У роботі приведено структурну схему, відповідну для опису практично будь-якої відомої на сьогоднішній день комбінованої, гібридної або декомпозиційної моделі прогнозування часових рядів. На основі даної схеми запропоновано методи ідентифікації розріджених нелінійних моделей взаємопов'язаних нестаціонарних часових рядів на основі методів «Гусениця»-SSA, швидкого ортогонального пошуку, методу групового урахування аргументів та моделей SARIMA

Ключові слова: прогнозування, структурна ідентифікація, метод «Гусениця»-SSA, метод групового урахування аргументів

В работе приведена структурная схема, подходящая для описания практически любой известной на сегодняшний день комбинированной, гибридной или декомпозиционной модели прогнозирования временных рядов. На основе данной схемы предложены методы структурной идентификации разреженных нелинейных моделей взаимосвязанных нестационарных временных рядов на основе методов «Гусеница»-SSA, быстрого ортогонального поиска, метода группового учёта аргументов и моделей SARIMA

Ключевые слова: прогнозирование, структурная идентификация, метод «Гусеница»-SSA, метод группового учёта аргументов

УДК 519.254
DOI: 10.15587/1729-4061.2015.40065

СИСТЕМНЫЙ ПОДХОД К СИНТЕЗУ МАТЕМАТИЧЕСКИХ МОДЕЛЕЙ ПРОГНОЗИРОВАНИЯ ВЗАИМОСВЯЗАННЫХ НЕСТАЦИОНАРНЫХ ВРЕМЕННЫХ РЯДОВ

В. Н. Щелкалин
Инженер
Кафедра прикладной математики
Харьковский национальный
университет радиоэлектроники
пр. Ленина, 14, г. Харьков, Украина, 61166
E-mail: vitalii.shchelkalin@gmail.com

1. Введение

Развитие методов прогнозирования определяется степенью математического описания процессов, имеющих место в различных отраслях науки и техники с учётом математических достижений, технических ограничений, качества и объёма выборки данных и ограничений на ресурсы, в том числе и временные, формирования математической модели.

Традиционные методы прогнозирования временных рядов (ВР) предназначены, как правило, для линейных и стационарных ВР и только в последние десятилетия начали активно развиваться методы прогнозирования нелинейных, но стационарных и детерминированных ВР, и линейных, но нестационарных ВР. Однако, линейное поведение свойственно только относительно простым процессам. Большинству естественным материальным процессам, как правило, присуще нелинейное нестационарное поведение. И при моделировании данных процессов используются определенные упрощения, особенно в отношении априорно устанавливаемого базиса преобразования ВР в новые, удобные для обработки и анализа метрические пространства. Необходимое условие корректного представления нелинейных и нестационарных ВР заключается в том, чтобы иметь возможность формирования адаптивного базиса, функционально зависимого от содержания самих данных [1].

Повышение точности прогнозирования ВР является важной, но зачастую сложной задачей, стоящей перед лицами, принимающими решения во многих областях науки и техники. Эффективным способом повышения точности прогнозирования может быть комбинирование нескольких моделей или использование гибридных моделей.

В настоящее время всё чаще возникает потребность не только в повышении точности моделирования, но и в создании качественно новых моделей, учитывающих нелинейность поведения реальных процессов исследования. Анализ подобных моделей намного сложнее, чем линейных, причём разработка методики и общих подходов к исследованию в настоящее время далеко от завершения. Являясь более богатым и сложным, мир нелинейных моделей представляется для современной науки более перспективным в плане открытия новых закономерностей и описания сложных явлений. Методы исследования нелинейных нестационарных моделей в настоящее время быстро прогрессируют [2].

Важным шагом в процессе идентификации моделей прогнозирования является отбор существенных переменных и их лаговых значений для того, чтобы получить самую простую модель, то есть модель, которая обеспечивает удовлетворительные прогнозы с наименьшим числом параметров. Правильный выбор этих переменных существенно влияет на точность получения результатов прогнозирования и на время, необходимое для определения модели.

Наиболее простым средством, широко применяемым для определения количества лаговых переменных при моделировании ВР, является автокорреляционная (АКФ) и частная автокорреляционная функции (ЧАКФ). Авторегрессионные модели, построенные с

использованием автокорреляционной функции, включают в себя все лаговые значения до выбранного порядка. Однако не все лаговые значения могут иметь существенное влияние на прогнозируемую величину. Поэтому полезной является процедура прореживания структуры авторегрессионных моделей.

Оптимальное прореживание – метод упрощения структуры регрессионной модели. Основная идея прореживания: элементы модели, которые оказывают малое влияние на ошибку аппроксимации ВР, можно исключить из модели без значительного ухудшения качества аппроксимации. Такой отбор переменных повышает эффективность обучения модели за счёт устранения избыточных и несущественных переменных.

Одним из методов, который может быть использован для оптимального прореживания, является метод полного перебора, в котором рассматриваются все комбинации подмножеств переменных. Этот метод гарантирует оптимальное решение, но задача обнаружения подмножества переменных имеет большие временные затраты, когда количество переменных является большим. Поэтому актуальными являются методы субоптимального прореживания, а модели с настроенными параметрами, доставляющие минимум заданному функционалу качества, называются моделями субоптимальной структуры.

Также, АКФ и ЧАКФ измеряют только степень линейной зависимости между переменными и их задержками и не отражают нелинейные отношения. Для построения адекватных нелинейных математических моделей достаточно эффективным является расширение множества переменных модели с помощью различных преобразований исходных прогнозируемого и экзогенных временных рядов. Например, путём добавления в регрессионную модель регрессоров в степенях и их комбинации. Однако это приводит к существенному повышению сложности модели. Поэтому необходимо использовать быстрые алгоритмы отбора переменных модели. Такое расширение множества переменных называется порождением переменных.

