)$ от среднего значения $\bar{Y}$ является величиной неизменной и характеризует свойство наблюдаемого процесса. Таким образом, из равенства (2.17) очевидно, что при $R^2\to max$, $S_{reg}\to min$.

Далее рассмотрим коэффициент линейной корреляции Пирсона ρ, определяемый выражением

$$\rho(X,Y)=\frac{\sum_{i=1}^{M}(X(i)-\bar{X})(Y(i)-\bar{Y})}{\sqrt{\sum_{i=1}^{M}(X(i)-\bar{X})^2\cdot\sum_{i=1}^{M}(Y(i)-\bar{Y})^2}}\in[-1,1]. \tag{2.18}$$

Согласно анализу, представленному в книге [19], связь двух рассматриваемых коэффициентов имеет следующий вид

$$\begin{gathered}\rho^2(\hat{Y},Y)=R^2\\ \rho(\hat{Y},Y)=sign(\alpha_1)\sqrt{R^2}.\end{gathered} \tag{2.19}$$

Таким образом, известно, что модуль величины $\rho(\hat{Y},Y)$ равен модулю величины R, а, следовательно, можно сформулировать следующее свойство

$$\begin{gathered}R^2\to max,\ S_{reg}\to min\\ |\rho(\hat{Y},Y)|\to max,\ S_{reg}\to min\end{gathered}. \tag{2.20}$$

Все подробности регрессионного анализа, а также связи рассматриваемых коэффициентов приведены в книге [19].

Вернемся к исследуемым выборкам временного ряда. Пусть дан временной ряд Z_1^T, для некоторой выборки Z_t^M, принадлежащей данному временному ряду, определим все значения $S_k^M(\alpha_1,\alpha_0)$ для $k\in\{1,2,\dots,t-1\}$, $M=const$. Далее, в множестве значений S_k^M найдем минимальное

$$S_{kmin}^M=min(S_{1,}^M S_{2,}^M \dots, S_{t-1}^M) \tag{2.21}$$

Согласно свойству (2.20) минимум ошибки регрессии S_{kmin}^M соответствует максимуму модуля коэффициента линейной корреляции (2.18). То есть если для $k\in\{1,2,\dots,t-1\}$ и $M=const$ определить множество значений модуля корреляции

$$\rho_k^M=|\rho(\hat{Z_t^M},Z_t^M)|=\frac{\left|\sum_{i=1}^{M}(\hat{Z}(t+i)-\bar{Z})(Z(t+i)-\bar{Z})\right|}{\sqrt{\sum_{i=1}^{M}(\hat{Z}(t+i)-\bar{Z})^2\cdot\sum_{i=1}^{M}(Z(t+i)-\bar{Z})^2}}\in[0,1], \tag{2.22}$$

а после определить максимальное значение полученного множества

$$\rho_{kmax}^M=max(\rho_{1,}^M \rho_{2,}^M \dots, \rho_{t-1}^M), \tag{2.23}$$

то задержка $kmin$ из выражения (2.21) и задержка $kmax$ из выражения (2.23) будут равны между собой, т. е. $kmin=kmax$. Проведенные расчеты подтверждают данное утверждение.

Определенную в (2.21) и (2.23) задержку, соответствующую минимуму ошибки регрессии S_{kmin}^M и максимуму модуля корреляции ρ_{kmax}^M обозначим $kmax$, а выборку Z_{t-kmax}^M назовем выборкой максимального подобия (most similar pattern). Выборка максимального подобия Z_{t-kmax}^M является выборкой, которая при подстановке в уравнение (2.2), дает в результате значения

выборки $\hat{Z}_t^M$, которая максимально точно описывает исходную выборку Z_t^M.

При реализации вычислений для определения выборки максимального подобия Z_{t-kmax}^M можно использовать как значения ошибки S_k^M, так и значения модуля корреляции ρ_k^M. В приведенном в диссертации примере использовалось значение коэффициента линейной корреляции (раздел 3.1.).

Гипотеза подобия. Если исходная выборка Z_t^M и модельная выборка $\hat{Z}_t^M$, полученная на основании (2.2) с использованием выборки Z_{t-kmax}^M, имеют значение величины ρ_{kmax}^M, близкое к единице, то для некоторых значений P и выборок Z_{t-kmax}^{M+P}, Z_t^{M+P} значение величины ρ_{kmax}^{M+P} также близко к единице.

Аналогичным образом можно сформулировать гипотезу подобия в случае учета ошибок регрессии S_k^M: если исходная выборка Z_t^M и модельная выборка $\hat{Z}_t^M$, полученная по формуле (2.2) на основании выборки Z_{t-kmin}^M, имеют минимальное значение ошибки $S_{kmin}^M = min(S_{1,}^M S_{2,}^M \dots, S_{t-1}^M)$, то для некоторых значений P и выборок Z_{t-kmin}^{M+P}, Z_t^{M+P} значение величины ошибки регрессии $S_{kmin}^{M+P} \to min(S_{1,}^{M+P} S_{2,}^{M+P} \dots, S_{t-1}^{M+P})$. Далее в работе использована первая формулировка гипотезы подобия.

Представленные в четвертой главе диссертации результаты прогнозирования подтверждают справедливость гипотезы для исследуемых временных рядов. Для временных рядов из других предметных областей справедливость гипотезы необходимо проверять.

2.1.4. Описание модели экстраполяции

Пусть дан временной ряд Z_1^T. Для данного временного ряда требуется определить значения Z_{T+1}^P. Используя свойство выборок, сформулированное

в разделе 2.1.3., выразим выборку $\boldsymbol{Z}_{T+1}^{P}$ через некоторую выборку $\boldsymbol{Z}_{\tau}^{P}$, лежащую внутри исходного временного ряда $\boldsymbol{Z}_{1}^{T}$

$$\hat{\boldsymbol{Z}}_{T+1}^{P}=\alpha_1 \boldsymbol{Z}_{\tau}^{P}+\alpha_0 \boldsymbol{I}^{P}. \qquad (2.24)$$

Алгоритм определения выборки $\boldsymbol{Z}_{\tau}^{P}$ состоит из двух шагов.

— Определение выборки $\boldsymbol{Z}_{kmax*}^{M}$.

— Вычисление $\boldsymbol{Z}_{\tau}^{P}$.

Рассмотрим подробно каждый шаг.

Определение выборки $\boldsymbol{Z}_{kmax*}^{M}$. На данном шаге для выборки $\boldsymbol{Z}_{T-M+1}^{M}$, содержащей значения процесса непосредственно перед моментом прогноза, находим выборку максимального подобия $\boldsymbol{Z}_{kmax*}^{M}$. Поиск выборки максимального подобия осуществляем перебором всех возможных значений задержек $k\in\{1,2,\dots,T-M-1\}$. Для каждого значения k из указанного диапазона решаем задачу аппроксимации (раздел 2.1.2.), в результате которой определяем коэффициенты α_1 и α_0, соответствующие k. Далее для найденной пары коэффициентов определяем значение модельной выборки $\hat{\boldsymbol{Z}}_{T-M+1}^{M}$, на основании которых вычисляем значение ρ_k^M (2.22). После того, как множество значений ρ_k^M для $k\in\{1,2,\dots,T-M-1\}$ получено, определяем значение ρ_{kmax}^{M} по выражению (2.23) и соответствующую выборку $\boldsymbol{Z}_{T-M+1-kmax}^{M}$. Для упрощения обозначим задержку $kmax*=T-M+1-kmax$ и выборку максимального подобия $\boldsymbol{Z}_{kmax*}^{M}$ (SimilarHistory).

Вычисление $\boldsymbol{Z}_{\tau}^{P}$. В соответствии с гипотезой подобия, сформулированной в разделе 2.1.3., в качестве выборки $\boldsymbol{Z}_{\tau}^{P}$ используем выборку $\boldsymbol{Z}_{kmax*+M}^{P}$, то есть выборку, расположенную на оси времени сразу за

выборкой максимального подобия. Выборку $\boldsymbol{Z}_{\tau}^{\boldsymbol{P}}$ назовем базовой историей (BaseHistory). На рисунке 2.6 представлено расположение всех рассмотренных выборок.

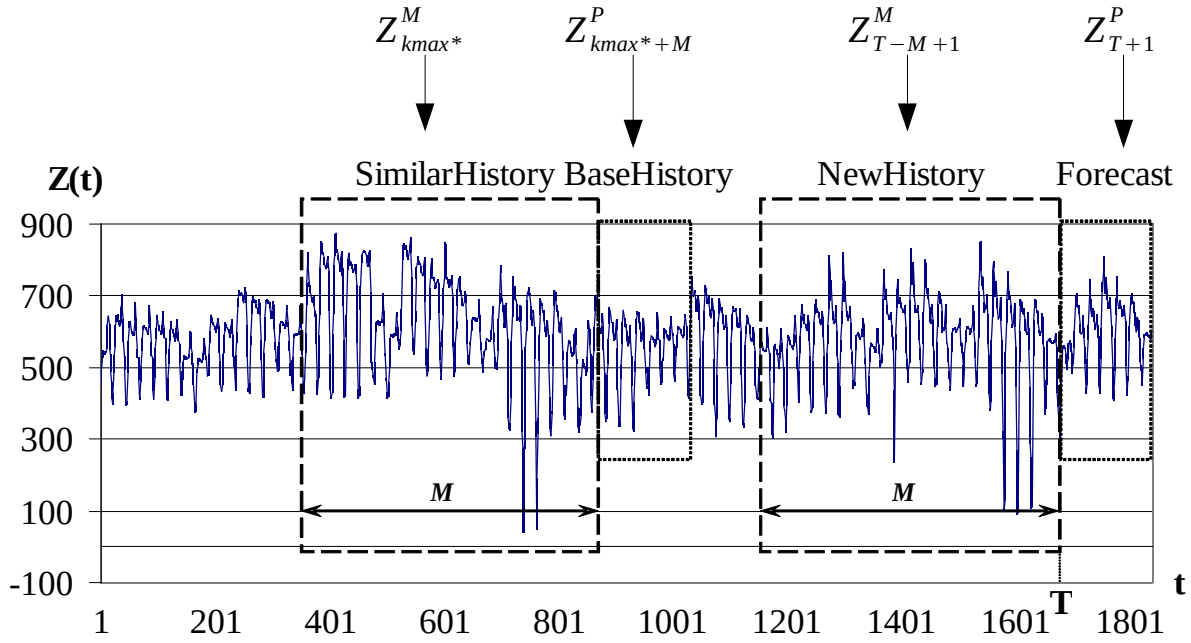


Рис. 2.6. Положение выборок Z_{kmax*}^{M}, $Z_{kmax*+M}^{P}$, Z_{T-M+1}^{M} и Z_{T+1}^{P}

Экстраполированные значения выборки $\hat{\boldsymbol{Z}}_{\boldsymbol{T+1}}^{\boldsymbol{P}}$ (прогноз, Forecast) определяются по формуле

$$\hat{\boldsymbol{Z}}_{\boldsymbol{T+1}}^{\boldsymbol{P}} = \alpha_1 \boldsymbol{Z}_{\boldsymbol{kmax*+M}}^{\boldsymbol{P}} + \alpha_0 \boldsymbol{I}^{\boldsymbol{P}} = EMMSP(M), \qquad (2.25)$$

которая представляет собой модель экстраполяции временных рядов по выборке максимального подобия (extrapolation model on most similar pattern, далее EMMSP). В работе [40] используется термин closest pattern, в работе [42] — closest piece, в работе [41] — most similar vector.

Особенности EMMSP:

— модель относится к классу авторегрессионных моделей прогнозирования;

— модель может работать с неравноотстоящими временными

рядами;

— модель работает со стационарными и нестационарными временными рядами;

— модель имеет один параметр M, определение которого подробно рассмотрено в третьей главе диссертации;

— экстраполяция P значений временного ряда производится за одну итерацию.

2.2. Модель с учетом внешних факторов

2.2.1. Выборки временных рядов

Пусть дан временной ряд $Z(t)$ и внешние факторы, представленные в виде временных рядов $X_1(t)$,..., $X_S(t)$, соответствующие исходному ряду по отметкам времени. Требуется построить модель прогнозирования исходного временного ряда $Z(t)$, которая будет учитывать влияние значений внешних факторов $X_1(t)$,..., $X_S(t)$.

Если внешние факторы $X_1(t)$,..., $X_S(t)$ имеют временное разрешение, отличное от разрешения исходного ряда $Z(t)$, то необходимо произвести дополнительные преобразования и привести внешние факторы в соответствие с исходным временном рядом $Z(t)$ по отметкам времени.

В основу модели экстраполяции с учетом внешнего фактора положено предположение о повторяемости выборок временного ряда (2.1.1.). Кроме того, известно, что для учета внешних факторов в авторегрессионных моделях вводятся дополнительные регрессоры (раздел 1.3.2.).

В настоящем разделе сначала рассматривается задача аппроксимации одной выборки при помощи другой с учетом выборок внешних факторов, а затем результаты аппроксимации применяются для построения модели

прогнозирования временного ряда.

2.2.2. Аппроксимация выборки

Пусть задана выборка исходного временного ряда Z_t^M и выборки $X_{(1)t}^M$, ..., $X_{(S)t}^M$ являются выборками внешних факторов, соответствующих Z_t^M по отметкам времени. Аппроксимируем выборку Z_t^M с учетом выборок $X_{(1)t}^M$,..., $X_{(S)t}^M$ по формуле

$$\boldsymbol{Z_t^M} = \alpha_{S+1}\boldsymbol{Z_{t-k}^M} + \alpha_S \boldsymbol{X_{(S)t}^M} + \dots \alpha_1 \boldsymbol{X_{(1)t}^M} + \alpha_0 \boldsymbol{I^M} + \boldsymbol{E^M}. \quad (2.26)$$

Здесь $\alpha_{S+1}, \alpha_S, ..., \alpha_0$ — коэффициенты, определяемые по методу наименьших квадратов. Вектор $\boldsymbol{E^M}$ — вектор значений ошибок аппроксимации, $\boldsymbol{I^M}$ — единичный вектор. Выражение (2.26) можно переписать в виде

$$\boldsymbol{\hat{Z}_t^M} = \alpha_{S+1}\boldsymbol{Z_{t-k}^M} + \alpha_S \boldsymbol{X_{(S)t}^M} + \dots \alpha_1 \boldsymbol{X_{(1)t}^M} + \alpha_0 \boldsymbol{I^M}. \quad (2.27)$$

Постановка задачи аппроксимации выборки. Пусть дана функциональная зависимость (2.27). Необходимо определить такие значения коэффициентов $\alpha_{S+1}, \alpha_S, ..., \alpha_0$, чтобы квадрат отклонений модельных значений выборки $\boldsymbol{\hat{Z}_t^M}$ от фактических $\boldsymbol{Z_t^M}$ был минимален

$$\sigma^2 = \sum_{i=0}^{M-1} (Z(t+i) - \hat{Z}(t+i))^2 \to min \quad (2.28)$$

После нахождения коэффициентов $\alpha_{S+1}, \alpha_S, ..., \alpha_0$ необходимо оценить вектор ошибок $\boldsymbol{E^M}$.

Решение. Пусть дана функциональная зависимость (2.26), тогда функция суммы квадратов имеет вид

$$S_k^M(\alpha_{S+1}, \dots, \alpha_0) = \sum_{i=0}^{M-1} (Z(t+i) - (\alpha_{S+1} Z(t-k+i) + \alpha_S X_S(t+i) + \dots \alpha_1 X_1(t+i) + \alpha_0))^2 \quad (2.29)$$

Повторяя рассуждения, приведенные в разделе 2.1.2., приведем задачу

к матричному уравнению

$$\boldsymbol{Z_X}\cdot\boldsymbol{A}=\boldsymbol{Z_Y}, \text{ где } \boldsymbol{A}=\begin{bmatrix}\alpha_{S+1}\\ \alpha_S\\ ...\\ \alpha_1\\ \alpha_0\end{bmatrix} \tag{2.30}$$

Решая матричное уравнение (2.30), определяем значения коэффициентов $\alpha_{S+1}, \alpha_S, ..., \alpha_0$, соответствующие минимуму функции $S_k^M(\alpha_{S+1},...,\alpha_0)$. Решение, как и в предыдущем случае, будет найдено при помощи обратной матрицы

$$\boldsymbol{Z_X^{-1}}\cdot\boldsymbol{Z_X}\cdot\boldsymbol{A}=\boldsymbol{Z_X^{-1}}\cdot\boldsymbol{Z_Y}, \tag{2.31}$$

$$\boldsymbol{A}=\boldsymbol{Z_X^{-1}}\cdot\boldsymbol{Z_Y}. \tag{2.32}$$

Оценка вектора ошибок аппроксимации $\boldsymbol{E^M}$ описана в разделе 2.1.2.

2.2.3. Подобие выборок

Пусть дана выборка исходного временного ряда Z_t^M и выборки внешних факторов $X_{(1)t}^M, ..., X_{(S)t}^M$. Модельная выборка $\hat{Z}_t^M$ определяется линейной зависимостью (2.27). Вычислим все значения $S_k^M(\alpha_{S+1},...,\alpha_0)$ для задержек $k\in\{1,2,...,t-1\}$ и $M=const$. Далее, в множестве значений S_k^M найдем минимальное по выражению (2.21). В случае множественной регрессии (2.27) равенство минимума ошибки регрессии (2.21) и максимума модуля линейной корреляции (2.23) не выполняется, а потому в данном случае следует рассматривать только ошибку регрессии $S_k^M(\alpha_{S+1},...,\alpha_0)$.

Определенную в выражении (2.21) задержку, соответствующую минимуму ошибки регрессии S_{kmin}^M обозначим по аналогии с предыдущим случаем $kmax$, а выборку Z_{t-kmax}^M назовем выборкой максимального подобия

(most similar pattern). Выборка максимального подобия Z^M_{t-kmax} является выборкой, которая при подстановке в уравнение (2.27), дает в результате значения выборки $\hat{Z}^M_t$, которая максимально точно описывает исходную выборку Z^M_t с учетом выборок внешних факторов $X^M_{(1)t},..., X^M_{(S)t}$. Величину минимальной ошибки S^M_{kmin}, соответствующую задержке *kmax* будем далее обозначать S^M_{kmax}.

Гипотеза подобия. Если исходная выборка Z^M_t и модельная выборка $\hat{Z}^M_t$, полученная на основании (2.27) с использованием выборки Z^M_{t-kmax} и выборок внешних факторов $X^M_{(1)t},..., X^M_{(S)t}$, имеют минимальное значение величины S^M_{kmax}, то для некоторых значений P и выборок Z^{M+P}_{t-kmax}, $X^{M+P}_{(1)t}$,..., $X^{M+P}_{(S)t}$, Z^{M+P}_t значение величины S^{M+P}_{kmax} также стремится к минимальному.

Представленные в четвертой главе диссертации результаты прогнозирования подтверждают справедливость гипотезы для исследуемых временных рядов. Для временных рядов из других предметных областей справедливость гипотезы необходимо проверять дополнительно.

2.2.4. Описание модели

Пусть дан исходный временной ряд Z^T_1 и внешние факторы $X^{T+P}_{(1)1},..., X^{T+P}_{(S)1}$. Для исходного временного ряда требуется определить значения Z^P_{T+1}, учитывая доступные значения Z^T_1 и $X^P_{(1)T+1},..., X^P_{(S)T+1}$. Используя свойство выборок, сформулированное в разделе 2.2.3., выразим выборку $\boldsymbol{Z^P_{T+1}}$ через некоторую выборку $\boldsymbol{Z^P_\tau}$, лежащую внутри исходного временного ряда $\boldsymbol{Z^T_1}$, и выборки $\boldsymbol{X^P_{(1)T+1}},..., \boldsymbol{X^P_{(S)T+1}}$ следующим образом

$$\hat{\boldsymbol{Z}}^{\boldsymbol{P}}_{\boldsymbol{T+1}} = \alpha_{S+1}\boldsymbol{Z}^{\boldsymbol{P}}_{\boldsymbol{\tau}} + \alpha_S \boldsymbol{X}^{\boldsymbol{P}}_{\boldsymbol{(S)T+1}} + ... + \alpha_1 \boldsymbol{X}^{\boldsymbol{P}}_{\boldsymbol{(1)T+1}} + \alpha_0 \boldsymbol{I}^{\boldsymbol{P}}. \qquad (2.33)$$

Алгоритм определения выборки $\boldsymbol{Z}_{\boldsymbol{\tau}}^{\boldsymbol{P}}$ состоит из двух шагов:

— определение выборки $\boldsymbol{Z}_{\boldsymbol{kmax*}}^{\boldsymbol{M}}$;

— вычисление $\boldsymbol{Z}_{\boldsymbol{\tau}}^{\boldsymbol{P}}$.

Определение выборки $\boldsymbol{Z}_{\boldsymbol{kmax*}}^{\boldsymbol{M}}$. На данном шаге, как в случае экстраполяции без учета внешних факторов, для выборки $\boldsymbol{Z}_{\boldsymbol{T-M+1}}^{\boldsymbol{M}}$, содержащей значения процесса непосредственно перед моментом прогноза, находим выборку максимального подобия $\boldsymbol{Z}_{\boldsymbol{kmax*}}^{\boldsymbol{M}}$. Поиск выборки максимального подобия осуществляем перебором всех возможных значений задержек $k \in \{1,2,\ldots,T-M-1\}$. Для каждого значения k из указанного диапазона решаем задачу аппроксимации (2.2.2.), в результате которой определяем коэффициенты $\alpha_{S+1}, \alpha_S, \ldots, \alpha_0$, соответствующие k. Далее для вычисленных коэффициентов определяем значение модельной выборки, на основании которых вычисляем значение ошибки регрессии S_k^M (2.22). После того, как множество значений S_k^M для $k \in \{1,2,\ldots,T-M-1\}$ получено, определяем значение S_{kmax}^M (2.23) и соответствующую выборку $\boldsymbol{Z}_{\boldsymbol{T-M+1-kmax}}^{\boldsymbol{M}}$. Как и ранее обозначим задержку $kmax* = T-M+1-kmax$, а выборку максимального подобия $\boldsymbol{Z}_{\boldsymbol{kmax*}}^{\boldsymbol{M}}$ (SimilarHistory).

Вычисление $\boldsymbol{Z}_{\boldsymbol{\tau}}^{\boldsymbol{P}}$**.** В соответствии с гипотезой подобия, сформулированной в разделе 2.2.3., в качестве выборки $\boldsymbol{Z}_{\boldsymbol{\tau}}^{\boldsymbol{P}}$ в выражении (2.33) используем выборку $\boldsymbol{Z}_{\boldsymbol{kmax*+M}}^{\boldsymbol{P}}$, то есть выборку, расположенную на оси времени сразу за выборкой максимального подобия. Для случая учета одного внешнего фактора положение выборок показано на рисунке 2.7.

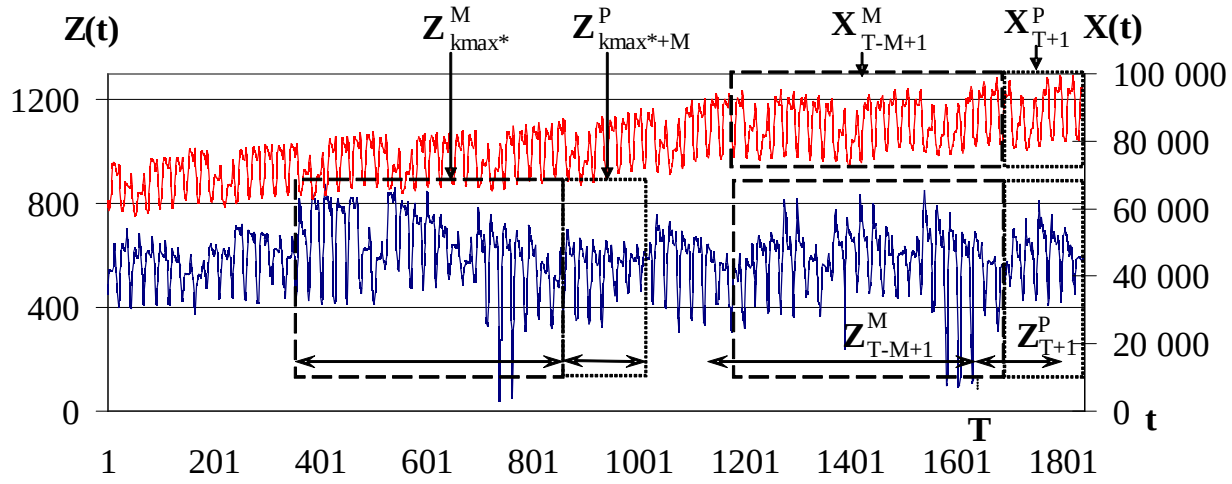


Рис. 2.7. Положение выборок на оси времени

Для удобства назовем выборки следующим образом:

— выборка новой истории $\boldsymbol{Z}^{\boldsymbol{M}}_{\boldsymbol{T-M+1}}$ (NewHistory);

— выборка максимального подобия $\boldsymbol{Z}^{\boldsymbol{M}}_{\boldsymbol{kmax*}}$ (SimilarHistory);

— выборка базовой истории $\boldsymbol{Z}^{\boldsymbol{P}}_{\boldsymbol{\tau}}$ (BaseHistory);

— выборка истории внешнего фактора $\boldsymbol{X}^{\boldsymbol{M}}_{\boldsymbol{T-M+1}}$ (FactorHistory);

— выборка прогноза внешнего фактора $\boldsymbol{X}^{\boldsymbol{P}}_{\boldsymbol{T+1}}$ (FactorForecast).

Экстраполированные значения выборки $\hat{\boldsymbol{Z}}^{\boldsymbol{P}}_{\boldsymbol{T+1}}$ (Forecast) определяем по формуле

$$\hat{\boldsymbol{Z}}^{\boldsymbol{P}}_{\boldsymbol{T+1}}=\alpha_{S+1}\boldsymbol{Z}^{\boldsymbol{P}}_{\boldsymbol{kmax*+M}}+\alpha_S\boldsymbol{X}^{\boldsymbol{P}}_{\boldsymbol{(S)T+1}}+\ldots+\alpha_1\boldsymbol{X}^{\boldsymbol{P}}_{\boldsymbol{(1)T+1}}+\alpha_0\boldsymbol{I}^{\boldsymbol{P}}=EMMSPX(M), \qquad (2.34)$$

которая представляет собой расширенную модель экстраполяции временных рядов по выборке максимального подобия (extrapolation model on most similar pattern extended, далее EMMSPX).

Необходимо заметить, что внешние факторы, как правило, оказывают влияние на исследуемый процесс в соответствующий момент времени. Например, из работы [9] известно, что на энергопотребление оказывает существенное влияние температура окружающей среды, а именно: при резких изменениях температуры скачкообразно меняется энергопотребление

(рис. 2.8). Таким образом, по свойству процессов учет внешних факторов необходим в соответствующий момент времени. В случае, если отсутствуют значения внешних факторов на будущий период, а именно, значения выборок $X^{P}_{(1)T+1}$,..., $X^{P}_{(S)T+1}$, то необходимо или спрогнозировать прежде внешние факторы, а потом исследуемый основной процесс, либо удалить внешние факторы из модели. В случае прогноза энергопотребления часто в качестве внешнего фактора используется прогноз температуры окружающей среды, формируемый Гидрометцентром России.

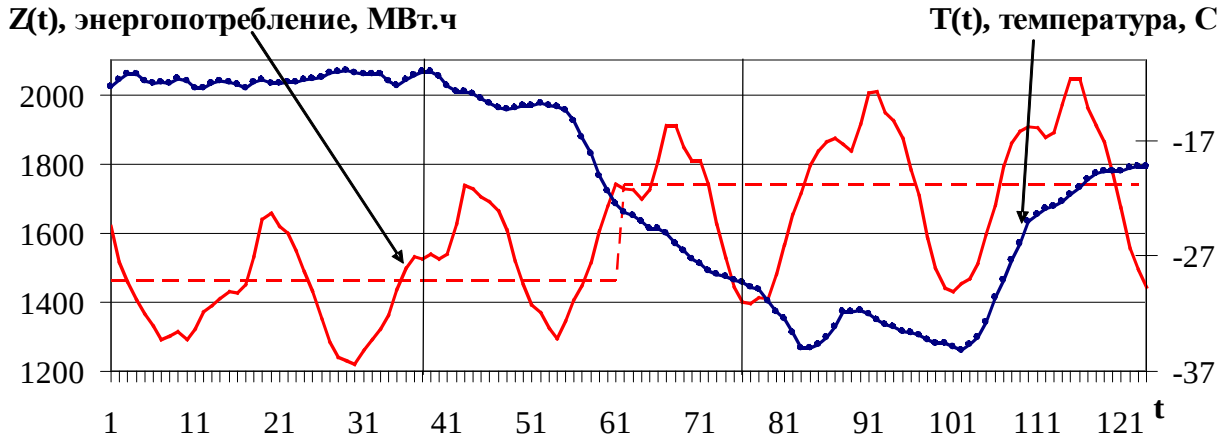


Рис. 2.8. Влияние температуры окружающей среды на энергопотребление Новосибирской области

На рисунке 2.8 представлены графики энергопотребления Новосибирской области, а также температура окружающей среды той же территории. Заметно, что резкое снижение температуры приводит к скачкообразному повышению энергопотребления. Таким образом, по свойству данных процессов учет температуры при прогнозе энергопотребления должен производиться в соответствующие моменты времени.

В завершении данного раздела необходимо отметить особенности EMMSPX:

— модель относится к классу авторегрессионных моделей прогнозирования, учитывающих дискретные внешние факторы;

— модель может учитывать несколько дискретных внешних факторов;

— модель эффективно работает с неравноотстоящими временными рядами;

— модель работает со стационарными и нестационарными временными рядами;

— модель имеет один параметр M, определение которого подробно рассмотрено в третьей главе диссертации;

— экстраполяция P значений временного ряда производится за одну итерацию.

2.3. Варианты моделей по выборке максимального подобия

Модель (2.34) можно разделить на две части — модель авторегрессии и модель внешнего фактора

$$\hat{\boldsymbol{Z}}^{\boldsymbol{P}}_{\boldsymbol{T+1}} = \underbrace{\alpha_2 \boldsymbol{Z}^{\boldsymbol{P}}_{\boldsymbol{kmax*+M}} + \alpha_0 \boldsymbol{I}^{\boldsymbol{P}}}_{\text{авторегрессия}} + \underbrace{\alpha_1 \boldsymbol{X}^{\boldsymbol{P}}_{\boldsymbol{T+1}}}_{\text{внешний фактор}} . \tag{2.35}$$

Модель авторегрессии и модель внешнего фактора могут быть модифицированы с целью повышения точности прогнозирования.

Авторегрессионная EMMSP со множеством выборок. Прогнозные значения $\hat{\boldsymbol{Z}}^{\boldsymbol{P}}_{\boldsymbol{T+1}}$ вычисляются как линейная комбинация нескольких выборок базовой истории с различными задержками

$$\hat{\boldsymbol{Z}}^{\boldsymbol{P}}_{\boldsymbol{T+1}} = \alpha_q \boldsymbol{Z}^{\boldsymbol{P}}_{\boldsymbol{\tau_q}} + \alpha_{q-1} \boldsymbol{Z}^{\boldsymbol{P}}_{\boldsymbol{\tau_{q-1}}} + \ldots + \alpha_1 \boldsymbol{Z}^{\boldsymbol{P}}_{\boldsymbol{\tau_1}} + \alpha_0 \boldsymbol{I}^{\boldsymbol{P}} . \tag{2.36}$$

В рамках диссертации проводились исследования эффективности

увеличения количества выборок, принимаемых в расчет. На практике количество используемых выборок исходного временного ряда не превышает двух, т. е. прогнозные значения $\hat{\boldsymbol{Z}}^{\boldsymbol{P}}_{\boldsymbol{T+1}}$ вычисляются как линейная комбинация двух выборок базовой истории с различными задержками

$$\hat{\boldsymbol{Z}}^{\boldsymbol{P}}_{\boldsymbol{T+1}} = \alpha_2 \boldsymbol{Z}^{\boldsymbol{P}}_{\boldsymbol{\tau}_2} + \alpha_1 \boldsymbol{Z}^{\boldsymbol{P}}_{\boldsymbol{\tau}_1} + \alpha_0 \boldsymbol{I}^{\boldsymbol{P}}. \tag{2.37}$$

Однако не исключено, то в ряде задач полезным будет использование трех-четырех выборок. Данная модель (2.37) применялась для прогнозирования временных рядов цен на электроэнергию [45].

Авторегрессионная EMMSP с использованием q-той степени значений выборок. Модель прогнозирования представляет собой линейную комбинацию степеней выборки максимального подобия

$$\hat{\boldsymbol{Z}}^{\boldsymbol{P}}_{\boldsymbol{T+1}} = \alpha_q (\boldsymbol{Z}^{\boldsymbol{P}}_{\boldsymbol{\tau}})^q + \alpha_{q-1} (\boldsymbol{Z}^{\boldsymbol{P}}_{\boldsymbol{\tau}})^{(q-1)} + \ldots + \alpha_1 \boldsymbol{Z}^{\boldsymbol{P}}_{\boldsymbol{\tau}} + \alpha_0 \boldsymbol{I}^{\boldsymbol{P}}. \tag{2.38}$$

Здесь $(\boldsymbol{Z}^{\boldsymbol{P}}_{\boldsymbol{\tau}})^n$ — выборка, значениями которой являются n-ные степени значений временного ряда

$$(\boldsymbol{Z}^{\boldsymbol{P}}_{\boldsymbol{\tau}})^n = \begin{bmatrix} Z^n(\tau) \\ \ldots \\ Z^n(\tau + P) \end{bmatrix} \tag{2.39}$$

В рамках диссертации проводились исследования эффективности повышения степени выборки максимального подобия, которые показали, что точность прогнозирования для некоторых временных рядов повышается при использовании второй степени, то есть модели

$$\hat{\boldsymbol{Z}}^{\boldsymbol{P}}_{\boldsymbol{T+1}} = \alpha_2 (\boldsymbol{Z}^{\boldsymbol{P}}_{\boldsymbol{\tau}})^2 + \alpha_1 \boldsymbol{Z}^{\boldsymbol{P}}_{\boldsymbol{\tau}} + \alpha_0 \boldsymbol{I}^{\boldsymbol{P}} \tag{2.40}$$

Дальнейшее увеличение степени выборки максимального подобия эффективности модели не повышает.

EMMSPX с использованием q-той степени значений внешних факторов. В случае доступности данных лишь по одному внешнему фактору

возможна модификация модели аналогично (2.38)

$$\hat{\boldsymbol{Z}}^{\boldsymbol{P}}_{\boldsymbol{T+1}} = \alpha_1 \boldsymbol{Z}^{\boldsymbol{P}}_{\boldsymbol{\tau}} + \alpha_0 \boldsymbol{I}^{\boldsymbol{P}} + \beta_q (\boldsymbol{X}^{\boldsymbol{P}}_{\boldsymbol{T}})^q + \beta_{q-1} (\boldsymbol{X}^{\boldsymbol{P}}_{\boldsymbol{T}})^{(q-1)} + \ldots + \beta_1 \boldsymbol{X}^{\boldsymbol{P}}_{\boldsymbol{T}}. \quad (2.41)$$

При этом значения выборки $(\boldsymbol{X}^{\boldsymbol{P}}_{\boldsymbol{T}})^n$ определяются как n-ные степени значений внешнего фактора (2.39).

Важно отметить, что при построении моделей (2.36) — (2.41) применяется единообразный подход к анализу и проектированию моделей, описанный в разделе 2.2. настоящей работы.

Ограничения применимости модели экстраполяции временных рядов по выборке максимального подобия. Предложенные в настоящей главе модели EMMSP и EMMSPX могут применяться для прогнозирования временных рядов на P значений вперед при выполнении набора условий.

— Длина временного ряда составляет не менее $500P - 700P$.

— Временной ряд равноотстоящий; в случае неравноотстоящего временного ряда применение модели возможно только при условии, что его равноотстоящие отрезки должны быть содержать не менее, чем $100P - 150P$ отсчетов.

— Временной ряд относится к классу временных рядов с длинной памятью.

— Задача прогнозирования на P значений вперед относится к классу краткосрочного или среднесрочного прогнозирования данного типа временного ряда. Не рекомендуется использовать разработанную модель для долгосрочного прогнозирования.

— В случае учета набора дискретных внешних факторов, их временное разрешение должно быть приведено к разрешению исходного временного ряда. Длина исходного временного ряда и временных рядов внешних факторов может не совпадать.

2.4. Выводы

1) В настоящей главе предложены модели экстраполяции временных рядов по выборке максимального подобия с учетом и без учета внешних факторов.

2) Предложенные модели относятся к классу авторегрессионных моделей прогнозирования и обладают всеми достоинствами, характерными для данного класса.

3) Предложенные модели устраняют существенный недостаток указанного класса — большое число свободных параметров, требующих идентификации. Обе модели экстраполяции по выборке максимального подобия имеют единственный параметр.

4) Разработаны варианты моделей по выборке максимального подобия, использование которых может приводить к повышению точности прогнозирования временного ряда.

ГЛАВА 3. МЕТОД ПРОГНОЗИРОВАНИЯ НА МОДЕЛИ ЭКСТРАПОЛЯЦИИ ПО ВЫБОРКЕ МАКСИМАЛЬНОГО ПОДОБИЯ

Метод прогнозирования на модели экстраполяции временных рядов по выборке максимального подобия содержит следующие алгоритмы.

1) Алгоритм экстраполяции временного ряда без учета внешних факторов.
2) Алгоритм экстраполяции временного ряда с учетом внешних факторов.
3) Алгоритм идентификации модели.
4) Алгоритм построения доверительного интервала для прогнозных значений.

Далее подробно рассмотрены все перечисленные выше алгоритмы.

3.1. Алгоритм экстраполяции временного ряда без учета внешних факторов

Модель экстраполяции временного ряда без учета внешних факторов сформулирована в разделе 2.1. диссертации. Алгоритм экстраполяции состоит из следующих шагов.

1) Определить выборку новой истории.
2) Определить выборку максимального подобия.
3) Определить выборку базовой истории.
4) Вычислить прогнозные значения.

Далее приведем описание каждого указанного выше шага, иллюстрируя расчеты решением следующей задачи. Пусть даны значения временного ряда цен на электроэнергию европейской территории РФ (№1 в таблице 7) с

01.09.2006 до 22.06.2009; длина временного ряда равна 24 624. Обозначим временной ряд $Z(t)$. Требуется определить 24 значения временного ряда цен за 23.06.2009. Считаем параметр модели $M=216$ заданным.

1) Определить выборку новой истории.

Выборкой новой истории является выборка временного ряда, значения которой предшествуют моменту прогноза T. В текущей постановке задачи выборка новой истории равна $\boldsymbol{Z}_{\boldsymbol{T-M+1}}^{\boldsymbol{M}}=\boldsymbol{Z}_{\boldsymbol{24409}}^{\boldsymbol{216}}$.

2) Определить выборку максимального подобия.

Для определения выборки максимального подобия необходимо определить значения модуля линейной корреляции ρ_k^M для выборки $\boldsymbol{Z}_{\boldsymbol{T-M+1}}^{\boldsymbol{M}}$ и всех выборок с задержкой $k\in\{1,2,\ldots,T-M-1\}$. При этом для каждого значения k из указанного диапазона требуется решить задачу аппроксимации выборки $\boldsymbol{Z}_{\boldsymbol{T-M+1}}^{\boldsymbol{M}}$ при помощи выборки $\boldsymbol{Z}_{\boldsymbol{T-M+1-k}}^{\boldsymbol{M}}$. Обозначим момент времени $T-M+1=t$ и решим данную задачу.

Вычислим аппроксимированные значения выборки

$$\hat{\boldsymbol{Z}}_{\boldsymbol{T-M+1}}^{\boldsymbol{M}}=\alpha_1\boldsymbol{Z}_{\boldsymbol{T-M+1-k}}^{\boldsymbol{M}}+\alpha_0\boldsymbol{I}^{\boldsymbol{M}}, \tag{3.1}$$

а с учетом обозначения

$$\hat{\boldsymbol{Z}}_{\boldsymbol{t}}^{\boldsymbol{M}}=\alpha_1\boldsymbol{Z}_{\boldsymbol{t-k}}^{\boldsymbol{M}}+\alpha_0\boldsymbol{I}^{\boldsymbol{M}}. \tag{3.2}$$

Согласно методу наименьших квадратов, коэффициенты аппроксимации определим, исходя из уравнения

$$\boldsymbol{Z}_{\boldsymbol{X}}\cdot\boldsymbol{A}=\boldsymbol{Z}_{\boldsymbol{Y}}, \tag{3.3}$$

где значение элементов матрицы $\boldsymbol{Z}_{\boldsymbol{X}}$ и $\boldsymbol{Z}_{\boldsymbol{Y}}$ определяются следующим образом:

$$\boldsymbol{Z}_X = \begin{bmatrix} \sum_{i=0}^{M-1} Z^2(k+i) & \sum_{i=0}^{M-1} Z(k+i) \\ \sum_{i=0}^{M-1} Z(k+i) & M \end{bmatrix};$$

$$, \boldsymbol{Z}_Y = \begin{bmatrix} \sum_{i=0}^{M-1} Z(k+i)\cdot Z(T-M+1+i) \\ \sum_{i=0}^{M-1} Z(T-M+1+i) \end{bmatrix}. \qquad (3.4)$$

Найденные коэффициенты аппроксимации подставим в (3.1), а далее определим значение модуля корреляции ρ_k^M по выражению (2.22). Повторяя вычисления для каждого значения k из указанного диапазона, определим множество значений $\rho_1^M, \rho_2^M, \rho_3^M, \ldots, \rho_{T-M-1}^M$. Значения ρ_k^M для $k \in \{1,2,\ldots,1000\}$ приведены на рисунке 3.1.

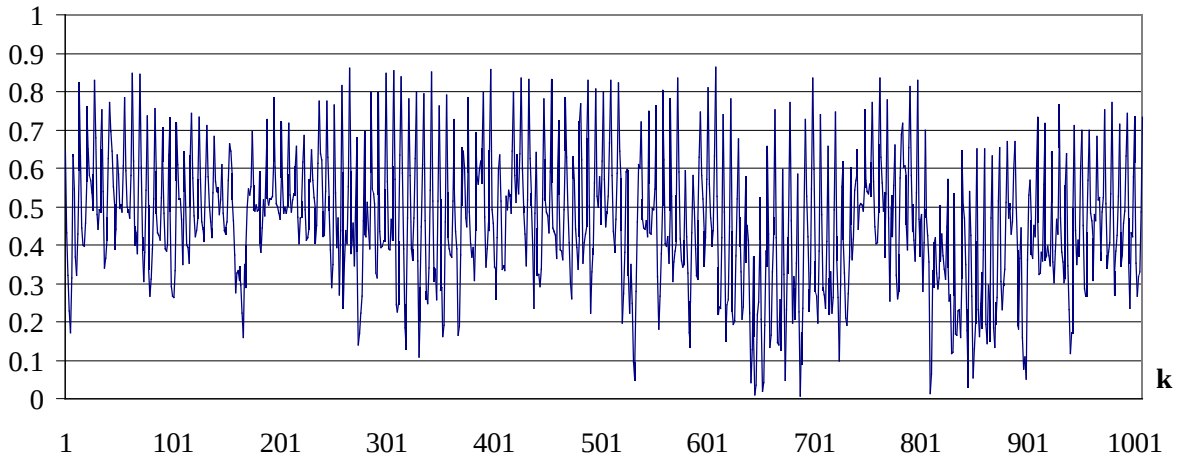


Рис. 3.1. Значения меры подобия ρ_k^M для $k \in \{1,2,\ldots,1000\}$

Далее на основании (2.23) определим значение максимума корреляции ρ_{kmax}^M и соответствующую задержку $kmax$. Для решаемой задачи значение максимума корреляции $\rho_{kmax}^M = 0.862$ и соответствует задержке $kmax = 9817$. Для задержки $kmax$ уравнение аппроксимации имеет вид

$$\hat{Z}_{24409}^{216} = 0.7759\, Z_{14592}^{216} + 172.7604\, I^{216}. \qquad (3.5)$$

Результаты аппроксимации представлены на рисунке 3.2.

Рис. 3.2. Пример аппроксимации

3) Определить выборку базовой истории.

Согласно гипотезе подобия (2.1.3.), в качестве выборки базовой истории Z_{τ}^{P} берем выборку следующую за выборкой максимального подобия Z_{14592}^{216}, то есть выборка базовой истории равна $Z_{\tau}^{P} = Z_{kmax*+M+1}^{P} = Z_{14809}^{24}$.

4) Вычислить прогнозные значения.

Вычислим значения выборки $\hat{Z}_{24625}^{24}$, согласно зависимости

$$\hat{Z}_{24625}^{24} = 0.7759\, Z_{14809}^{24} + 172.7604\, I^{24}. \qquad (3.6)$$

Результат экстраполяции представлен на рисунке 3.3.

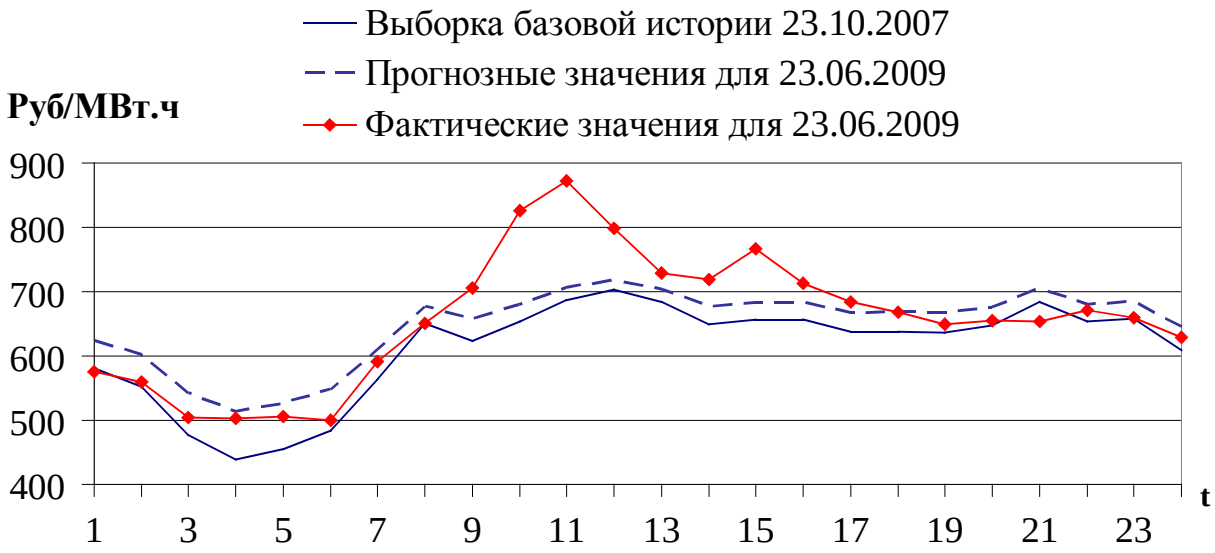


Рис. 3.3. Результат экстраполяции $\hat{\boldsymbol{Z}}_{24625}^{24}$, выборка базовой истории $\boldsymbol{Z}_{14808}^{24}$ и фактические значения цен на электроэнергию

Значение MAPE (2.10) аппроксимации (3.5) равно 5.39%, значение MAE (2.9) равно 35.69 руб/МВт·ч. Оценки точности для модели экстраполяции (3.6): MAPE = 6.21%; MAE = 43.16 руб/МВт·ч. Результаты показывают, что ошибка аппроксимации близка, но не равна ошибке экстраполяции. Зависимость ошибок экстраполяции и аппроксимации рассмотрена в разделе 3.4.

После описания шагов алгоритма экстраполяции необходимо провести оценку времени вычислений прогнозных значений при его программной реализации.

Оценка времени расчета прогнозных значений. Разработанный алгоритм экстраполяции реализован при помощи программного комплекса MATLAB [46]. Эксперименты проводились на персональном компьютере следующей модификации:

— процессор Intel Core 2 Duo E7400 2.80 ГГц, 2ГБ DDR2,

— материнская плата AsusP5KPL-CM.

Оценка производительности данного персонального компьютера при помощи теста Java Micro Benchmark составляет 828 единиц. Оценки производительности компьютеров и серверов при помощи данного теста колеблются в широком диапазоне от 95 до 22 054 единиц [47].

Время расчета прогнозных значений временного ряда t_p зависит от длины временного ряда T и производительности компьютера. Одной из особенностей модели экстраполяции является то, что P прогнозных значений определяются за один пробег алгоритма, например, время расчета одного значения вперед равно времени расчета 24 значений вперед. Экспериментальная зависимость времени расчета t_p от длины временного ряда T для указанного персонального компьютера представлена на рисунке 3.4, значения представлены в таблице 2.

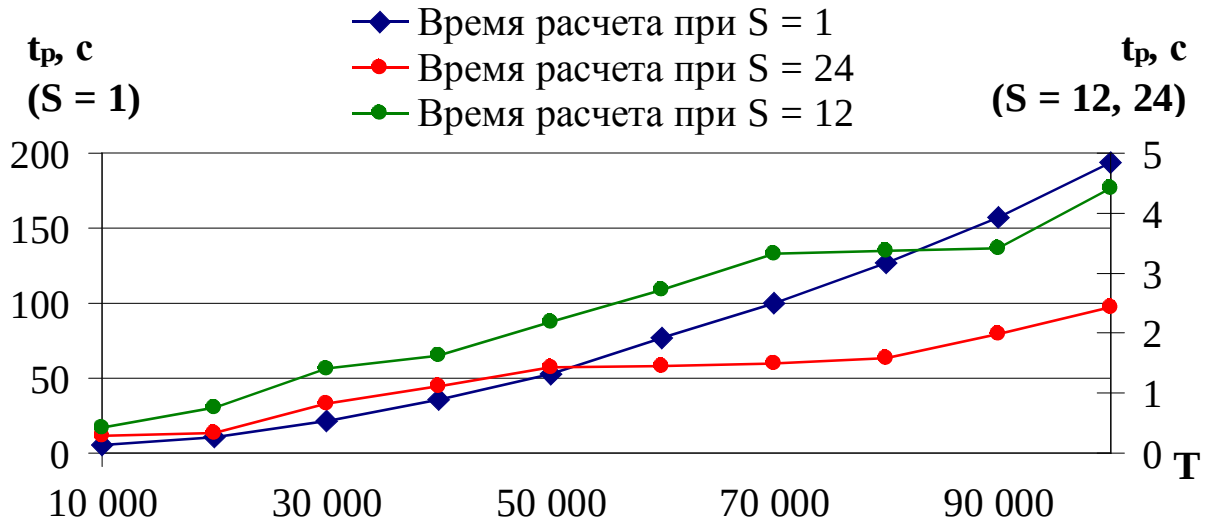


Рис. 3.4. Экспериментальная зависимость времени расчета t_p от длины временного ряда T

На рисунке 3.4 приведены три зависимости времени расчета на 24 значения вперед t_p от длины временного ряда T при условии, что множество значений модуля корреляции ρ_k^M определяется с шагом

1) $S=1$, т. е. вычисляются значения $\rho_1^M, \rho_2^M, \rho_3^M, \ldots, \rho_{T-M-1}^M$;

2) $S=12$, т. е. вычисляются значения $\rho_{12}^M, \rho_{24}^M, \rho_{36}^M, \ldots, \rho_{T-M-1}^M$;

3) $S=24$, т. е. вычисляются значения $\rho_{24}^M, \rho_{48}^M, \rho_{72}^M, \ldots, \rho_{T-M-1}^M$.

Таблица 2.

Значения времени расчета при различной длине временного ряда

Длина временного ряда T	**Время расчета** t_p**, с**		
	$S=1$	$S=12$	$S=24$
10 000	5.08	0.43	0.29
20 000	11.12	0.76	0.33
30 000	21.80	1.41	0.83
40 000	35.73	1.63	1.11
50 000	52.87	2.19	1.43
60 000	77.17	2.72	1.45
70 000	99.61	3.32	1.49
80 000	126.68	3.38	1.59
90 000	156.85	3.42	1.99
100 000	193.90	4.42	2.44

Шаг $S=24$ применяется для прогнозирования энергопотребления, цен энергорынка РФ (4.1.). Шаг перебора значений модуля корреляции ρ_k^M определяется экспертом и может варьироваться в зависимости от сезонности временного ряда.

Согласно оценке [48], высокой считается скорость, при которой вычисление 24 прогнозных значений занимает не более 20 минут. Время расчета 24 значений временного ряда длинной 100 000 значений и переборе значений ρ_k^M с шагом $S=1$ составляет около 200 секунд на указанном персональном компьютере. При шаге $S=12$ аналогичное время расчетов не превышает 5 секунд; при $S=24$ — 2.5 секунд (таблица 2). Данные оценки

подтверждают высокую скорость вычислений предложенного алгоритма экстраполяции временного ряда без учета внешних факторов.

3.2. Алгоритм экстраполяции временного ряда с учетом внешних факторов

Модель экстраполяции временного ряда с учетом внешних факторов сформулирована в разделе 2.2. диссертации. Алгоритм экстраполяции состоит из следующих шагов.

1) Определить выборку новой истории.
2) Определить выборку максимального подобия.
3) Определить выборку базовой истории.
4) Вычислить прогнозные значения.

Далее приведем описание каждого указанного шага, иллюстрируя расчеты решением следующей задачи. Пусть даны значения временного ряда цен на электроэнергию европейской территории РФ (№1 в таблице 7) с 01.09.2006 до 22.06.2009; а также значения временного ряда энергопотребление европейской территории РФ (№1 в таблице 11) с 01.09.2006 до 22.06.2009. Обозначим временной ряд цен на электроэнергию $Z(t)$, временной ряд энергопотребления – $X(t)$. Требуется определить 24 значения временного ряда $Z(t)$ за 23.06.2009 с учетом влияния временного ряда энергопотребления $X(t)$. Считаем параметр модели $M = 216$ заданным.

В связи с тем, что значения временного ряда $X(t)$ доступны до той же отметки времени, что и значения временного ряда $Z(t)$, необходимо на первом этапе определить значения временного ряда $\hat{X}(t)$ в сутках 23.06.2009 по алгоритму, рассмотренному подробно в разделе 3.1. На втором

этапе полученные экстраполированные значения $\hat{X}(t)$ использовать при вычислении экстраполированных значений $\hat{Z}(t)$.

Алгоритм расчета $\hat{X}(t)$ рассмотрен в предыдущем разделе. Модель экстраполяции временного ряда $X(t)$ имеет вид

$$\hat{\boldsymbol{X}}_{\mathbf{24625}}^{\mathbf{24}} = 1.0789\, \boldsymbol{X}_{\mathbf{24073}}^{\mathbf{24}} - 4689.78\, \boldsymbol{I}^{\mathbf{24}} \qquad (3.7)$$

и оценку точности: MAPE = 1.07%; MAE = 782 МВт·ч. Считаем задачу определения $\hat{X}(t)$ решенной.

1) Определить выборку новой истории.

Выборкой новой истории временного ряда $Z(t)$ является выборка временного ряда, значения которой предшествуют моменту прогноза T. В текущей постановке задачи выборка новой истории равна $\boldsymbol{Z}_{\boldsymbol{T-M+1}}^{\boldsymbol{M}} = \boldsymbol{Z}_{\mathbf{24409}}^{\mathbf{216}}$. Выборка новой истории соответствует значениям цены на электроэнергию за период с 14.06.2009 до 22.06.2009.

2) Определить выборку максимального подобия.

Для определения выборки максимального подобия необходимо определить значения ошибки регрессии S_k^M (2.29) для выборки $\boldsymbol{Z}_{\boldsymbol{T-M+1}}^{\boldsymbol{M}}$ и всех выборок с задержкой $k \in \{1,2,\ldots,T-M-1\}$. При этом для каждого значения k из указанного диапазона требуется решить задачу аппроксимации выборки $\boldsymbol{Z}_{\boldsymbol{T-M+1}}^{\boldsymbol{M}}$ при помощи выборок $\boldsymbol{Z}_{\boldsymbol{T-M+1-k}}^{\boldsymbol{M}}$ и $\boldsymbol{X}_{\boldsymbol{T-M+1}}^{\boldsymbol{M}}$. Обозначим момент времени $T-M+1=t$ и решим данную задачу.

Вычислим аппроксимированные значения выборки

$$\hat{\boldsymbol{Z}}_{\boldsymbol{T-M+1}}^{\boldsymbol{M}} = \alpha_2 \boldsymbol{Z}_{\boldsymbol{T-M+1-k}}^{\boldsymbol{M}} + \alpha_1 \boldsymbol{X}_{\boldsymbol{T-M+1}}^{\boldsymbol{M}} + \alpha_0 \boldsymbol{I}^{\boldsymbol{M}}, \qquad (3.8)$$

а с учетом обозначения

$$\hat{\boldsymbol{Z}}_{\boldsymbol{t}}^{\boldsymbol{M}} = \alpha_2 \boldsymbol{Z}_{\boldsymbol{t-k}}^{\boldsymbol{M}} + \alpha_1 \boldsymbol{X}_{\boldsymbol{t}}^{\boldsymbol{M}} + \alpha_0 \boldsymbol{I}^{\boldsymbol{M}}. \qquad (3.9)$$

Согласно методу наименьших квадратов, коэффициенты аппроксимации определим исходя из уравнения

$$\boldsymbol{Z}_X \cdot \boldsymbol{A} = \boldsymbol{Z}_Y, \tag{3.10}$$

где матрицы $\boldsymbol{Z}_X$ и $\boldsymbol{Z}_Y$ определяются следующим образом:

$$\boldsymbol{Z}_X = \begin{bmatrix} \sum_{i=0}^{M-1} Z^2(t-k+i) & \sum_{i=0}^{M-1} Z(t-k+i)\cdot X(t+i) & \sum_{i=0}^{M-1} Z(t-k+i) \\ \sum_{i=0}^{M-1} Z(t-k+i)\cdot X(t+i) & \sum_{i=0}^{M-1} X^2(t+i) & \sum_{i=0}^{M-1} X(t+i) \\ \sum_{i=0}^{M-1} Z(t-k+i) & \sum_{i=0}^{M-1} X(t+i) & M \end{bmatrix};$$

$$\boldsymbol{Z}_Y = \begin{bmatrix} \sum_{i=0}^{M-1} Z(t-k+i)\cdot Z(t+i) \\ \sum_{i=0}^{M-1} Z(t+i)\cdot X(t+i) \\ \sum_{i=0}^{M-1} Z(t+i) \end{bmatrix}. \tag{3.11}$$

Например, для задержки $k = 11\,689$ коэффициенты аппроксимации равны

$$\boldsymbol{A} = \begin{bmatrix} \alpha_2 \\ \alpha_1 \\ \alpha_0 \end{bmatrix} = \boldsymbol{Z}_X^{-1} \cdot \boldsymbol{Z}_Y = \begin{bmatrix} -0.1544 \\ 0.0178 \\ -476.7678 \end{bmatrix}. \tag{3.12}$$

Найденные коэффициенты аппроксимации подставим в выражение (3.8), а далее определим значение ошибки регрессии S_k^M (2.29). Повторяя вычисления для каждого значения k из указанного диапазона, определим множество значений $S_1^M, S_2^M, S_3^M, \ldots, S_{T-M-1}^M$.

Далее на основании (2.21) определим значение минимума ошибки регрессии S_{kmax}^M и соответствующую задержку $kmax$. Для решаемой задачи минимальная ошибка $S_{kmax}^M = 137\,337.65$ соответствует задержке $kmax = 11\,689$. Для задержки $kmax$ уравнение аппроксимации имеет вид

$$\hat{\boldsymbol{Z}}_t^{216} = -0.1544\,\boldsymbol{Z}_{t-11689}^{216} + 0.0178\,\boldsymbol{X}_t^{216} - 476.7678\,\boldsymbol{I}^{216}. \tag{3.13}$$

Выборка $Z_{t-11689}^{216} = Z_{12720}^{216}$ соответствует значениям временного ряда за период с 05.06.2008 по 13.06.2008. Результаты аппроксимации представлены на рисунке 3.5.

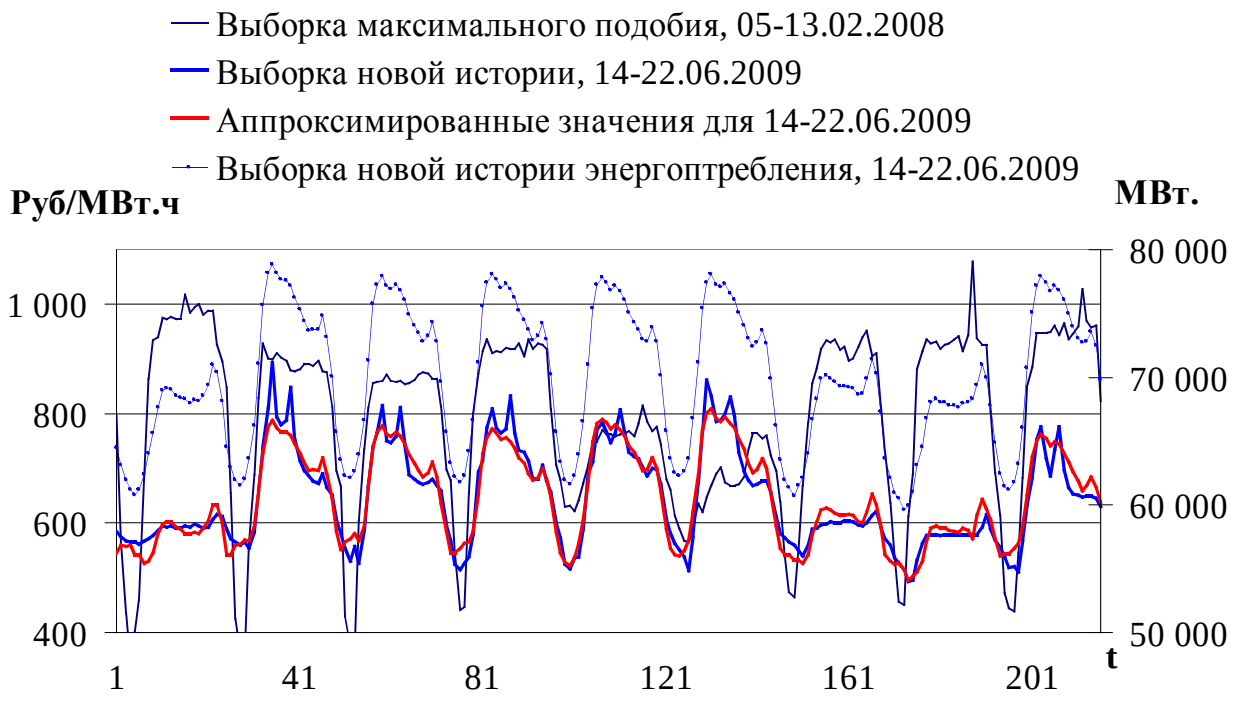


Рис. 3.5. Пример аппроксимации с учетом внешнего фактора

3) Определить выборку базовой истории.

Согласно гипотезе подобия (2.2.3.), в качестве выборки базовой истории Z_{τ}^{P} берем выборку, следующую за выборкой максимального подобия Z_{12720}^{216}, то есть выборка базовой истории равна $Z_{\tau}^{P} = Z_{12936}^{24}$.

4) Вычислить прогнозные значения.

Вычислим значения выборки $\hat{Z}_{24625}^{24}$, согласно зависимости

$$\hat{Z}_{24625}^{24} = -0.1544\, Z_{12936}^{24} + 0.02\, \hat{X}_{24625}^{24} - 476.77\, I^{24}. \qquad (3.14)$$

Результат экстраполяции представлен на рисунке 3.6.

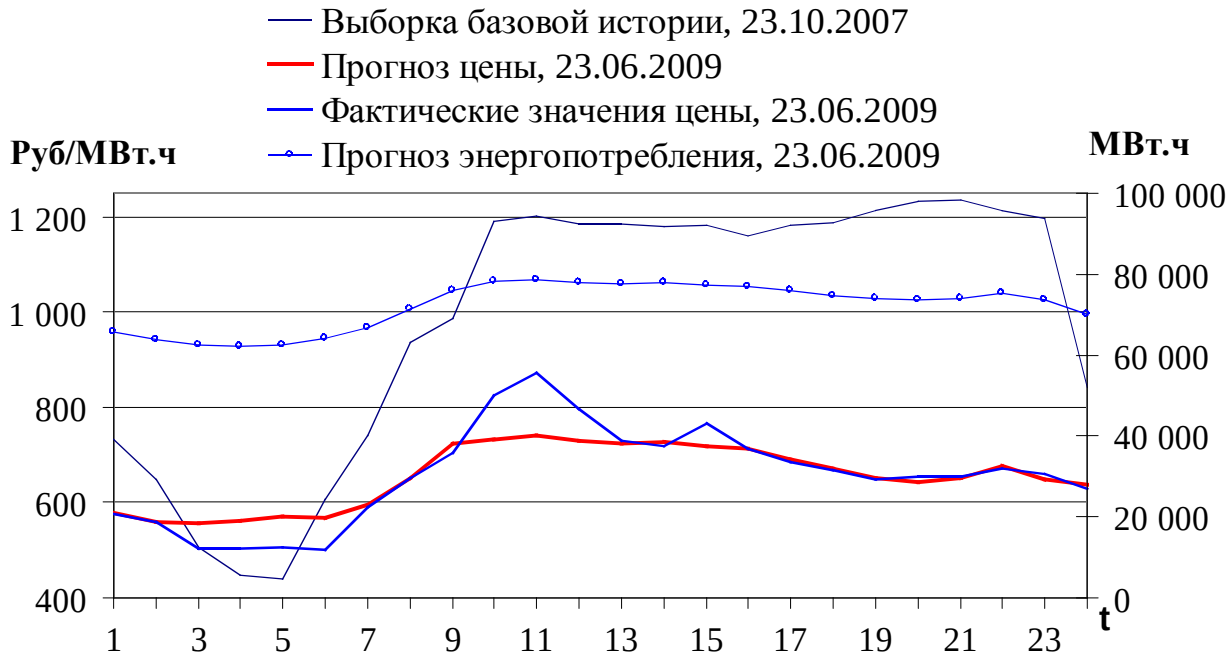


Рис. 3.6. Результат экстраполяции цен на электроэнергию с учетом энергопотребления

Оценки ошибки аппроксимации: значение MAPE аппроксимации (3.8) равно 2.11%, значение MAE = 12.24 руб/МВт·ч. Оценки точности для модели экстраполяции (3.14): MAPE = 4.42%; MAE = 28.29 руб/МВт·ч. Результаты показывают, что ошибка аппроксимации близка, но не равна ошибке экстраполяции. Зависимость ошибок экстраполяции и аппроксимации рассмотрена в разделе 3.4.

На основании сравнения результатов расчета прогнозных значений цен на электроэнергию в предыдущем и текущих разделах диссертации, т. е. без учета и с учетом одного внешнего фактора, утверждаем, что учет внешнего фактора может приводить к повышению точности прогнозирования, как это получилось в рассмотренной задаче. Подробнее данная особенность модели рассмотрена в разделе 4.1.2.

После описания алгоритма экстраполяции необходимо произвести оценку времени вычисления прогнозных значения при программной

реализации данного алгоритма.

Оценка времени расчета прогнозных значений. Разработанный алгоритм экстраполяции временного ряда с учетом внешних факторов реализован при помощи программного комплекса MATLAB. Эксперименты проводились на указанном в предыдущем разделе персональном компьютере.

Время расчета t_p будущий значений временного ряда зависит от длины временного ряда T, количества внешних факторов и производительности компьютера. Для модели с учетом внешних факторов время t_p не зависит от времени упреждения P.

Зависимость времени расчета t_p от длины временного ряда T при экстраполяции с учетом одного внешнего фактора с использованием указанного персонального компьютера представлена на рисунке 3.7, значения представлены в таблице 3.

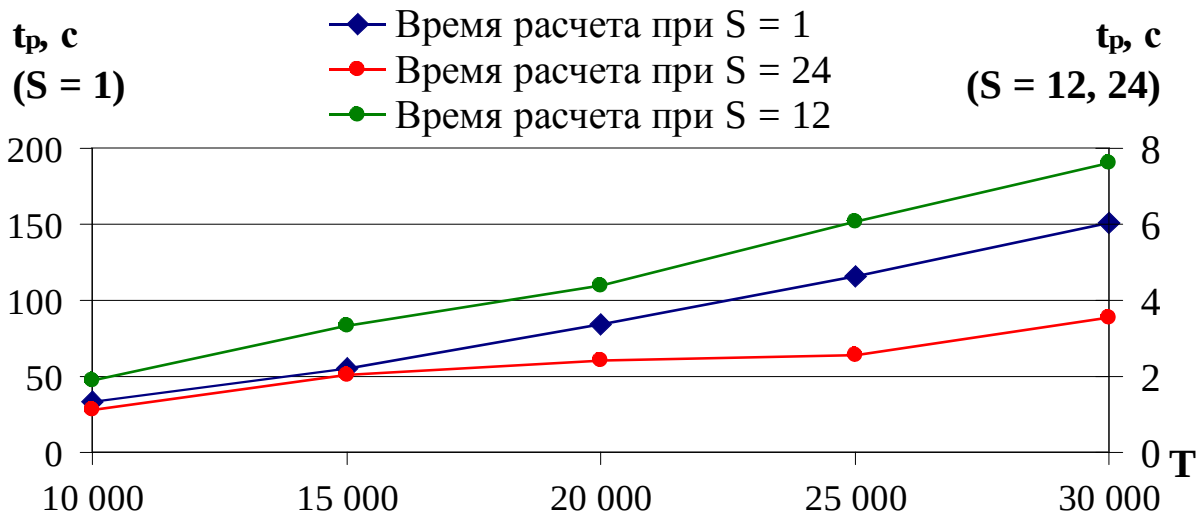


Рис. 3.7. Зависимость времени расчета t_p от длины временного ряда T при экстраполяции с учетом одного внешнего фактора

Таблица 3.

Значения времени расчета при различной длине временного ряда для экстраполяции с учетом одного внешнего фактора

Длина временного ряда, T	**Время расчета** t_p**, с**		
	$S=1$	$S=12$	$S=24$
10 000	33.28	1.91	1.14
15 000	55.29	3.32	2.04
20 000	84.41	4.40	2.43
25 000	115.40	6.06	2.57
30 000	150.95	7.62	3.53

На рисунке 3.7 приведены три зависимости времени расчета 24 значений вперед t_p на модели с учетом одного внешнего фактора от длины временного ряда T при условии, что множество значений S_k^M определяется с шагом $S=1$, $S=12$, $S=24$ (раздел 3.1.). Полученные оценки значений t_p находятся в диапазоне от 33 до 150 секунд для $S=1$; для $S=12$, $S=24$ величина t_p не превышает 8 секунд при длине временного ряда 30 000 значений (таблица 3). На основании полученных оценок t_p сделаем вывод о высокой скорости вычислений экстраполированных значений временного ряда с учетом одного внешнего фактора [48].

Рассмотрим зависимость расчета t_p от длины временного ряда T при экстраполяции с учетом двух внешних факторов при вычислениях на указанном персональном компьютере (рис. 3.8).

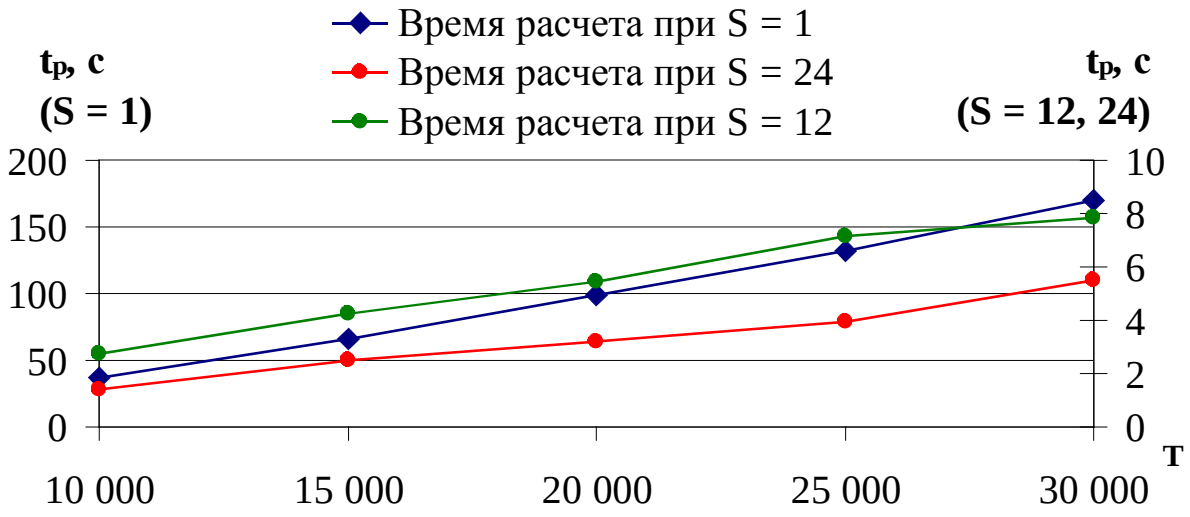


Рис. 3.8. Зависимость времени расчета t_p от длины временного ряда T при экстраполяции с учетом двух внешних факторов

Таблица 4.

Значения времени расчета при различной длине временного ряда для экстраполяции с учетом двух внешних факторов

Длина временного ряда T	**Время расчета** t_p**, с**		
	$S=1$	$S=12$	$S=24$
10 000	37.44	2.77	1.39
15 000	65.86	4.26	2.52
20 000	98.51	5.47	3.21
25 000	131.97	7.13	3.95
30 000	170.39	7.87	5.48

На рисунке 3.8 приведены три зависимости времени расчета 24 значений вперед t_p на модели экстраполяции с учетом двух внешних факторов от длины временного ряда T при условии, что множество значений S_k^M определяется с шагом $S=1$, $S=12$, $S=24$ (раздел 3.1.). Полученные оценки значений t_p находятся в диапазоне от 37 до 170 секунд для $S=1$; для

$S=12$, $S=24$ величина t_p не превышает 8 секунд.

Отметим, что при увеличении числа внешних факторов с одного до двух время расчета t_p для аналогичных значений T существенно не изменилось. Например, при $T=30\,000$ и $S=1$ при одном внешнем факторе значение $t_p=151$ секунда; при двух внешних факторах аналогичное значение — $t_p=170$ секунд. Таким образом, разница составила менее 20 секунд.

На основании полученных оценок значений t_p заключим высокую скорость вычисления будущих значений предложенного алгоритма экстраполяции временного ряда с учетом внешних факторов [48].

3.3. Алгоритм идентификации модели

3.3.1. Описание алгоритма

В настоящем разделе предложен алгоритм идентификации моделей EMMSP (2.25) и EMMSPX (2.34). Задача идентификации состоит в нахождении параметра модели M, который определяет длину выборок новой истории и максимального подобия (рис. 2.6, 2.7). Выполним идентификацию обеих моделей по одному алгоритму.

1) Определить тестовый и контрольный периоды временного ряда.

2) Определить время упреждения P и диапазон возможных значений параметра M.

3) Прогнозировать тестовый период на P значений вперед при всех значениях параметра M из установленного диапазона.

4) Построить зависимость ошибки прогнозирования от M.

5) Экспертно определить окончательное значение параметра M.

Рассмотрим указанные шаги идентификации подробнее.

1) Определить тестовый и контрольный периоды временного ряда.

На данном шаге исходный временной ряд $Z(t)$ разделим на три части в пропорции, например, 1:1:1. Полученные части назовем базовый период (33%), тестовый период (33%) и контрольный период (34%) временного ряда, соответственно, как показано на рисунке 3.9.

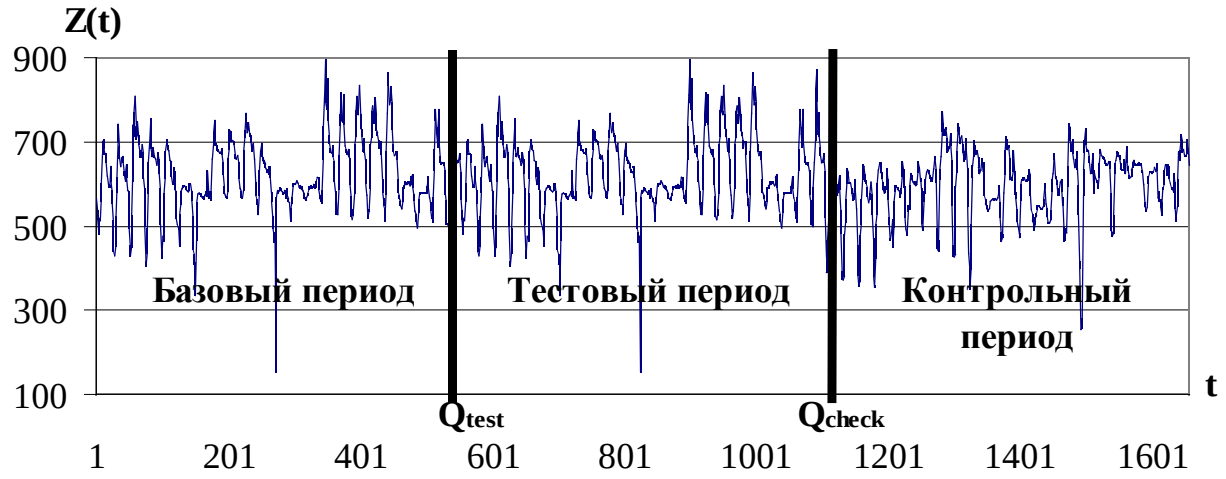


Рис. 3.9. Разделение временного ряда на периоды

2) Определить время упреждения P и диапазон возможных значений параметра M.

Далее, исходя из решаемой задачи прогнозирования временного ряда, требуется определить время упреждения P. а также диапазон возможных значений параметра M. Первоначально рекомендуем брать широкий диапазон возможных значений M, например, $M \in \{2P, \ldots, 15P\}$. Значения M можно перебирать с шагом $S=1$, т. е. $M \in \{P, P+1, P+2, \ldots, 15P\}$. Однако практика решения данной задачи показала, что удобнее перебирать значения параметра M с шагом $S=P$ или $S=0.5P$, т. е. $M \in \{P, 1.5P, 2P, \ldots, 15P\}$. Данный подход сокращает время

идентификации, существенно не влияя на точность последующего прогнозирования.

3) Прогнозировать тестовый период на P значений вперед при всех значениях параметра M из установленного диапазона.

Для каждого значения параметра M из установленного диапазона, прогнозируем значения временного ряда на P значений вперед внутри тестового периода. Рекомендуем устанавливать длину тестового периода в диапазоне от $150P$ до $300P$.

По результатам прогнозирования для каждого значения M определяем среднюю абсолютную ошибку прогноза для всего тестового периода

$$MAE(M)=\frac{1}{K}\sum_{t=Q_{test}}^{Q_{test}+K}\left|\hat{Z}(t)-Z(t)\right|, \qquad (3.15)$$

где K – количество значений временного ряда, попавших внутрь тестового периода, Q_{test} – начало тестового периода на оси времени (рис. 3.9), $\hat{Z}(t)$ – прогнозные значения, полученные при прогнозировании с параметром модели M.

4) Построить зависимость ошибки прогнозирования от M.

Строим график зависимости $MAE(M)$ для тестового периода и определяем диапазон значений M, соответствующий устойчивому минимуму $MAE(M)$.

Рассмотрим в качестве примера зависимость $MAE(M)$ для временного ряда энергопотребления европейской территории РФ, представленного на рисунке 3.10. Первоначально был выбран диапазон $M\in\{36,48,\ldots,360\}$, внутри которого выделен диапазон устойчивого минимума ошибки $M\in\{168,180,\ldots,240\}$.

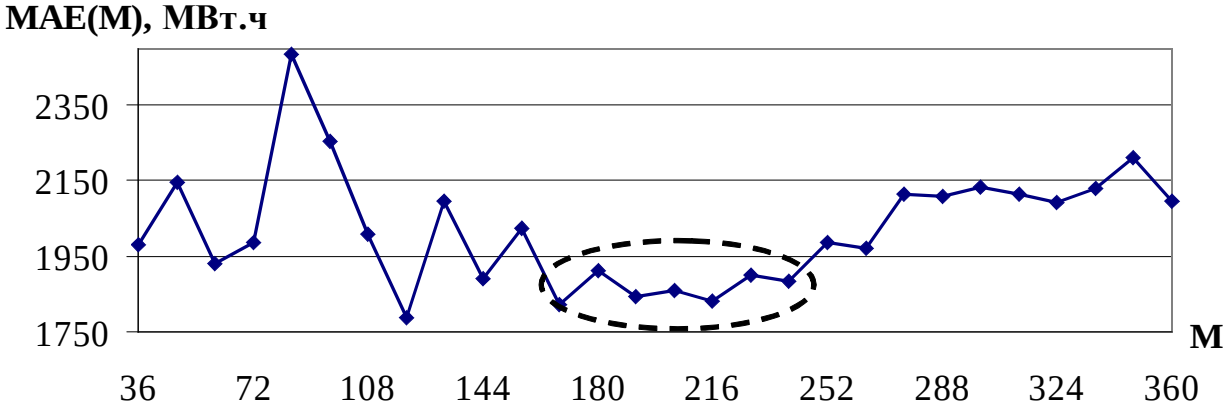


Рис. 3.10. Зависимость $MAE(M)$ для $M \in \{36,48,...,360\}$

На рисунке 3.10 представлен типичный вид зависимости $MAE(M)$, при малых и больших значениях M из диапазона $\{1.5P, 2P, ..., 15P\}$ ошибка, как правило, велика. Однако существует стабильный минимум значений $MAE(M)$ для некоторых промежуточных значений M. В приведенном примере значения параметра M перебирались с шагом $S = 0.5P$.

5) Экспертно определить значение параметра M.

На последнем шаге экспертом выбирается значение параметра модели M внутри диапазона устойчивого минимума.

Рекомендуем выбирать значение M таким образом, чтобы значения $MAE(M)$ для соседних точек примерно совпадали со значением $MAE(M)$ для выбранного значения M. В случае, если устойчивый минимум значений $MAE(M)$ лежит близко к левой границе исследуемого диапазона M, то рекомендуем брать наибольшее значение из диапазона устойчивого минимума.

После выполнения всех шагов алгоритма задача идентификации решена и модель может применяться для прогнозирования контрольного периода.

3.3.2. Распараллеливание вычислений

Наиболее ресурсоемкий шаг в смысле времени и вычислительных мощностей в предложенном алгоритме идентификации модели состоит в прогнозировании тестового периода при различных значениях параметра модели M из установленного диапазона (шаг 3). В связи с тем, что результаты прогнозирования при одном значении параметра M не зависят от результатов прогнозирования при другом значении M, данный процесс может быть распараллелен следующим образом.

1) Разделить тестовый период временного ряда на непересекающиеся подпериоды, объединение которых дает тестовый период.

2) Производить параллельно вычисления прогнозных значений для подпериодов при различных значениях параметра M.

В таблице 5 проиллюстрирована схема распараллеливания вычислений.

Таблица 5.

Схема распараллеливания вычисления при идентификации

Значение параметра M	**Тестовый подпериод №1**	**Тестовый подпериод №2**	**...**	**Значение** $MAE(M)$
M_1	$MAE_{№1}(M_1)$	$MAE_{№2}(M_1)$	...	$MAE(M_1)$
M_2	$MAE_{№1}(M_2)$	$MAE_{№2}(M_2)$	...	$MAE(M_2)$
M_3	$MAE_{№1}(M_3)$	$MAE_{№2}(M_3)$	...	$MAE(M_3)$
...	...	...	...	...

Для значения параметра модели M_1 производится прогноз внутри тестового подпериода №1 и определяется величина ошибки $MAE_{№1}(M_1)$. Параллельно с данными вычислениями производятся оценки остальных величин $MAE_{№q}(M_i)$. После завершения всех вычислений итоговые

значения $MAE(M_1)$ определяются как средние величины от множества значений оценок ошибки $MAE_{№1}(M_1)$, $MAE_{№2}(M_1)$,..., соответствующих рассматриваемому значению параметра M_1.

В случае ужесточения требований по скорости идентификации модели количество тестовых подпериодов должно быть увеличено.

3.3.3. Наборы моделей

Решением задачи идентификации модели экстраполяции является значение параметра M. В случае если сформированная модель $EMMSP(M)$ или $EMMSPX(M)$ имеет недостаточную точность прогнозирования, возможно повышение точности за счет использования различных моделей для различных отрезков временного ряда.

Пусть необходимо прогнозировать временной ряд $Z(t)$. При этом искомый временной ряд $Z(t)$ можно очевидным образом разбить на некоторые чередующие отрезки, например, временной ряд, имеющий почасовое разрешение, может быть разбит по дням недели, месяцам и т. д. Тогда в процессе идентификации модели необходимо определять параметр модели M для каждого установленного отрезка отдельно. Иллюстрацией такого набора моделей для прогнозирования временного ряда, имеющего почасовое разрешение, является таблица 6.

Таблица 6.

Набор моделей для прогнозирования временного ряда, имеющего почасовое разрешение

День недели	**Параметр модели**
Понедельник	M_1
Вторник	M_2
Среда	M_3
Четверг	M_4
Пятница	M_5
Суббота	M_6
Воскресенье	M_7

В данном примере каждому дню недели соответствует собственное значение параметра M: при прогнозе значений понедельника используется модель $EMMSP(M_1)$, при прогнозе вторника — $EMMSP(M_2)$ и т. д.

Получить набор моделей для временного ряда можно на основании результатов идентификации модели следующим образом.

1) Разбить результаты прогнозирования тестового периода при различных значениях параметра M в соответствии с установленными заранее отрезками временного ряда (например, по дням недели).

2) Определить зависимость $MAE(M)$ для каждого отрезка.

3) Определить диапазон устойчивого минимума и окончательное значение параметра M для каждого отрезка отдельно.

Наборы моделей применялись для прогнозирования временных рядов энергорынка РФ. Результаты прогнозирования цен на электроэнергию и энергопотребление, представленные в разделах 4.1.2. и 4.1.3. диссертации, подтверждают, что применение наборов повышает точность прогнозирования

(таблицы 10 и 14).

3.3.4. Оценка времени идентификации

Идентификация модели, предложенная в настоящем разделе диссертации, имеет следующую оценку времени вычислений на шаге №3 (остальные шаги не требуют оценки времени)

$$t_{id} = \frac{T_{TEST}}{P} \cdot t_p \cdot N_M \cdot \frac{828}{PC}. \qquad (3.16)$$

Здесь t_{id} – время идентификации в секундах, T_{TEST} – длина тестового периода временного ряда, P – время упреждения, N_M – количество возможных значений параметра M, PC – оценка производительности компьютера по тесту Java Micro Benchmark [47]. Время расчета t_p определяется по таблицам 2, 3 и 4. В случае если длина временного ряда превышает диапазоны, заданные в таблицах, то необходимо линейно экстраполировать значения t_p при соответствующем шаге перебора значений задержки k при определении множества значений модуля корреляции.

Для иллюстрации приведенной оценки рассмотрим пример идентификации временного ряда энергопотребления европейской территории РФ (№1 в таблице 11). Временной ряд содержит значения за период с 01.09.2006 по 07.08.2011, длина временного ряда $T = 43\,224$ значения. В качестве тестового периода выбираем с 07.08.2009 по 06.08.2010, в качестве контрольного периода выбираем период с 07.08.2010 по 07.08.2011. При этом длина тестового и контрольного периодов равна 8 760 значений.

Для рассматриваемого примера $T_{TEST} = 8\,760$; в связи с тем, что идентификации выполнялась на указанном персональном компьютере, оценка его производительности $PC = 828$. Время упреждения $P = 24$, количество значений параметра M в оцениваемом диапазоне $N_M = 28$ (рис.

3.10). Начало тестового периода определяется $Q_{test} = 43\,224 - 8760 \cdot 2 = 25\,704$, таким образом прогнозирование временного ряда производится в диапазоне $\{25\,704, \ldots, 34\,464\}$. Перебор значений задержки k производится с шагом $S = 12$. Обратившись к таблице 2, находим оценку значения $t_p = 1.63$ и получаем

$$t_{id} = \frac{8760}{24} \cdot 1.63 \cdot 28 \cdot \frac{828}{828} = 16\,659\,c = 4.63\,ч\,. \qquad (3.17)$$

Таким образом, для идентификации временного ряда энергопотребления европейской территории РФ необходимо 4.63 часа работы указанного в разделе 3.1. типа персонального компьютера.

В случае если произведенная оценка времени идентификации слишком высока, то существенное сокращение времени может быть достигнуто применением параллельных вычислений (раздел 3.3.2.).

3.4. Алгоритм построения доверительного интервала

В настоящем разделе рассмотрен алгоритм построения доверительного интервала прогнозных значений для разработанной модели экстраполяции.

Выборки максимального подобия $\boldsymbol{Z}_{t-k}^{M}$ и новой истории $\boldsymbol{Z}_{t}^{M}$ связаны соотношением

$$\boldsymbol{Z}_{t}^{M} = \alpha_1 \boldsymbol{Z}_{t-k}^{M} + \alpha_0 \boldsymbol{I}^{M} + \boldsymbol{E}_{app}^{M}\,, \qquad (3.18)$$

а выборки прогноза $\boldsymbol{Z}_{T+1}^{P}$ и базовой истории $\boldsymbol{Z}_{kmax*+M}^{P}$ – соотношением

$$\boldsymbol{Z}_{T+1}^{P} = \alpha_1 \boldsymbol{Z}_{kmax*+M}^{P} + \alpha_0 \boldsymbol{I}^{M} + \boldsymbol{E}_{ext}^{P}\,. \qquad (3.19)$$

Исследования векторов ошибок аппроксимации $\boldsymbol{E}_{app}$ и экстраполяции $\boldsymbol{E}_{ext}$ показали, что

— распределения значений ошибок $\boldsymbol{E}_{app}$ и $\boldsymbol{E}_{ext}$ не являются

нормальными (проверялось по критерию Пирсона при уровне значимости pValue = 0.05),

— распределение значений ошибок аппроксимации E_{app} не согласуется с распределением значений ошибок экстраполяции E_{ext} (проверялось по критерию Колмогорова-Смирнова при уровне значимости pValue = 0.05),

— среднее значение обеих ошибок близко к нулю,

— дисперсия ошибки аппроксимации E_{app} отлична от дисперсии ошибки экстраполяции E_{ext},

— асимметрия обеих ошибок отлична от нуля, колеблется в диапазоне от 0.5 до 1.8 для исследуемых временных рядов,

— эксцесс обеих ошибок больше единицы, колеблется в диапазоне от 3 до 15 для исследуемых временных рядов.

На рисунке 3.11 представлены гистограммы выборок E_{app} и E_{ext} для временного ряда цен на электроэнергию европейской территории РФ.

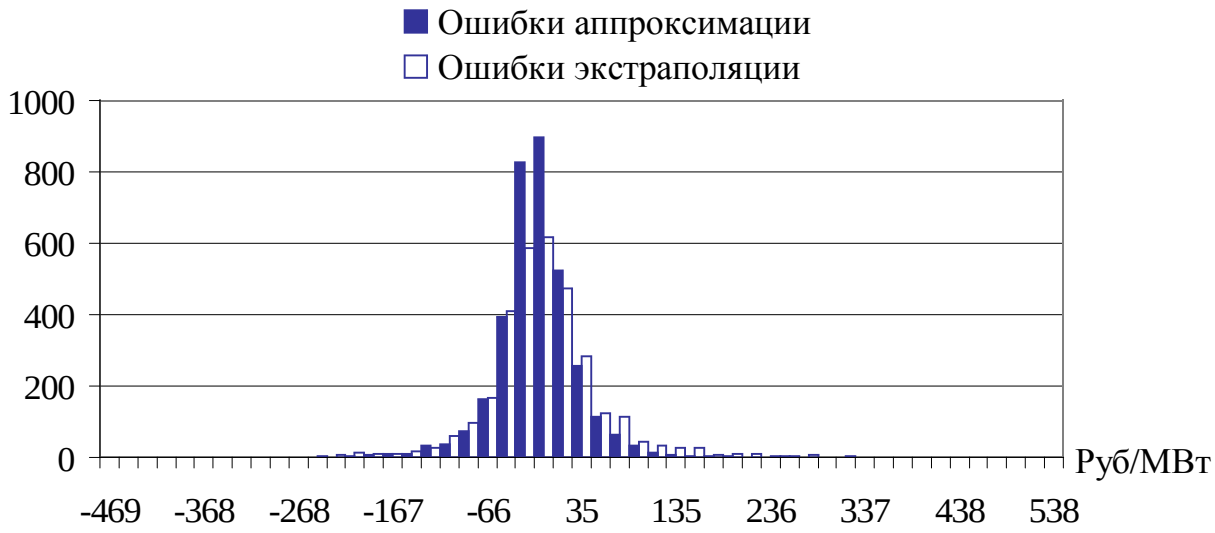


Рис. 3.11. Гистограммы выборок E_{app} и E_{ext}

Постановка задачи. В процессе прогнозирования значений временного

ряда $Z(t)$ вычисляем вектор ошибок аппроксимации $\boldsymbol{E}_{app}^{M}$ (2.1.2. и 2.2.2.). Требуется определить асимметричный доверительный интервал для прогнозных значений $\hat{Z}(T+1),\ldots,\hat{Z}(T+P)$ для заданной вероятности p, используя значения ошибок аппроксимации $\boldsymbol{E}_{app}^{M}$

$$Z_T^P=[\hat{Z}_T^P-\sigma_L^{ext};\hat{Z}_T^P+\sigma_R^{ext}]. \qquad (3.20)$$

Здесь значения σ_L^{ext} и σ_R^{ext} соответствуют левому (индекс L) и правому (индекс R) доверительному интервалу. Обозначим значение вероятности малой буквой p, чтобы не путать с временем упреждения P.

Решение задачи. Построим гистограмму значений $\boldsymbol{E}_{app}$. Пример гистограммы приведен на рисунке 3.12.

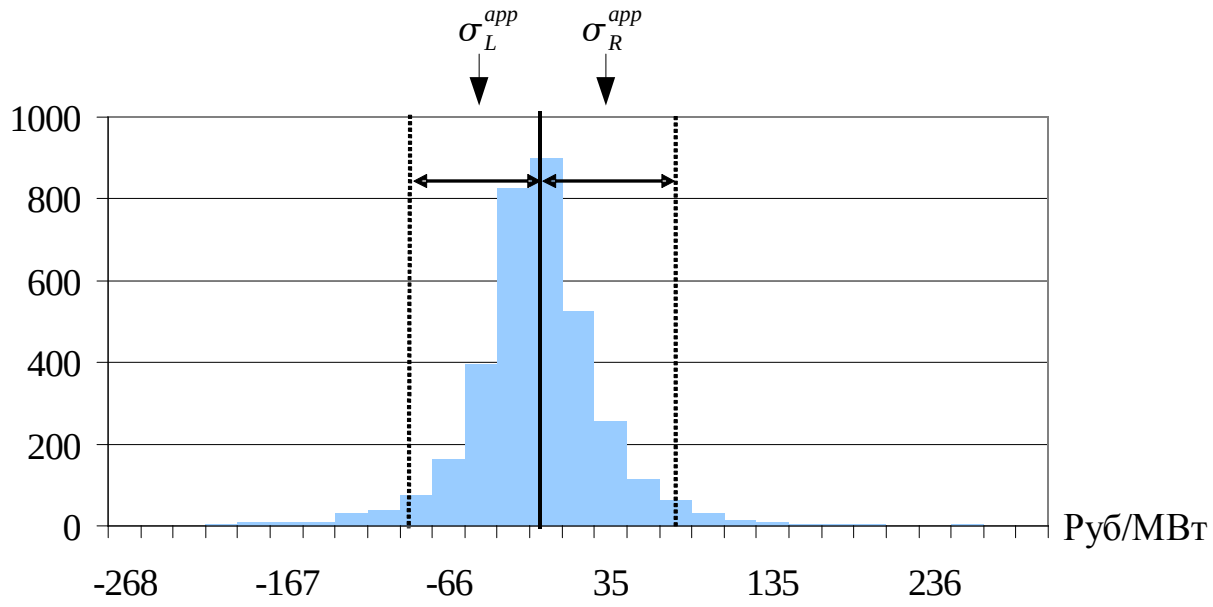


Рис. 3.12. Гистограмма выборки $\boldsymbol{E}_{app}$

Полную площадь гистограммы можно определить как сумму левой и правой половин относительно среднего значения ошибки

$$S_{TOTAL}=S_L+S_R. \qquad (3.21)$$

Для заданной вероятности p площадь гистограммы ограниченна σ_L^{app}

и σ_R^{app}, а потому

$$S(p) = p \cdot S_{TOTAL} = p \cdot (S_L + S_R) = p \cdot S_L + p \cdot S_R . \quad (3.22)$$

Определим значение левой границы доверительного интервала σ_L^{app} таким образом, чтобы площадь, ограниченная σ_L^{app} и средним значением, была равна $p \cdot S_L$. Аналогичным образом определим правую границу доверительного интервала σ_R^{app}. Используя тот же подход, определим значения границ доверительных интервалов σ_L^{ext}, σ_R^{ext} для вектора ошибок экстраполяции $\boldsymbol{E}_{\boldsymbol{ext}}$.

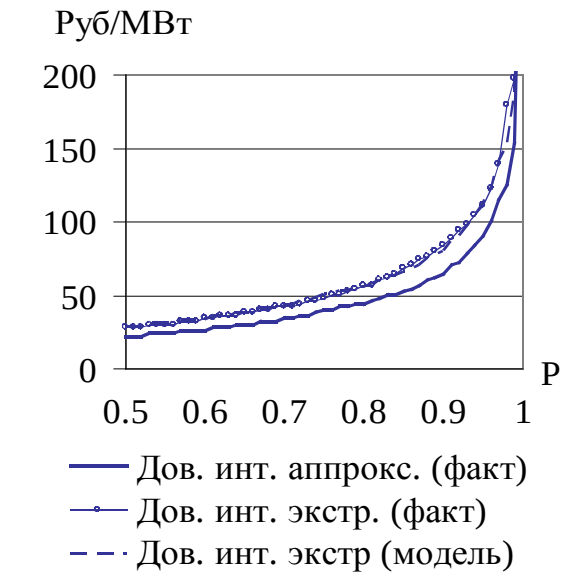
Исследование границ доверительных интервалов ошибок аппроксимации σ_L^{app}, σ_R^{app} и экстраполяции σ_L^{ext}, σ_R^{ext} показали, что значения $\hat{\sigma}_L^{ext}$, $\hat{\sigma}_R^{ext}$ могут быть с высокой точностью определены при помощи линейных зависимостей

$$\begin{aligned} \hat{\sigma}_L^{ext} &= \gamma_1^L \sigma_L^{app} + \gamma_0^L , \\ \hat{\sigma}_R^{ext} &= \gamma_1^R \sigma_R^{app} + \gamma_0^R , \end{aligned} \quad (3.23)$$

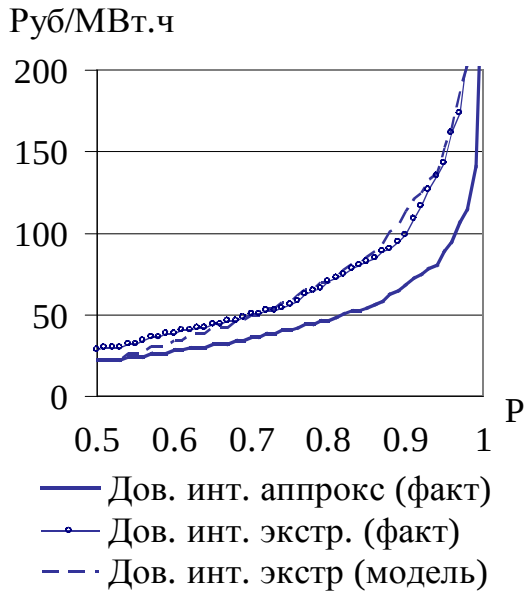
где γ_1^L, γ_0^L и γ_1^R, γ_0^R – коэффициенты.

При прогнозировании тестового периода определим фактические значения ошибок аппроксимации $\boldsymbol{E}_{\boldsymbol{app}}$ и экстраполяции $\boldsymbol{E}_{\boldsymbol{ext}}$ для результирующего значения параметра модели M. На основании данных векторов для всех вероятностей $p \in [0;1]$ с шагом 0.01 определим пары значений $\sigma_R^{app}, \sigma_R^{ext}$ и $\sigma_L^{app}, \sigma_L^{ext}$. Далее при помощи метода наименьших квадратов вычислим значения коэффициентов γ_1^L, γ_0^L и γ_1^R, γ_0^R, соответственно.

На рисунке 3.13 представлены фактические и модельные значения доверительных интервалов для вероятностей $p \in \{0.50, 0.51, \ldots, 1\}$ для временного ряда цен на электроэнергию европейской территории РФ.



Левый дов. интервал прогнозных значений $\hat{\sigma}_L^{ext} = \gamma_1^L \sigma_L^{app} + \gamma_0^L$



Правый дов. интервал прогнозных значений $\hat{\sigma}_R^{ext} = \gamma_1^R \sigma_R^{app} + \gamma_0^R$

Рис. 3.13. Моделирование доверительных интервалов

Среднее отклонение в процентах (2.10) рассчитанных значений $\hat{\sigma}_L^{ext}$ от фактических σ_L^{ext} и отклонений $\hat{\sigma}_R^{ext}$ от σ_R^{ext} границ доверительного интервала для исследованных временных рядов колеблется в диапазоне от 1% до 8%. Полученная точность моделирования границ доверительного интервала достаточна для работы с исследуемыми временными рядами.

Алгоритм построения доверительного интервала для прогнозных значений состоит из двух частей – определение модели границ доверительного интервала (3.23) и вычисление границ для новых прогнозных значений.

Для определения моделей границ доверительного интервала (3.23) необходимо выполнить следующие шаги.

1) Произвести прогнозирование тестового периода и оценить ошибки аппроксимации $\boldsymbol{E}_{app}$ и экстраполяции $\boldsymbol{E}_{ext}$, соответствующие окончательному значению параметра M.

2) На основании векторов ошибок для всех вероятностей $p \in \{0.50, 0.51, \ldots, 1\}$ определить пары значений границ доверительных интервалов $\sigma_R^{app}, \sigma_R^{ext}$ и $\sigma_L^{app}, \sigma_L^{ext}$, как описано выше.

3) При помощи метода наименьших квадратов вычислить коэффициенты зависимости (3.23).

Модели границ доверительного интервала также как и параметр модели M определяются в процессе идентификации модели экстраполяции. Для вычисления границ доверительного интервала новых прогнозных значений необходимо выполнить следующие шаги.

1) Произвести прогнозирование временного ряда и получить вектор ошибок аппроксимации $\boldsymbol{E}_{\boldsymbol{app}}^{\boldsymbol{M}}$.

2) На основании полученного вектора $\boldsymbol{E}_{\boldsymbol{app}}^{\boldsymbol{M}}$ определить значения границ доверительного интервала $\sigma_L^{app}, \sigma_R^{app}$.

3) На основании зависимости (3.23), вычислить границы доверительного интервала для экстраполированных значений $\hat{\sigma}_L^{ext}$ и $\hat{\sigma}_R^{ext}$.

После того, как определены доверительные интервалы прогнозных значений, задача прогнозирования временного ряда считается решенной (раздел 1.2.).

3.5. Выводы

1) Разработан метод прогнозирования на базе модели экстраполяции по выборке максимального подобия.

2) Разработаны алгоритмы экстраполяции временного ряда с учетом и без учета внешних факторов. Исследования показали высокую скорость вычислений прогнозных значений.

3) Предложен алгоритм идентификации параметра модели M. Алгоритм идентификации содержит вычисления, которые могут выполняться параллельно по предложенной схеме.

4) Произведена оценка времени последовательных вычислений для решения задачи идентификации модели.

5) Разработан алгоритм построения доверительного интервала для прогнозных значений.

ГЛАВА 4. ПРОГРАММНАЯ РЕАЛИЗАЦИЯ И ОЦЕНКА ЭФФЕКТИВНОСТИ МОДЕЛИ ЭКСТРАПОЛЯЦИИ ПО ВЫБОРКЕ МАКСИМАЛЬНОГО ПОДОБИЯ

4.1. Прогнозирование показателей энергорынка РФ

Оптовый рынок электроэнергии и мощности РФ существует с сентября 2009 года. Как и всякий либерализованный энергорынок рынок РФ устроен таким образом, чтобы каждый его участник — генерирующая компания или потребитель электроэнергии, планировали как можно точнее объемы своей выработки и энергопотребления. От точности планирования своей работы на рынке зависит финансовый результат участника. Во многих компаниях существуют подразделения планирования или прогнозирования, которые решают поставленные рынком задачи. Для более эффективного решения задач планирования в качестве входной информации требуются не только показатели собственного потребления или выработки, но и общерыночные.

Одним из поставщиков общерыночной информации является ЗАО «РусПауэр», созданное в 2008 году. Компания предоставляет участникам энергорынка информацию в виде специальных отчетов, сгруппированных по так называемым продуктам. На сегодняшний день одним из продуктов «РусПауэр» являются «Прогнозы» — набор отчетов, содержащих прогнозные значения по 19 временным рядам общих показателей энергорынка РФ для трех горизонтов [49]. Разработка программного комплекса для прогнозирования общих показателей рынка с целью формирования продукта «Прогнозы» по заказу компании «РусПауэр» является частью настоящей диссертации.

4.1.1. Программная реализация

Для создания аналитического продукта «Прогнозы», содержащего прогнозные значения 19 показателей оптового рынка электроэнергии РФ, компанией «РусПаэур» была поставлена задача реализации алгоритмов прогнозирования как самостоятельного серверного приложения, способного работать без вмешательства экспертов. На сегодняшний день подавляющее большинство серверов используют семейство операционных систем UNIX. При выборе языка программирования и системы управления базами данных (СУБД) принимались в расчет следующие требования:

— разработка и эксплуатация серверного приложения под управлением широкого набора операционных систем;

— наличие готовых библиотек, содержащих реализацию известных математических функций, а также библиотек, реализующих взаимодействие приложения с различными источниками данных, без лицензионных ограничений;

— наличие библиотек для обращения к серверу системы управления базами данных без лицензионных ограничений.

Для разработки был выбран компилируемый язык программирования JAVA [50], который

— работает под управлением наибольшего числа операционных систем,

— широко применяется для создания серверных приложений и имеет в открытом доступе набор требуемых документированных библиотек без лицензионных ограничений,

— предоставляет средства для разработки приложений.

В качестве сервера СУБД был выбран MySQL [51], также работающий по управлением широкого набора операционных систем и имеющий высокую

производительность. На сегодняшний день MySQL является наиболее распространенной сервером СУБД, не имеющим лицензионных ограничений.

Согласно требованиям компании «РусПауэр» был создан программный комплекс, состоящий из функциональных блоков, представленных на рисунке 4.1.

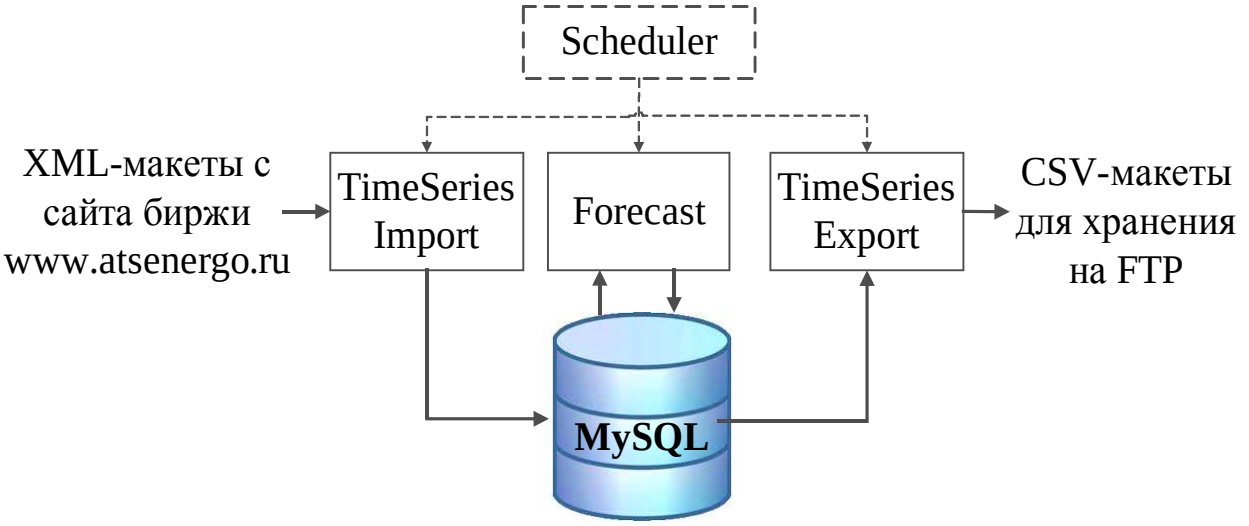


Рис. 4.1. Структура программного комплекса для прогнозирования показателей энергорынка РФ

Блок «Scheduler» выполняет управление процессом ежедневного прогнозирования. Для каждого временного ряда и каждого времени упреждения (сутки, неделя и месяц вперед) создаются три программные задачи:

— импорт фактических данных временного ряда за прошедшие сутки из XML-макетов с сайта биржи;

— прогнозирование временного ряда на модели EMMSP;

— экспорт прогнозных значений ряда в формат CSV.

Блок «Scheduler» управляет запуском по времени и проверкой корректности завершения каждой задачи в текущих сутках, а также создает новые задачи для расчета в будущих сутках. Созданная система прогнозирования не имеет графического интерфейса, работает автономно

и не требует вмешательства эксперта.

Блок «TimeSeries Import» извлекает из XML формата фактические значения для всех временных рядов энергопотребления и цен на электроэнергию, и загружает полученные значения временных рядов в базу данных.

Блок «Forecast» выполняет прогнозирование 19 временных рядов показателей энергорынка РФ при помощи модели EMMSP. В рамках данного модуля реализованы алгоритмы экстраполяции временных рядов с учетом и без учета внешних факторов (разделы 3.1., 3.2.). На сегодняшний день в виду текущей постановки задачи прогнозирования используется только модель EMMSP. Однако блок «Forecast» содержит реализованный алгоритм для прогнозирования на модели EMMSPX, который планируется использовать в дальнейшем.

Алгоритмы идентификации модели и оценки доверительных интервалов (3.3. и 3.4.) реализованы в программной среде MATLAB. Согласно требованиям компании «РусПауэр», разработанное серверное приложение должно предоставлять лишь прогнозные значения показателей энергорынка без оценки доверительного интервала.

Блок «TimeSeries Export» создает CSV файл, содержащий прогнозные значения временных рядов показателей энергорынка РФ и размещает файлы на FTP-сервере.

Разработанная система отвечает требованиям ЗАО «РусПауэр». В разделах 4.1.2. и 4.1.3. представлены результаты прогнозирования показателей энергорынка РФ, в том числе результаты прогнозирования выполнены по заказу компании «РусПауэр».

4.1.2. Прогнозирование цен на электроэнергию

Целью прогнозирования цен на электроэнергию – цен рынка на

сутки вперед и цен балансирующего рынка – является определение будущих значений, которые необходимы участникам энергорынка для планирования работы. Генерирующие компании на основании прогноза цен на месяц вперед планируют расход топлива (в первую очередь газа и угля) на выработку электроэнергии; на основании прогноза цен на неделю вперед генерирующие компании планируют состав включенного оборудования и на основании прогноза на сутки вперед планируют краткосрочный график нагрузки станции [52]. Компаниям-потребителям прогноз цен необходим для финансового планирования [53]. Обоим типам компаний прогноз цен на электроэнергию необходим для оценки и хеджирования (скрытия) финансовых рисков.

Задача прогнозирования цен на электроэнергию является новой для России в связи с тем, что отечественный рынок является одним из самых молодых рынков электроэнергии и мощности. Особенность задачи прогнозирования цен для России состоит в том, что по мере реформирования энергорынка алгоритм расчета цен подвергается изменениям. Цены с 01.09.2006 до 01.01.2008 рассчитывались по одному алгоритму, затем алгоритм был изменен.

Исходные временные ряды цен энергорынка РФ предоставлены Открытым акционерным обществом «Системный оператор Единой энергетической системы» (далее «СО ЕЭС») и компанией «РусПауэр».

Временные ряды цен рынка на сутки вперед содержат почасовые равноотстоящие значения в руб/МВт·ч за период с 01.09.2006 по 07.08.2011, их параметры приведены в таблице 7 (№1 – 7). Временные ряды индексов хабов, т. е·цен в специально определенных зонах, (№8 – 12 таблицы 7) содержат значения за период с 15.06.2010 по 15.06.2011. Временные ряды цен балансирующего рынка для европейской территории содержат значения за

период 01.01.2007 по 15.12.2009 (№13 таблицы 7), для сибирской территории за период с 21.02.2008 по 15.12.2009 (№14 таблицы 7)

Таблица 7.

Параметры временных рядов цен на электроэнергию в руб/МВт·ч

№	Временной ряд	Длина ряда	Среднее значе-ние	Стандарт. отклоне-ние	Мин. знач.	Макс. знач.
1	Цена РСВ ЕЦЗ	43224	717	237	0	2135
2	Цена РСВ СЦЗ	43224	423	168	0	1030
3	Цена РСВ ОЭС Урала	43224	696	217	0	2715
4	Цена РСВ ОЭС Средней Волги	43224	717	251	0	2128
5	Цена РСВ ОЭС Юга	43224	795	272	0	2396
6	Цена РСВ ОЭС Северо-Запада	43224	686	247	0	2220
7	Цена РСВ ОЭС Центра	43224	728	255	0	2268
8	Индекс Центр	10032	971	171	48	2216
9	Индекс Юг	10032	1044	198	9	2366
10	Индекс Урал	10032	907	144	530	1879
11	Индекс Восточная Сибирь	10032	484	74	87	726
12	Индекс Западная Сибирь	10032	563	87	282	955
13	Цена БР ЕЦЗ	25920	629	250	0	3309
14	Цена БР СЦЗ	15936	481	173	0	1372
15	Цена БР ОЭС Урала	25920	622	242	0	3046
16	Цена БР ОЭС Средней Волги	25920	621	258	0	3095

Таблица 7. – окончание

№	Временной ряд	Длина ряда	Среднее значение	Стандарт. отклонение	Мин. знач.	Макс. знач.
17	Цена БР ОЭС Юга	25920	702	300	0	4033
18	Цена БР ОЭС Северо-Запада	25920	616	270	0	3401
19	Цена БР ОЭС Центра	25920	634	265	0	3172
Абб.: РСВ – рынок на сутки вперед; ЕЦЗ – европейская ценовая зона; СЦЗ – сибирская ценовая зона; ОЭС – объединенная энергосистема; БР – балансирующий рынок.						

Прогнозирование временных рядов цен осуществлялось на неделю вперед и на сутки вперед. Время упреждения указано в таблице 8, содержащей результаты расчетов:

— время упреждения $P=24$ – прогнозирование на сутки вперед;

— время упреждения $P=168$ – прогнозирование на неделю вперед.

Прогнозирование цен на электроэнергию и энергопотребления на месяц вперед выполняется на модели ARIMA и не является частью диссертации. Контрольный период для каждого временного ряда указан в таблице 8.

Наборы моделей №1–12 созданы по дням недели: каждый день недели имеет собственный параметр модели M (приложение 1, таблицы 17 – 28).

Таблица 8.

Результаты прогнозирования цен рынка на сутки вперед

№	Временной ряд	Контрольный период	Время упреждения	Параметр модели *M*	MAE (MAPE)
1	Цена РСВ ЕЦЗ	01.09.10 – 07.08.11 (более 8000 значений)	24	Набор №1	47 (4.84%)
			168	288	50 (5.07%)
2	Цена РСВ СЦЗ		24	Набор №2	39 (7.22%)
			168	228	56 (10.14%)
3	Цена РСВ ОЭС Урала	01.04.11 – 07.08.11 (более 3000 значений)	24	Набор №3	45 (4.69%)
			168	384	60 (6.32%)
4	Цена РСВ ОЭС Средней Волги		24	Набор №4	42 (4.21%)
			168	216	54 (5.49%)
5	Цена РСВ ОЭС Юга		24	Набор №5	61 (15.85%)
			168	264	84 (17.61%)
6	Цена РСВ ОЭС Северо-Запада		24	Набор №6	72 (7.86%)
			168	216	102 (11.14%)
7	Цена РСВ ОЭС Центра		24	Набор №7	45 (5.58%)
			168	144	61 (7.01%)
8	Индекс Центр	01.04.11 – 07.08.11 (более 3000 значений)	24	Набор №8	44 (5.16%)
			168	96	62 (6.98%)
9	Индекс Юг		24	Набор №9	60 (9.3%)
			168	240	83 (11.8%)
10	Индекс Урал		24	Набор №10	44 (4.64%)
			168	216	57 (6.01%)

Таблица 8. – окончание

№	Временной ряд	Контроль-ный период	Время упрежде-ния	Параметр модели *M*	MAE (MAPE)
11	Индекс Восточная Сибирь	01.04.11 – 07.08.11 (более 3000 значений)	24	Набор №11	42 (8.25%)
			168	360	52 (10.47%)
12	Индекс Западная Сибирь		24	Набор №12	50 (8.55%)
			168	144	78 (13.45%)

Полученные для краткосрочного прогнозирования цен значения MAPE лежат в диапазоне от 4.21% до 15.85% для прогнозирования на сутки вперед; в диапазоне от 5.07% до 17.61% для прогнозирования на неделю вперед.

Прогнозирование цен балансирующего рынка выполнено двумя моделями: EMMSP и EMMSPX. Бизнес-процессы энергорынка РФ устроены таким образом, что при прогнозировании цен балансирующего рынка доступны фактические значения цен рынка на сутки вперед, а также объем планового энергопотребления. Прогнозирование осуществлялось на 24 значения вперед, контрольным периодом являлся период с 01.03.2009 по 30.09.2009 (более 5 000 значений). Результаты прогнозирования представлены в таблице 9.

Таблица 9.

Результаты прогнозирования цен балансирующего рынка

№	Временной ряд	Модель	Параметр M	Внешний фактор	MAE (MAPE)
1	Цена БР ЕЦЗ	EMMSP	360	–	81.45 (13%)
		EMMSPX	312	Цена РСВ ЕЦЗ	44.71 (7%)
			156	Энергопотребление ЕЦЗ	63.16 (10%)
			708	Цена РСВ ЕЦЗ и энергопотребление ЕЦЗ	42.59 (7%)
2	Цена БР СЦЗ	EMMSP	84	–	100.01 (21%)
		EMMSPX	336	Цена РСВ ЕЦЗ	81.79 (17%)
			288	Энергопотребление ЕЦЗ	96.25 (20%)
			444	Цена РСВ ЕЦЗ и энергопотребление ЕЦЗ	80.63 (17%)
3	Цена БР ОЭС Урала	EMMSP	360	–	73.2 (12%)
		EMMSPX	192	Цена РСВ ЕЦЗ	53.77 (9%)
			168	Энергопотребление ЕЦЗ	64.06 (10%)
			612	Цена РСВ ЕЦЗ и энергопотребление ЕЦЗ	52.1 (8%)

Таблица 9. – продолжение

№	Временной ряд	Модель	Параметр M	Внешний фактор	MAE (MAPE)
4	Цена БР ОЭС Средней Волги	EMMSP	72	–	80.55 (13%)
		EMMSPX	360	Цена РСВ ЕЦЗ	44.76 (7%)
			228	Энергопотребление ЕЦЗ	62.25 (10%)
			708	Цена РСВ ЕЦЗ и энергопотребление ЕЦЗ	44.94 (7%)
5	Цена БР ОЭС Центра	EMMSP	360	–	74.39 (12%)
		EMMSPX	312	Цена РСВ ЕЦЗ	50.15 (8%)
			204	Энергопотребление ЕЦЗ	67.16 (11%)
			684	Цена РСВ ЕЦЗ и энергопотребление ЕЦЗ	46.08 (7%)
6	Цена БР ОЭС Северо-Запада	EMMSP	348	–	80.32 (13%)
		EMMSPX	348	Цена РСВ ЕЦЗ	59.18 (10%)
			228	Энергопотребление ЕЦЗ	69.46 (11%)
			468	Цена РСВ ЕЦЗ и энергопотребление ЕЦЗ	58.02 (9%)

Таблица 9. – окончание

№	Временной ряд	Модель	Параметр M	Внешний фактор	MAE (MAPE)
7	Цена БР ОЭС Юга	EMMSP	336	–	112.48 (16%)
		EMMSPX	360	Цена РСВ ЕЦЗ	71.23 (10%)
			244	Энергопотребление ЕЦЗ	82.86 (12%)
			612	Цена РСВ ЕЦЗ и энергопотребление ЕЦЗ	69.77 (10%)
Абб.: БР – балансирующий рынок, ЕЦЗ – европейская ценовая зона, СЦЗ – сибирская ценовая зона, РСВ – рынок на сутки вперед.					

Результаты прогнозирования цен балансирующего рынка показывают, что учет внешних факторов в модели благоприятно сказывается на точности прогнозирования. Для всех временных рядов из таблицы 9 ошибка прогнозирования снизилась на 4 – 7% при учете двух внешних факторов в сравнению с моделью без учета внешних факторов. Таким образом, утверждается, что разработанная модель экстраполяции временных рядов, как и некоторые другие модели (1.3.1., 1.3.2., 1.3.4. и 1.3.6.) способны эффективно учитывать влияние внешних факторов, повышая точность прогнозирования.

На сегодняшний день результаты прогнозирования цен энергорынка РФ, представленные в диссертации, являются одними из первых, находящихся в открытом доступе. Кроме результатов, приведенных в таблицах 8 и 9, в статьях [54-57] приведены результаты краткосрочного прогнозирования цен на электроэнергию за другие контрольные периоды.

Сравнение эффективности EMMLS с моделью ANN. Очевидно, что

корректно сравнить эффективность различных моделей прогнозирования можно только при условии, что эти модели используются для прогнозирования одного и того же временного ряда при единой постановке задачи и едином контрольном периоде. В рамках диссертации в ряде случаев такое сравнение проводилось.

Точность прогнозирования модели EMMLS сравнивалась с аналогичной точность прогнозирования модели ANN, разработанной Обществом с ограниченной ответственностью «BIGROUP LABS», (далее BIGroup Labs). Компания BIGroup Labs была создана в 2004 году для продвижения новейших информационных технологий на энергорынок РФ. В 2009 году BIGroup Labs разработала и программно реализовала специализированную модель на базе нейронных сетей (ANN) для прогнозирования цен РСВ и энергопотребления, которая внедрена в ряде энергосбытовых компаний, а также на промышленных предприятиях.

Временной ряд цен РСВ ЕЦЗ (временной ряд №1 в таблице 7) прогнозировался на обеих моделях на 24 значения вперед на контрольном периоде с 01.03.2009 по 30.09.2009 (более 5 000 значений). Программный комплекс прогнозирования цен РСВ, разработанный «BIGroup Labs», носит название BI EnergoPrice [58]. Сравнение точности прогнозирования двух моделей представлено в таблице 10.

Таблица 10.

Сравнение точности EMMSP и ANN для цен РСВ ЕЦЗ

№	Модель	Параметр M	MAE, руб/МВт·ч	MAPE, %	Кол-во часов точнее
1	EMMSP	360	36.98	6.68	–
2	EMMSP	Набор №13	31.88	5.97	48%
3	ANN	–	31.27	6.10	52%
4	ANN + EMMSP	Формула (4.1)	28.22	5.40	–

Модель №1 в таблице 10 имеет один параметр $M=360$, ее точность ниже остальных моделей. Модель №2 является набором моделей, представленной в приложении (таблица 29); ошибка данной модели ниже на 0.71% ошибки модели №1. Результаты прогнозирования подтверждают, что применение наборов способно повысить точность прогнозирования (раздел 3.3.3.). Модель №3 разработана «BIGroup Labs».

Сравнение ошибок проводилось между моделью №2 и №3 и показало, что

— значения MAE моделей №2 и №3 практически одинаковы,

— значения MAPE моделей №2 и №3 также практически одинаковы,

— в 52% часов исследуемого периода модель №3 была точнее, в 48% часов – была точнее модель №2.

Модель №4 представленная в таблице 10 является суммой прогнозных значений модели №2 и №3. Исследование ошибки показало, что наибольшую точность имеет комбинация

$$\hat{Z}(t)=0.53\,\hat{Z}(t)_{№2}+0.48\,\hat{Z}(t)_{№3}-9.80\,. \qquad (4.1)$$

Здесь окончательный результат прогнозирования получается как

линейная комбинация результатов прогнозирования моделей №2 и №3. Модель №4 имеет максимальную точность, что показывает высокую эффективность подхода использования двух различных моделей для прогнозирования временного ряда цен рынка на сутки вперед европейской территории РФ.

Аналогичный подход к прогнозированию часто применяется в экономике, политике и других предметных областях и называется консенсус-прогноз (consensus forecast) [59]. При формировании консенсус-прогноза в расчет принимают два и более прогноза, выполняемых независимыми организациями или моделями. В работе [59] утверждается, что точность консенсус-прогноза может быть выше точности каждого из прогнозов, принимаемых во внимание. В рассмотренном случае консенсус-прогноз определяется как линейная комбинация двух независимых прогнозов.

На основании результатов работы по сравнению эффективности моделей заключим, что если две модели прогнозирования имеют приблизительно одинаковую точность, то имеет смысл исследовать модели, являющейся суммой прогнозных значений аналогично модели №4. Данный подход к прогнозированию энергопотребления продемонстрирован в разделе 4.1.3.; к прогнозированию сахара крови – в разделе 4.2.. Во всех трех случаях линейная комбинация прогнозных значений двух различных моделей давала результат точнее, чем каждая из моделей в отдельности.

Сравнение точности прогнозирования цен на электроэнергию с точностью прогнозирования цен энергорынков Испании, Скандинавии и Онтарио (Канада) проводилось в рамках оценки эффективности предложенной модели прогнозирования. Выше отмечалось, что задача экстраполяции цен энергорынка РФ является новой, потому нет возможности

выполнить широкое сравнение точности краткосрочного прогнозирования.

В работах [13,60-62] исследуются цены энергорынка Испании. В работе [60] полученные значения MAPE при прогнозировании временного ряда цен 2000 года при помощи модели GARCH на 24 значения вперед колеблются от 2.90% до 10.40%. Во второй работе [61], также относящейся к ценам энергорынка Испании 2000 года приведены значения MAPE в диапазоне от 4.62% до 19.93%. В более поздней работе [62] для краткосрочного прогнозирования цен энергорынка Испании в 2001 году применялась регрессионная модель. Значения MAPE, полученные автором [62], колеблются в диапазоне от 4.93% до 8.31%. В самой поздней работе из рассматриваемых [13] использовалась комбинированная модель на основании ARIMA и вейвлет-преобразования. Значения MAPE для цен Испании 2002 года находятся в пределах от 4.78% до 13.78% для различных недель года.

В работе [5] проводилось исследование цен рынка Скандинавии Nordpool 2004 года. Полученные значения MAPE находятся в диапазоне от 2.54% до 13.40% для различных недель. В работе заключается, что полученные оценки MAPE подтверждают эффективность разработанной модели прогнозирования для цен энергорынка Nordpool.

В двух работах [63,64] исследуются цены энергорынка Онтарио 2004 года (Ontario electricity market, Канада). В первой работе [63] исследовались как линейные так и нелинейные модели. Наибольшую адекватность показала модель, для которой при краткосрочном прогнозировании значение MAPE составило 16.10%. Во второй работе [64] для аналогичного контрольного периода, что и в работе [63], применялась модель на основании адаптивных регрессионных сплайнов. Оценка MAPE для второй модели находится в диапазоне от 8.60% до 13.90%, что точнее, чем в первой работе.

Во всех указанных работах отмечается, что прогноз цен на

электроэнергию с оценкой MAPE до 10 – 15% является достаточно эффективным для использования в планировании. Таким образом сравнение с западными оценками для аналогичных задач показывает, что разработанная в диссертации модель EMMSP эффективна для прогнозирования цен на электроэнергию, так как полученные значения MAPE практически для всех исследованных временных рядов находится в пределах указанного диапазона (таблицы 8, 9 и 10). Отдельно отметим, что в рассмотренных работах о прогнозировании цен западных рынков оценки MAPE давались, как правило, для контрольного периода, содержащего несколько десятков или сотен значений (отдельно взятые дни или недели); в рамках диссертации оценки MAPE приводятся для контрольных периодов в несколько тысяч значений.

По итогам прогнозирования цен на электроэнергию можно сделать следующие выводы.

1) Модели EMMSP и EMMSPX применялись для прогнозирования 19 временных рядов цен энергорынка РФ общей длиной более 500 000 значений. Точности прогнозирования указанных временных рядов, представленные в работе, являются первыми полными опубликованными в открытом доступе по энергорынку РФ.

2) Сравнение эффективности разработанной в диссертации модели с программным комплексом компании BIGroup Labs показало сравнимую эффективность модели EMMSP и ANN для исследуемого временного ряда.

3) Сравнение точности прогнозирования цен рынка на сутки вперед с точностью аналогичных западных рынков показало высокую эффективность разработанной модели. В большинстве случаев ошибка краткосрочного прогнозирования цен при помощи EMMSP не выходит за границы 10%, что по оценкам западных специалистов является

высокоэффективным.

4) В настоящее время ежедневно компания «РусПауэр» при помощи разработанного в рамках диссертации программного комплекса формирует прогнозы по 12 временным рядам (№1 — 12 таблицы 7) цен рынка на сутки вперед в виде аналитического продукта, используемого участниками энергорынка РФ в повседневной работе [49].

4.1.3. Прогнозирование энергопотребления

Целью прогнозирования энергопотребления является поддержание надежной работы единой энергосистемы РФ. Прогноз энергопотребления необходим, в первую очередь, системному оператору для балансирования энергосистемы РФ. С введением рынка электроэнергии и мощности взамен общему планированию каждая компания-потребитель самостоятельно прогнозирует собственное энергопотребление. Система финансовых расчетов на энергорынках устроена таким образом, чтобы мотивировать потребителей как можно точнее планировать собственное потребление: чем точнее прогноз энергопотребления, тем выше финансовый результат. В связи с этим каждая компания-потребитель заинтересована в предельно точном прогнозе собственного потребления.

Задача прогнозирования энергопотребления отличается от задачи прогнозирования цен на электроэнергию. При прогнозировании цен существует экспертно полученная оценка точности (значение MAPE 10 – 15%, 4.1.), при достижении которой прогнозные значения можно использовать для решения последующих задач и далее не заниматься усовершенствованием модели. При прогнозировании энергопотребления борьба за десятые доли процентов ведется постоянно, так как расходы компании на покупку электроэнергии напрямую зависят от точности прогноза собственного энергопотребления. В работе [53] отмечается, что при

повышении средней точности прогнозирования энергопотребления ОАО «СахаЭнерго» с 5% до 4.3% предприятие в год экономит 20.4 млн. руб.

Исходные временные ряды энергопотребления предоставлены «АТС», «СО ЕЭС», «РусПауэр» и Открытым акционерным обществом «Сибирьэнерго» (далее «Сибирьэнерго»).

Временные ряды энергопотребления содержат почасовые равноотстоящие значения в МВт·ч за период с 01.09.2006 по 07.08.2011, их параметры приведены в таблице 11 (№1 – 7). Временной ряд энергопотребления «Сибирьэнерго» (№8 таблицы 11) содержит значения за период с 01.01.2005 по 19.05.2008.

Таблица 11.

Параметры временных рядов энергопотребления в МВт·ч

№	Временной ряд	Длина ряда	Среднее значение	Стандарт. отклонение	Мин. знач.	Макс. знач.
1	Энергопотребление ЕЦЗ	43224	82348	11133	57847	111723
2	Энергопотребление СЦЗ	43224	22373	3070	15329	30666
3	Энергопотребление ОЭС Урала	43224	27347	2969	19959	35099
4	Энергопотребление ОЭС Средней Волги	43224	11179	1947	6085	16640
5	Энергопотребление ОЭС Юга	43224	8726	1409	5757	12990

Таблица 11. – окончание

№	Временной ряд	Длина ряда	Среднее значе-ние	Стандарт. отклоне-ние	Мин. знач.	Макс. знач.
6	Энергопотребление ОЭС Северо-Запада	43224	7614	1249	4825	11374
7	Энергопотребление ОЭС Центра	43224	24559	4194	15604	36171
8	Энергопотребление «Сибирьэнерго»	29640	1300	369	516	2 312
Аббревиатуры: ЕЦЗ – европейская ценовая зона; СЦЗ – сибирская ценовая зона; ОЭС – объединенная энергосистема.						

Прогнозирование временных рядов энергопотребления осуществлялось на два горизонта – на неделю вперед и на сутки вперед. Для каждой модели в таблице 12 указано время упреждения. Наборы моделей приведены в приложении (таблицы 30 – 37).

Таблица 12.

Результаты прогнозирования временных рядов энергопотребления

№	Временной ряд	Контроль-ный период	Время упрежде-ния	Параметр модели *M*	MAE (MAPE)
1	Энергопотребление ЕЦЗ	01.09.10 – 07.08.11 (более 8000 значений)	24	Набор №14	970 (1.12%)
			168	264	1011 (1.31%)
2	Энергопотребление СЦЗ		24	Набор №15	373 (1.65%)
			168	120	391 (1.99%)
3	Энергопотребление ОЭС Урала	01.04.11 – 07.08.11 (более 3000 значений)	24	Набор №16	234 (0.91%)
			168	360	324 (1.26%)
4	Энергопотребление ОЭС Средней Волги		24	Набор №17	170 (1.65%)
			168	144	193 (1.8%)

Таблица 12. – окончание

№	Временной ряд	Контроль-ный период	Время упрежде-ния	Параметр модели *M*	MAE (MAPE)
5	Энергопотребление ОЭС Юга	01.04.11 – 07.08.11 (более 3000 значений)	24	Набор №18	152 (1.83%)
			168	168	280 (3.3%)
6	Энергопотребление ОЭС Северо-Запада		24	Набор №19	117 (1.67%)
			168	264	216 (3.04%)
7	Энергопотребление ОЭС Центра		24	Набор №20	327 (1.53%)
			168	264	643 (2.9%)
8	Энергопотребление «Сибирьэнерго»	19.05.07 – 19.05.08 (более 8000 значений)	24	Набор №21	41 (3.19%)

Полученные для краткосрочного прогнозирования энергопотребления значения MAPE лежат в диапазоне от 0.91% до 1.83% для прогнозирования на сутки вперед; в диапазоне от 1.26% до 3.30% для прогнозирования на неделю вперед.

Кроме результатов, приведенных в таблице 12, в статьях [54-57] приведены результаты краткосрочного прогнозирования энергопотребления за другие контрольные периоды.

Сравнение достигнутой точности с оценками точности прогнозирования энергопотребления, представленными в научных работах за последние годы. На сегодняшний день существует множество моделей для решения задачи прогнозирования энергопотребления, например:

— западные работы [9,11,29,37,48,65-67];

— российские работы [53,68-70].

Во всех указанных работах отмечена важность задачи прогнозирования энергопотребления.

Таблица 13.

Обзор работ по прогнозированию энергопотребления

№	Работа, год публикации	Временной ряд энергопотребления	Полученная точность, MAPE
1	[9], 2006	Шанхайская энергосистема (Shanghai Power Grid)	2.8 – 3.4% в зависимости от модели
2		Подстанции Франкфурта (Frankfurt Substation)	2.04%
3	[65], 2010	Обзор методов прогнозирования	1.26 – 4.81% в зависимости от модели
4	[29], 2009	Энергосистема Виктории (Victorian Power System, Австралия)	2.64%
5	[48], 2008	Энергосистема штата Орисса (Восточная Индия)	2.96 – 5.27% в зависимости от алгоритма обучения модели
6	[67], 2008	Энергосистемы различных стран Европы (10 стран)	0.80 – 2.90% для различных моделей, стран и времени упреждения
7	[11], 2010	Энергосистема Малайзии	0.99%
8	[68], 2010	Энергосистема Костромской области	2 – 5% в зависимости от модели
9	[53], 2009	Потребление поселка Жиганск, Республика Саха (Якутия)	3 – 5% для различных моделей

Таблица 13. – окончание

№	Работа, год публикации	Временной ряд энергопотребления	Полученная точность, MAPE
10	[69], 2011	Энергопотребление ОАО «Мордовская энергосбытовая компания»	1.43 – 2.75% для различных дней недели
11	[70], 2007	Энергопотребление ОАО «Костромская энергосбытовая компания»	2 – 5% для различных дней недели

Из таблицы видно, что в настоящее время задача краткосрочного прогнозирования энергопотребления актуальна для различных стран. Приведенные в работах оценки значений MAPE колеблются в диапазоне от 0.80% до 5.27% в то время как точности, достигнутые с использованием EMMSP, колеблются в диапазоне от 0.91% до 1.83%. На этом основании заключим, что разработанная в диссертации модель прогнозирования энергопотребления является высокоэффективной.

Сравнение EMMSP с моделью ARIMA от компании iRM (www.irm.at, Австрия) производилось на основании результатов конкурса «Сибирьэнерго» в 2008 году.

В рамках конкурсной задачи компания «Сибирьэнерго» предоставляла данные по собственному энергопотреблению, а участники конкурса на ежедневной основе формировали прогнозные значения. Закрытое акционерное общество «Верисел Проекты» (www.vp.ru, Россия) принимало участие в данном конкурсе с продуктом компании iRM для решения задачи прогнозирования энергопотребления iOPT PRO на базе модели ARIMAX. Компания iRM со своим продуктом iOPT является одним из европейских лидеров по разработке программных продуктов для автоматизации торговли, прогнозирования, управления рисками на энергорынках Европы.

Контрольный период длился с 5 по 19 мая 2008 года, то есть составил 360 часов. В рамках диссертации произведен прогноз указанного временного диапазона на разработанной модели EMMLS.

По итогам конкурса по оценке «Сибирьэнерго» продукт iOPT PRO занял второе место среди участников конкурса, уступив 0.1% компании-победителю. Таким образом, модель ARIMAX показала высокую эффективность при прогнозе энергопотребления «Сибирьэнерго».

Сравнение результатов прогнозирования продукта iOPT PRO и результатов прогнозирования на EMMSP представлены в таблице 14.

Таблица 14.

Сравнение точности EMMSP и ARIMAX при прогнозе энергопотребления «Сибирьэнерго»

№	**Модель**	**Параметр** M	**MAE, МВт·ч**	**MAPE, %**	**Кол-во часов точнее**
1	EMMSP	132	46.20	4.48%	–
2	EMMSP	Набор №21	45.05	4.32%	50%
3	ARIMAX	–	44.63	4.21%	50%
4	ARIMA + EMMSP	Формула (4.2)	31.79	3.01%	–

Модель №1 из таблицы 14 имеет один параметр $M=132$, эта модель показала наименее точный результат. Модель №2 является набором моделей, представленным в приложении (таблица 37). Как и в случае с ценами энергорынка модель №2 имеет более высокую точность прогнозирования, что показывает эффективность применения наборов. Результаты прогнозирования модели №3 предоставлены компанией iRM.

Сравнение ошибок проводилось между моделью №2 и №3 и показало,

что

— значения MAE моделей №2 и №3 практически одинаковы,

— значения MAPE модели №2 немного выше значения MAPE модели №3,

— ровно в половине часов исследуемого периода модель №3 была точнее, в другой половине – была точнее модель №2.

Модель №4, представленная в таблице 14, является консенсус-прогнозом моделей №2 и №3. Исследование ошибки показало, что наибольшую точность имеет комбинация

$$\hat{Z}(t) = 0.4830\,\hat{Z}(t)_{№2} + 0.4603\,\hat{Z}(t)_{№3} + 38.5605\,. \qquad (4.2)$$

В модели №4 окончательный результат прогнозирования получается как линейная комбинация результатов прогнозирования моделей №2 и №3. Как и в случае с ценами, рассмотренном в предыдущем разделе, консенсус-прогноз имеет максимальную точность, значительно улучшая показатели модели №2 и №3.

Проведенные исследования показали, что модель прогнозирования, являющаяся линейной комбинацией двух независимых моделей имеет наибольшую эффективность при прогнозировании энергопотребления.

По итогам прогнозирования энергопотребления можно сделать следующие выводы.

1) В рамках диссертации были исследованы 8 временных рядов энергопотребления, общая длина которых составляет более 300 000 значений.

2) Полученные значения MAPE при краткосрочном прогнозировании энергопотребления лежат в диапазоне от 0.91% до 1.83% и доказывают высокую эффективность применения EMMSP для решения данной задачи.

3) Значения MAPE для среднесрочного прогнозирования лежат

в диапазоне от 1.26% до 3.30% и сравнимы со значениями MAPE краткосрочного прогнозирования, приведенными в ряде новейших работ: 0.80 — 5.27%. Произведенные оценки точности результатов доказывают высокую эффективность применения разработанной модели для среднесрочного прогнозирования энегопотребления.

4) Сравнение результатов прогнозирования EMMSP и модели ARIMAX от компании iRM показало сравнимую эффективность моделей.

5) В настоящее время ежедневно компания «РусПауэр» при помощи разработанного в рамках диссертации программного комплекса формирует прогнозы по 8 временным рядам (№1 — 7 таблицы 11) энергопотребления в виде аналитического продукта, используемого участниками энергорынка РФ в повседневной работе [49].

4.2. Прогнозирование других временных рядов

В рамках диссертации были решены задачи прогнозирования других временных рядов:

— уровня сахара крови человека;

— скорости движения транспорта по городу Москва;

— финансовых показателей.

Программная реализация. Для решения задач по прогнозированию каждого из приведенных выше временных рядов было разработано отдельное программное обеспечение на базе специализированного комплекса MATLAB. Программный продукт MATLAB содержит пакеты прикладных программ для решения задач технических вычислений и одноимённый язык программирования, используемый в этом пакете. MATLAB используют более 1 000 000 инженерных и научных работников, он работает на большинстве

современных операционных систем, включая Linux, Mac OS, Solaris. На сегодняшний день MATLAB является наиболее широко применяемой математической средой для реализации технических вычислений, не требующих интеграции в информационное пространство предприятия [46].

4.2.1. Уровень сахара крови человека

Сахарный диабет первого типа – это метаболическое заболевание, вызванное абсолютным дефицитом секреции инсулина и характеризующееся неспособностью организма поддерживать уровень глюкозы в крови (BG – Blood Glucose) в целевом интервале 4 – 6 ммоль/л – в обычном состоянии и до 9 ммоль/л – после еды. Диабет вызывает множество опасных осложнений, избежать которые можно только путем контроля уровня BG и его удержания в физиологичном интервале. Основным путем решения этой задачи в настоящее время является введение в кровь пациента искусственных препаратов (генноинженерных человеческих инсулинов), которые могут симулировать действие эндогенного инсулина, вырабатываемого β-клетками здоровой поджелудочной железы [8].

Для решения задачи созданы системы непрерывного измерения уровня BG – Continuous Glucose Monitoring Systems (CGM-системы), а также системы непрерывного подкожного введения инсулина (инсулиновые помпы – insulin pumps). На основе CGM-систем и инсулиновых помп разработаны и интенсивно разрабатываются системы автоматического управления уровнем глюкозы в крови пациента. С алгоритмической точки зрения эти системы включают в себя две следующие основные подсистемы: подсистема прогнозирования уровня BG; подсистема определения оптимального времени и требуемой дозы инсулина.

В рамках настоящей работы проводилось сравнение эффективности прогнозирования двух моделей – модели на основе нейронных сетей

и модель экстраполяции временных рядов по выборке максимального подобия [71].

При прогнозе с помощью нейронных сетей используется одна нейронная сеть, на вход которой подаются предыдущие значения глюкозы, инсулина, принятых с пищей углеводов и физической нагрузки. На выходе нейронной сети получаем прогнозируемое значение глюкозы [71]. При прогнозе на EMMSP используются только фактические значения временного ряда BG без учета остальных показателей.

Рассматривался временной ряд уровней BG. Характеристики временного ряда приведены в таблице 15. Ставилась задача прогноза значений этого ряда на 20, 60 и 90 минут вперед (на 4, 12, 18 отсчетов).

Для сравнения моделей был произведен прогноз контрольного периода длиной около 3 800 значений на модели EMMSP и модели искусственной нейронной сети (ANN).

ANN-модель представляла собой нейронную сеть прямого распространения, принимающая на вход пять параметров (BG, инсулин, углеводы, гликемический индекс и физическую нагрузку) с использованием разреженно-суммирующей линии задержки длинной, равной три (т.е. общее число входов равно 15). Сеть обучалась алгоритмом Левенберга-Маркадта.

Таблица 15.

Параметры временного ряда BG

Временное разрешение	**Временной ряд**	**Длина ряда**	**Среднее значе-ние**	**Стандарт. отклоне-ние**	**Мин. знач.**	**Макс. знач.**
5 мин	BG, ммоль/л	29640	2.25	14.24	2.20	21.90

Сравнение результатов прогнозирования моделей приведено

в таблице 16.

Прогноз на моделях EMMSP(18) и EMMSP(78) для времени упреждения 20 минут и час, соответственно, в среднем имеет точность выше аналогичного прогноза с использованием ANN. Как видно из таблицы 16, значение MAPE для модели EMMSP(18) при $P=4$ составляет 5.07%, а при $P=12$ – 11.33%. Аналогично, MAPE для модели ANN при $P=4$ составляют 8.09%, а при $P=12$ – 13.21%. Важно отметить, что результаты прогнозирования позволяют выявить сильные изменения уровня BG (резкие увеличения и снижения этого уровня).

Таблица 16.

Точность прогноза моделей ANN и EMMSP

Модель	**P**	**MAE, ммоль/л**	**MAPE, %**	**Число точек точнее, %**	**Время экстраполяции, час**	**Время идентификации, час**
EMMSP(18)	4	0.36	5.07	45	0.35	0.5
ANN		0.30	4.12	55	80	80
EMMSP(78)	12	0.79	11.33	52	0.52	0.85
ANN		0.91	12.21	48	80	80
EMMSP(180)	18	0.97	14.70	58	0.10	2.5
ANN		1.12	17.02	42	80	80

При прогнозировании на полтора часа модель EMMSP(180) показала точность более, чем на 2% превышающую точность, достигнутую с помощью ANN.

В рамках работы с временным рядом BG для прогнозирования на полтора часа вперед исследовалась модель вида

$$\hat{Z}(t)=0.5416\,\hat{Z}(t)_{EMMSP}+0.5808\,\hat{Z}(t)_{ANN}-0.7326\,. \qquad (4.3)$$

Здесь, как и в разделах 4.1.2. и 4.1.3., консенсус-прогноз является линейной комбинацией результатов прогнозирования, полученных на моделях EMMSP и ANN соответственно. Исследование показало, что использованием комбинации (4.3) при прогнозировании на полтора часа позволяет существенно повысить точность и обеспечивает значение MAPE, равное 12.13%; значение MAE, равное 0.77 ммоль/л.

Проведенное сравнение моделей показало, что при краткосрочном прогнозировании уровня BG более точный результат дает нейронная сеть, при увеличении горизонта прогнозирования – модель EMMSP.

Важно отметить, что на качественном уровне прогнозирование, как с помощью нейронной сети, так и с помощью модели EMMSP, верно предсказывает факт роста/снижения уровня BG, а также скорость и пределы изменения этого уровня.

По результатам исследования можно сделать вывод, что экстраполяция методом максимального подобия, особенно в комбинации с прогнозированием с помощью нейронных сетей, дает достаточно точный прогноз уровня глюкозы в крови пациента. Этот прогноз может быть использован для принятии решения об оптимальной дозе инсулина, которая должна быть введена в данный момент времени [71].

4.2.2. Скорость движения транспорта по дорогам Москвы

С 1 марта по 16 мая 2010 года в рамках проекта «Интернет-Математика 2010» компания «Яндекс» проводила математический конкурс. В качестве конкурсной была предложена задача прогнозирования скорости движения транспорта по автомобильным дорогам города Москвы внутри одного дня на основе исторических данных [72].

По условиям конкурса исторические данные о скорости движения

транспорта (СДТ) охватывали 31 день: первые 30 дней содержали данные за период с 16:00 до 22:00 часов, для последнего дня — с 16:00 до 18:00. Файл с исходными данными содержал около 30 млн. значений СДТ. Согласно заданию необходимо было спрогнозировать около 700 тыс. значений СДТ для 29 335 дорог за период с 18:00 до 22:00 для последнего дня.

При использовании EMMSP для решения данной задачи каждая дорога рассматривалась как отдельный временной ряд без учета их взаимного влияния. Для каждого временного ряда из задания была создана модель прогнозирования EMMSP, после чего определены прогнозные значения. Точность прогнозирования участников конкурса компания «Яндекс» оценивала специальным индексом. Точность, которую удалось достичь при использовании EMMSP, составила 64.93 единицы.

В конкурсе «Интернет-Математика 2010» приняла участие 191 команда. Лучшим оказался результат с итоговой оценкой 58.92 единицы, алгоритм расчетов описан в статье [73]. По итогам конкурса результат EMMSP, равный 64.93, занял 38 место среди всех участников. Экспертная оценка точности прогнозирования СДТ аналитиков компании «Яндекс», так называемая Baseline, составляет 77.88 единиц. Таким образом, модель экстраполяции по выборке максимального подобия, не являющаяся специализированной моделью для решения поставленной задачи, показала эффективность, сравнимую с эффективностью специализированных решений для данной отрасли.

Постановка задачи прогнозирования СДТ, исходные файлы, а также итоговый рейтинг участников находится в открытом доступе [72].

4.2.3. Финансовые временные ряды

В рамках диссертации были также исследованы финансовые временные ряды:

— фьючерсные цены на природный газ на Нью-Йоркской товарной бирже (NYMEX, www.nymex.com) за период с 01.10.2007 по 01.05.2009 (около 7 месяцев);

— валютная пара GBP/USD (www.forex.com) за период с 03.11.2008 по 06.11.2009 (12 месяцев).

Результаты прогнозирования упомянутых временных рядов приведены в работах [74,75].

4.3. Выводы

1) Разработанный метод прогнозирования на базе модели экстраполяции по выборке максимального подобия реализован в виде серверного приложения, выполняющего прогнозирование показателей энергорынка РФ без участия эксперта на ежедневной основе.

2) Прогнозирование временных рядов цен на электроэнергию энергорынка РФ показало, что ошибка прогнозирования в большинстве случаев лежит в диапазоне 5 — 9%, что по оценкам специалистов западных рынков является высокоэффективным. Сравнение точности прогнозирования разработанной модели и нейросетевой модели от компании BIGroup Labs показало сравнимую эффективность моделей.

3) Прогнозирование временных рядов энергопотребления показало высокую эффективность реализованной модели: значения оценок ошибки краткосрочного и среднесрочного прогнозирования лежат в диапазоне от 0.91% до 3.30%, что сравнимо и точнее значений аналогичных оценок точности прогнозирования энергопотребления, приведенных в ряде новейших работ.

4) Реализация предложенной модели прогнозирования при помощи

математического пакета MATLAB показала высокую точность прогнозирования временного ряда уровня сахара крови человека на один и полтора часа вперед в сравнении с нейросетевой моделью. Реализация предложенной модели при помощи MATLAB для прогнозирования скорости движения транспорта показала сравнимую со специализированными моделями точность прогнозирования.

5) Проведенные эксперименты по формированию консенсус-прогноза на основании линейной комбинации двух независимых исследуемых прогнозов во всех трех случаях приводили к существенному повышению точности прогнозирования.

ВЫВОДЫ

1) Задача прогнозирования временных рядов актуальна и решается на основании модели прогнозирования. Одним из наиболее используемых классов моделей прогнозирования является класс авторегресионных моделей. Установлено, что основным недостатком данного класса является большое число свободных параметров, требующих определения. Определено перспективное направление развития моделей прогнозирования, позволяющее устранить указанный недостаток.

2) Разработана новая модель прогнозирования временных рядов по выборке максимального подобия для двух видов постановки задачи прогнозирования временного ряда — с учетом и без учета внешних факторов. Новая модель относится к авторегрессионному классу моделей и имеет единственный параметр, что упрощает задачу идентификации модели, устраняя основной недостаток моделей данного класса.

3) Разработан новый метод прогнозирования на основе предложенной модели, содержащий набор алгоритмов для экстраполяции временных рядов, идентификации модели и построения доверительного интервала прогнозных значений. Произведена оценка времени последовательных вычислений при решении задач экстраполяции временного ряда и идентификации модели. Предложена схема параллельных вычислений, позволяющая сократить время расчета при решении задачи идентификации.

4) Выполнена программная реализация разработанных алгоритмов средствами математического пакета MATLAB. По заказу компании «РусПауэр» создано специализированное серверное приложения для прогнозирования показателей энергорынка РФ на ежедневной основе. Приложение работает в автоматическом режиме и предоставляет прогнозные значения показателей без вмешательства эксперта.

5) Произведена оценка эффективности новой модели прогнозирования. Применение новой модели для прогнозирования показателей энергорынка РФ показало высокую эффективность предложенной модели. Применение новой модели для прогнозирования временных рядов уровня сахара крови больных сахарным диабетом первого типа и скорости движения транспорта по дорогам г. Москва показали эффективность, сравнимую со специализированными моделями для данных областей.

ЛИТЕРАТУРА

1. Бокс Дж., Дженкинс Г.М. Анализ временных рядов, прогноз и управление. М.: Мир, 1974. 406 с.

2. Егошин А.В. Анализ и прогнозирование сложных стохастических сигналов на основе методов ведения границ реализаций динамических систем: Автореферат диссертации … канд. техн. наук. Санкт-Петербург, 2009. 19 с.

3. Gheyas I.A., Smith L.S. A Neural Network Approach to Time Series Forecasting // Proceedings of the World Congress on Engineering, London, 2009, Vol 2 [электронный ресурс]. P. 1292 – 1296. URL: www.iaeng.org /publication/WCE2009/WCE2009_pp1292-1296.pdf (дата обращения 28.08.2011).

4. Morariu N., Iancu E., Vlad S. A neural network model for time series forecasting // Romanian Journal of Economic Forecasting. 2009, No. 4. P. 213 – 223.

5. Mazengia D.H. Forecasting Spot Electricity Market Prices Using Time Series Models: Thesis for the degree of Master of Science in Electric Power Engineering. Gothenburg, Chalmers University of Technology, 2008. 89 p.

6. Нормативные системы в прогнозировании развития предпринимательского сектора экономики / Л.И. Муратова [и др.] // Управление экономическими системами [электронный ресурс]. 2009, №20. URL: http://uecs.mcnip.ru/modules.php?name=News&file=print&sid=145 (дата обращения 28.08.2011).

7. Parzen E. Long memory of statistical time series modeling // NBER-NSF Time Series Conference, USA, Davis, 2004 [электронный ресурс]. 10 p. URL: http://www.stat.tamu.edu/~eparzen/Long%20Memory%20of%20Statistical %20Time%20Series%20Modeling.pdf (дата обращения 28.08.2011).

8. Методы прогнозирования оптимальных доз инсулина для больных сахарным диабетом I типа. Обзор / С.А. Чернецов [и др.] // Наука и образование [электронный ресурс]. 2009, №9. URL: http://technomag.edu.ru/ doc/ 119663.html (дата обращения 28.08.2011).

9. Jingfei Yang M. Sc. Power System Short-term Load Forecasting: Thesis for Ph.d degree. Germany, Darmstadt, Elektrotechnik und Informationstechnik der Technischen Universität, 2006. 139 p.

10. Extrapolation // The free encyclopedia «Wikipedia» [электронный ресурс]. URL: http://en.wikipedia.org/wiki/Extrapolation (дата обращения 28.08.2011).

11. Norizan M., Maizah Hura A., Zuhaimy I. Short Term Load Forecasting Using Double Seasonal ARIMA Model // Regional Conference on Statistical Sciences, Malaysia, Kelantan, 2010. P. 57 – 73.

12. Collantes-Duarte J., Rivas-Echeverriat F. Time Series Forecasting using ARIMA, Neural Networks and Neo Fuzzy Neurons // WSEAS International Conference on Neural Networks and Applications, Switzerland, 2002 [электронный ресурс]. 6 p. URL: www.wseas.us/e-library/conferences/ switzerland2002/papers/464.pdf (дата обращения 28.08.2011).

13. Day-Ahead Electricity Price Forecasting Using the Wavelet Transform and ARIMA Models / A.J. Conejo [at al.] // IEEE transaction on power systems. 2005, Vol. 20, No. 2. P. 1035 – 1042.

14. Тихонов Э.Е. Прогнозирование в условиях рынка. Невинномысск, 2006. 221 с.

15. Леоненков А. Нечеткое моделирование в среде MATLAB и fuzzyTECH. СПб: БХВ-Петербург, 2005. 736 с.

16. Armstrong J.S. Forecasting for Marketing // Quantitative Methods in Marketing. London: International Thompson Business Press, 1999. P. 92 – 119.

17. Семенов В.В. Математическое моделирование динамики транспортных

потоков мегаполиса. М.: ИПМ им. М.В.Келдыша РАН, 2004. 44 с.

18. Self-organization in leaky threshold systems: The influence of near-mean field dynamics and its implications for earthquakes, neurobiology, and forecasting / J.B. Rundle [at al.] // Colloquium of the National Academy of Sciences, Irvine, USA, 2002. P. 2514 – 2521.

19. Draper N., Smith H. Applied regression analysis. New York: Wiley, In press, 1981. 693 p.

20. Maximum likelihood // The free encyclopedia «Wikipedia» [электронный ресурс]. URL: http://en.wikipedia.org/wiki/Maximum_likelihood (дата обращения 28.08.2011).

21. Ивахненко А.Г. Обзор задач, решаемых по алгоритмам Метода Группового Учета Аргументов (МГУА) // Group Method of Data Handling [электронный ресурс]. URL: http://www.gmdh.net/articles/rus/obzorzad.pdf (дата обращения 28.08.2011).

22. Autoregressive conditional heteroskedasticity // The free encyclopedia «Wikipedia» [электронный ресурс]. URL: http://en.wikipedia.org/wiki /Autoregressive_conditional_heteroskedasticit (дата обращения 28.08.2011).

23. Эконометрия: Учебное пособие / В.И. Суслов [и др.] Новосибирск: Издательство СО РАН, 2005. 744 с.

24. Prajakta S.K. Time series Forecasting using Holt-Winters Exponential Smoothing // Kanwal Rekhi School of Information Technology Journal [электронный ресурс]. 2004. 13 p. URL:http://www.it.iitb.ac.in/~praj/acads/ seminar/04329008_ExponentialSmoothing.pdf (дата обращения 28.08.2011).

25. Хайкин С. Нейронные сети: полный курс. М.: ООО «И. Д. Вильямс», 2006. 1104 с.

26. Pradhan R.P., Kumar R. Forecasting Exchange Rate in India: An Application of Artificial Neural Network Model // Journal of Mathematics Research. 2010, Vol.

2, No. 4. P. 111 – 117.

27. Yildiz B., Yalama A.,Coskun M. Forecasting the Istanbul Stock Exchange National 100 Index Using an Artificial Neural Network // An International Journal of Science, Engineering and Technology. 2008, Vol. 46. P.36 – 39.

28. An Artificial Neural Network Approach for Day-Ahead Electricity Prices Forecasting / J. Catalao [at al.] // 6th WSEAS international conference on Neural networks, USA, Stevens Point, 2005. P. 80 – 83.

29. Kumar M. Short-term load forecasting using artificial neural network techniques: Thesis for Master of Science degree in Electrical Engineering. India, Rourkela, National Institute of Technology, 2009. 48 p.

30. Zhu J., Hong J., Hughes J.G. Using Markov Chains for Link Prediction in Adaptive Web Sites // 1st International Conference on Computing in an Imperfect World, UK, London, 2002. P. 60 – 73.

31. Hannes Y.Y., Webb P. Classification and regression trees: A User Manual for IdentifyingIndicators of Vulnerability to Famine and Chronic Food Insecurity // International Food Policy Research Institute [электронный ресурс]. 1999. 59 p. URL: http://www.fao.org/sd/erp/toolkit/BOOKS/classification_and_regression_trees_intro.pdf (дата обращения 28.08.2011).

32. Huanga W., Nakamoria Y., Wangb S. Forecasting stock market movement direction support vector machine // Elsevier: computers and operation research. 2005, Vol. 32. P. 2513 – 2522.

33. Support vector machine // The free encyclopedia «Wikipedia» [электронный ресурс]. URL: http://en.wikipedia.org/wiki/Support_vector_machine (дата обращения 28.08.2011).

34. Mahfoud S., Mani G. Financial Forecasting Using Genetic Algorithms // Applied Artificial Intelligence. 1996, Vol. 10, No.6. P. 543 – 565.

35. Nogales F.J., Conejo A.J. Electricity price forecasting through

transferfunction models // Journal of the Operational Research Society. 2006, Vol. 57, No. 4. P. 350 – 356.

36. Alfares H.K., Nazeeruddin M. Electric load forecasting: literature survey and classifcation of methods // International Journal of Systems Science. 2002, Vol 33. P. 23 – 34.

37. Hinman J., Hickey E. Modeling and forecasting sort term electricity load using regression analysis // Journal of IInstitute for Regulatory Policy Studies [электронный ресурс]. 2009. 51 p. URL: http://www.irps.ilstu.edu/research /documents/LoadForecastingHinman-HickeyFall2009.pdf (дата обращения 28.08.2011).

38. Fogler H.R. A pattern recognition model for forecasting // Management science. 1974, No.8. P. 1178 – 1189.

39. Discovering Patterns in Electricity Price Using Clustering Techniques / F. Martínez Álvarez [at al.] // ICREPQ International Conference on Renewable Energies and Power Quality, Spain, Sevilla, 2007 [электронный ресурс]. 8 p. URL: http://www.icrepq.com/icrepq07/245-martinez.pdf (дата обращения 28.08.2011).

40. Singh S. Pattern Modelling in Time-Series Forecasting // Cybernetics and Systems-AnInternational Journal. 2000, Vol. 31, No. 1. P. 49 – 65.

41. Scherer Perlin M. Nearest neighbor method // Revista Eletrônica de Administração [электронный ресурс]. 2007, Vol. 13, No. 2. 15 p. URL: http://read.adm.ufrgs.br/edicoes/pdf/artigo_495.pdf (дата обращения 28.08.2011).

42. Fernández-Rodríguez F., Sosvilla-Rivero S., Andrada-Félix J. Nearest-Neighbour Predictions in Foreign Exchange Markets // Fundacion de Estudios de Economia Aplicada [элекронный ресурс]. 2002, No.5. 36 p. URL: http://www.fedea.es/pub/Papers/2002/dt2002-05.pdf (дата обращения 28.08.2011)

43. Трофимов А. Г., Скругин В. И. Адаптивный классификатор

многомерных нестационарных сигналов на основе анализа динамических паттернов // Наука и образование [электронный ресурс]. 2010, №8. URL: http://technomag.edu.ru/doc/151934.html (дата обращения 28.08.2011).

44. Мерков А.Б. Распознавание образов: Введение в методы статистического обучения. М.:Едиториал УРСС, 2011. 254 с.

45. Чучуева И.А. Модель экстраполяции по максимуму подобия (ЭМП) для временных рядов цен и объемов на рынке на сутки вперед ОРЭМ (Оптовом рынке электроэнергии и мощности) // Наука и образование [электронный ресурс]. 2010. № 1. URL: http://technomag.edu.ru /doc/135870.html (дата обращения 28.08.2011).

46. Иглин С.П. Математические расчеты на базе Matlab. СПб.: BHV-Санкт-Петербург, 2005. 640 с.

47. Тест «Java Micro Benchmark» // Java [электронный ресурс]. URL: http://infoscreens.org/benchmark_en.html (дата обращения 28.08.2011).

48. Mishra S. Short term load forecasting using computation intelligence methods: Thesis for the degree of Master of technology electronics and communication engineering. India, Rourkela, National Institute Of Technology, 2008. 89 p.

49. Продукт «Прогнозы» // Закрытое акционерное общество «РусПауэр». URL: http://www.ruspower.ru/products/forecast (дата обращения 28.08.2011).

50. Java Help Center [электронный ресурс] // URL: http://www.java.com/en/ download/help/index.xml (дата обращения 28.08.2011).

51. Документация по MySQL [электронный ресурс] // URL: http://www.mysql.ru/docs/ (дата обращения 28.08.2011).

52. Рыжкова Ж. В. Методические подходы к формированию стратегий генерирующий компаний на рынках энергии и мощности: Автореферат дисс. … канд. эконом. наук. Москва, 2010. 20 с.

53. Многофакторное прогнозирование потребления электроэнергии в промышленном и бытовом секторах / Т. Кирилова [и др.] // Энергорынок 2009, №11. С. 40 – 43.

54. Павлов Ю. Н., Чучуева И. А. Экстраполяция псевдослучайных процессов по максимуму подобия // Наука и образование [электронный ресурс]. 2009. №7. URL: http://technomag.edu.ru/doc/129712.html (дата обращения 28.08.2011).

55. Pavlov J. N., Chuchueva I. A. Extrapolation of pseudorandom number sequence on maximum likeness // Наука и образование [электронный ресурс]. 2009. №7. URL: http://technomag.edu.ru/en/doc/129712.html (дата обращения 28.08.2011).

56. Чучуева И. А. Модель экстраполяции временных рядов по выборке максимального подобия // Информационные технологии. 2010. №12. С. 43 – 47.

57. Чучуева И. А., Павлов Ю. Н. Сезонно-регрессионная модель прогнозирования в решении задачи прогнозирования цен РСВ (рынок на сутки вперед) // Энерго-Info. 2009. №4. С. 46 – 49.

58. BI EnergoPrice: Прогнозирование цен на электроэнергию.// Общество с ограниченной ответственностью «BIGroupLabs» [электронный ресурс]. URL: http://www.bi-grouplabs.ru/Rech/electricity/BI_EnergoPrice.html (дата обращения 28.08.2011).

59. Oliva R., Watson N. Managing Functional Biases in Organizational Forecasts: A Case Study of Consensus Forecasting in Supply Chain Planning // HBS Working Paper [электронный ресурс]. 2006, No.10. P. 7 – 24. URL: http://www.hbs.edu/research/pdf/07-024.pdf (дата обращения 28.08.2011).

60. Reinaldo C. Garcia A. GARCH Forecasting Model to Predict Day-Ahead Electricity Prices // Workshop of Applied Infrastructure, Germany, Berlin, 2003

[электронный ресурс]. 14 p. URL: http://www.wip.tu-berlin.de/typo3 /fileadmin/documents/infraday/2003/papers/Contreras-Garcia-Garcia2003-paper-Garch_Models_to_Predict_Electricity_Prices.pdf (дата обращения 28.08.2011).

61. A GARCH Forecasting Model to Predict Day-Ahead Electricity Prices / R.C. Garcia [at al.] // IEEE Transactions on Power Systems. 2005, Vol. 20, No. 2. P. 867 – 874.

62. Day-ahead electricity prices forecasting based on time series models: a comparison / R. Espinola [at al.] // 14th Power Systems Computation Conference, Spain, Sevilla, 2002, Session 15, Paper 6 [электронный ресурс]. 8 p. URL: http://www.pscc-central.org/uploads/tx_ethpublications/s15p06.pdf (дата обращения 28.08.2011).

63. Zareipour H., Bhattacharya K., Canizares C.A. Forecasting the Hourly Ontario Energy Price byMultivariate Adaptive Regression Splines // IEEE Power Engineering Society General Meeting, Canada, Montreal, 2006. 7 p.

64. Zareipour H. Price Forecasting and Optimal Operation of Wholesale Customers in a Competitive Electricity Market: Thesis for Ph.D degree. Canada, Ontario, 2006. 201 p.

65. Bunnoon P., Chalermyanont K., Limsakul C. A Computing Model of Artificial Intelligent Approaches to Mid-term Load Forecasting: a state-of-the-art-survey for the researcher // IACSIT International Journal of Engineering and Technology. 2010, No.1. P. 94 – 100.

66. Basaran Filik U.,Kurban M. A New Approach for the Short-Term Load Forecasting with Autoregressive and Artificial Neural Network Models // International Journal of Computational Intelligence Research. 2007, No.3. P. 66 – 71.

67. Taylor J.W., McSharry P.E. Short-Term Load Forecasting Methods: An Evaluation Based on European Data // IEEE Transactions on Power Systems

2008, Vol.22. P. 2213 – 2219.

68. Сидоров С.Г., Никологорская А.В. Анализ временных рядов как метод построения потребления электроэнергии // Вестник ИГЭУ. 2010, Вып. 3. С. 81 – 83.

69. Соломкин А.В. Краткосрочное прогнозирование потребления электроэнергии с помощью нейросетевых методов // Электроника и информационные технологии [электронный ресурс]. 2011, №1. 5 с. URL: http://fetmag.mrsu.ru/2009-3/pdf/Forecasting_electricity_consumption.pdf (дата обращения 28.08.2011).

70. Староверов Б.А., Изотов В.А., Мормылев М.А. Повышение точности оперативных прогнозов потребления электроэнергии с помощью нейронных сетей за счет объединения процессов классификации и аппроксимации суточных профилей // Вестник ИГЭУ. 2007, Вып. 4. С. 91 – 93.

71. Чернецов С. А., Чучуева И. А. Прогнозирование уровня глюкозы в крови больных инсулинозависимым диабетом нейронными сетями и методом экстраполяции по выборке максимального подобия // Наука и образование [электронный ресурс]. 2010. №11. URL: http://technomag.edu.ru/doc/162847.html (дата обращения 28.08.2011).

72. Конкурс «Интернет математика 2010» // Компания «Яндекс» [электронный ресурс]. URL: http://imat2010.yandex.ru (дата обращения 28.08.2011).

73. Гуда С.А., Рябов Д.С. Прогнозирование пробок на улицах по известным данным о скорости автомобилей // IV Российская летняя школа по информационному поиску: Труды четвертой российской конференции молодых ученых по информационному поиску. Воронеж, 2010. С. 52–63.

74. Чучуева И. А. Прогнозирование временных рядов при помощи модели экстраполяции по выборке максимального подобия // Наука и современность:

сборник материалов Международной научно-практической конференции. Новосибирск, 2010. С. 187 – 192.

75. Chuchueva I. The time series extrapolation model based on maximum likeness set // Математическое моделирование социальной и экономической динамики: труды III Международной конференции. М., 2010. С. 281–283.

ПРИЛОЖЕНИЕ

Таблица 17. Набор моделей № 1 из таблицы 8

День недели	**Параметр модели** M
Понедельник	144
Вторник	144
Среда	360
Четверг	168
Пятница	216
Суббота	96
Воскресенье	216

Таблица 18. Набор моделей №2 из таблицы 8

День недели	**Параметр модели** M
Понедельник	360
Вторник	384
Среда	384
Четверг	168
Пятница	264
Суббота	96
Воскресенье	312

Таблица 19. Набор моделей №3 из таблицы 8

День недели	**Параметр модели** M
Понедельник	144
Вторник	240
Среда	120
Четверг	72
Пятница	288
Суббота	336
Воскресенье	96

Таблица 20. Набор моделей №4 из таблицы 8

День недели	**Параметр модели** M
Понедельник	120
Вторник	96
Среда	168
Четверг	168
Пятница	72
Суббота	216
Воскресенье	168

Таблица 21. Набор моделей №5 из таблицы 8

День недели	**Параметр модели** M
Понедельник	288
Вторник	96
Среда	192
Четверг	72
Пятница	384
Суббота	168
Воскресенье	216

Таблица 22. Набор моделей №6 из таблицы 8

День недели	**Параметр модели** M
Понедельник	360
Вторник	360
Среда	384
Четверг	240
Пятница	168
Суббота	96
Воскресенье	48

Таблица 23. Набор моделей №7 из таблицы 8

День недели	**Параметр модели** M
Понедельник	144
Вторник	96
Среда	336
Четверг	336
Пятница	312
Суббота	72
Воскресенье	168

Таблица 24. Набор моделей №8 из таблицы 8

День недели	**Параметр модели** M
Понедельник	144
Вторник	168
Среда	312
Четверг	96
Пятница	72
Суббота	96
Воскресенье	216

Таблица 25. Набор моделей №9 из таблицы 8

День недели	**Параметр модели** M
Понедельник	360
Вторник	192
Среда	384
Четверг	240
Пятница	48
Суббота	144
Воскресенье	216

Таблица 26. Набор моделей №10 из таблицы 8

День недели	**Параметр модели** M
Понедельник	144
Вторник	120
Среда	312
Четверг	72
Пятница	72
Суббота	96
Воскресенье	48

Таблица 27. Набор моделей №11 из таблицы 8

День недели	**Параметр модели** M
Понедельник	48
Вторник	216
Среда	336
Четверг	240
Пятница	384
Суббота	168
Воскресенье	264

Таблица 28. Набор моделей №12 из таблицы 8

День недели	**Параметр модели** M
Понедельник	336
Вторник	216
Среда	264
Четверг	192
Пятница	360
Суббота	48
Воскресенье	312

Таблица 29. Набор моделей №13 из таблицы 10

День недели	**Параметр модели** M
Понедельник	180
Вторник	108
Среда	300
Четверг	180
Пятница	228
Суббота	324
Воскресенье	60

Таблица 30. Набор моделей №14 из таблицы 12

День недели	**Параметр модели** M
Понедельник	160
Вторник	348
Среда	132
Четверг	348
Пятница	156
Суббота	108
Воскресенье	228

Таблица 31. Набор моделей №15 из таблицы 12

День недели	**Параметр модели** M
Понедельник	84
Вторник	180
Среда	136
Четверг	132
Пятница	160
Суббота	136
Воскресенье	252

Таблица 32. Набор моделей №16 из таблицы 12

День недели	**Параметр модели** M
Понедельник	144
Вторник	144
Среда	312
Четверг	120
Пятница	144
Суббота	96
Воскресенье	384

Таблица 33. Набор моделей №17 из таблицы 12

День недели	**Параметр модели** M
Понедельник	96
Вторник	192
Среда	240
Четверг	72
Пятница	360
Суббота	48
Воскресенье	48

Таблица 34. Набор моделей №18 из таблицы 12

День недели	**Параметр модели** M
Понедельник	96
Вторник	144
Среда	168
Четверг	360
Пятница	96
Суббота	48
Воскресенье	192

Таблица 35. Набор моделей №19 из таблицы 12

День недели	**Параметр модели** M
Понедельник	48
Вторник	360
Среда	336
Четверг	264
Пятница	144
Суббота	48
Воскресенье	312

Таблица 36. Набор моделей №20 из таблицы 12

День недели	**Параметр модели** M
Понедельник	144
Вторник	72
Среда	360
Четверг	120
Пятница	120
Суббота	168
Воскресенье	240

Таблица 37. Набор моделей №21 из таблицы 12

День недели	**Параметр модели** M
Понедельник	72
Вторник	360
Среда	216
Четверг	264
Пятница	196
Суббота	48
Воскресенье	324

Закрытое акционерное общество «РусПауэр»

АКТ ВНЕДРЕНИЯ

результатов кандидатской диссертационной работы
Чучуевой Ирины Александровны
«Модель прогнозирования временных рядов по выборке максимального подобия»

Согласно договору 1Л от 30 мая 2011 года Чучуева Ирина Александровна несет ответственность за надежную ежедневную работу программного обеспечения для прогнозирования 19 временных рядов цен на электроэнергию и энергопотребления Оптового рынка электроэнергии и мощности Российской Федерации.

Программное обеспечение для прогнозирования временных рядов показателей энергорынка представляет собой серверное приложение, формирующие требуемые прогнозы на основании разработанной в диссертации модели экстраполяции временных рядов по выборке максимального подобия. Прогнозирование на три горизонта – сутки, неделю и месяц вперед, осуществляется в автоматическом режиме без участия эксперта.

На основании прогнозов показателей энергорынка РФ компания «РусПауэр» формирует аналитический продукт «Прогнозы» в виде специализированных отчетов, содержащих прогнозные значения (http://www.ruspower.ru/products/forecast).

ЗАО «РусПауэр» подтверждает, что приведенные в диссертационной работе И.А. Чучуевой точности прогнозирования 19 временных рядов цен на электроэнергию и энергопотребления являются оценкой точности прогнозов, составляющих аналитический продукт «Прогнозы».

Генеральный директор ЗАО «РусПауэр»
В. В. Озеров

«РусПауэр»

12.09.2011