Проблема мультиколлинеарности является основной при порождении признаков и может приводить к неустойчивости оценок параметров модели и их дисперсии. Признаками наличия мультиколлинеарности являются: значительные изменения в оценках параметров при добавлении или удалении параметра модели, превышение некоторого порога абсолютным значением корреляции между переменными, близость к нулю определителя матрицы парных корреляций признаков [3]. Основными методами устранения мультиколлинеарности являются либо выбор признаков, либо введение ограничений на параметры модели [4, 5].

Современный этап развития методов прогнозирования характеризуется все более расширяющимся применением сложных математических моделей и методов. В настоящее время эффективное моделирование сложных процессов предполагает использование различных приемов декомпозиции модели. Декомпозиция позволяет реализовать общую модель как совокупность иерархически взаимосвязанных более простых моделей разного уровня иерархии. Такая структура модели позволяет повысить точность и адекватность моделирования в случае многомерных, нелинейных и нестационарных процессов, упростить и повысить устойчивость процесса идентификации [6].

Исследователи в области построения математических моделей систем, объектов, процессов стремятся к созданию универсальной, обобщённой методики решения этой задачи. Идея применения системного подхода как методологической основы постановки и решения проблем идентификации модели не является новой для научной литературы. Одним из первых её высказал в 1984 г. В. Я. Ротач: «...задача построения математической модели объекта является системной задачей, требующей для своего решения системного подхода» [7]. В системном подходе, в отличие от традиционного, анализ ведется от системы к элементам, от сложного к простому.

2. Анализ литературных данных и постановка проблемы

Одними из наиболее главных наработок в прогнозировании ВР на протяжении последнего десятилетия являются комбинации математических моделей и гибридные модели. Гибридные математические модели и методы были использованы в различных приложениях. Математические модели и методы, использующие искусственные нейронные сети (ИНС) и алгоритмы поиска наиболее значимых переменных, являются наиболее часто используемыми. Не менее популярными среди гибридных математических моделей являются модели на основе метода «Гусеница»-SSA и моделей сезонной авторегрессии – проинтегрированного скользящего среднего (SARIMA). Использование компонент разложения метода «Гусеница»-SSA является достаточно эффективным способом порождения переменных [8–10].

Среди гибридных математических моделей следует также выделить модели, объединяющие ИНС с моделями ARIMA [11–14], метод группового учёта аргументов (МГУА) с ИНС [14], МГУА с различными преобразованиями [15], МГУА с LSSVM (least squares support vector machines) [16].

Гибридные модели, предложенные в [8–10] включают в себя большое количество лаговых переменных. Это приводило к значительным временным затратам на обучение модели. Кроме того, точность модели падает из-за присутствия в модели значительного количества незначимых переменных. Поэтому актуальной задачей является разработка методов для получения моделей с разряженной структурой. Основной целью статьи является предложить метод отбора существенных переменных для моделирования временного ряда с заданной точностью.

Анализ литературы [17–22] позволяет сделать вывод, что по мере развития прогностика существенно видоизменяется, возникают новые методологические подходы, совершенствуются методы разработки прогнозов, приобретают более четко определенный вид, расширяются сферы объектов прогнозирования, уровень и эффективность использования прогнозов.

Множества переменных моделей, приведенных в [8–10] бывает недостаточно для построения модели удовлетворительного качества. В этом случае требуется расширить множество переменных с помощью

преобразований исходных переменных с целью уменьшения недоопределённости линейной модели.

В некоторых методах поиска наиболее существенных переменных модели, таких как, например, генетические алгоритмы, необходимо одновременно оценивать множество решений на каждой итерации алгоритма, что требует больших вычислительных затрат [23].

Достаточно много методов было предложено для решения задачи отбора переменных модели, такие как эволюционные алгоритмы [24, 25] и ортогональный метод наименьших квадратов [11, 26]. Эволюционные методы опираются на генетические алгоритмы при выборе адекватной модели. Алгоритм случайного поиска для отбора переменных и их задержек рассматривает множество переменных и произвольно формирует группы переменных. Данный вид поиска обычно применяется в сочетании с многослойными нейронными сетями [12], с нейронными сетями с радиально-базисными функциями (RBF) [13] или с методом опорных векторов (SVM) [27] в качестве руководства для поиска оптимального подмножества переменных. Ортогональные методы используют набор ортогональных переменных-кандидатов, уменьшающих среднеквадратическую ошибку аппроксимации модели. Также популярными среди методов отбора значимых переменных являются: методы индуктивного построения регрессионных моделей, шаговые методы, Лассо (Least absolute shrinkage and selection operator), алгоритм ступенчатой регрессии, метод наименьших углов (LARS, Least Angle Regression) и пр. [3].

Тем не менее, проблема поиска подмножества подходящих переменных нередко может стать трудноразрешимой [28]. Проблема выбора переменных как трудноразрешимая охарактеризована в [29].

3. Цели и задачи исследования

Целью проведенных исследований является:

1. Рассмотреть системный подход к построению математических моделей прогнозирования взаимосвязанных нестационарных ВР.

2. Используя системный подход, синтезировать гибридные разреженные нелинейные математические модели прогнозирования взаимосвязанных нестационарных ВР.

Для достижения поставленных целей решались следующие задачи:

1. Предложена структурная схема моделей прогнозирования взаимосвязанных нестационарных ВР, подходящая для описания практически любой, известной на сегодняшний день, комбинированной, гибридной или декомпозиционной модели прогнозирования ВР.

2. Предложены эффективные методы порождения и отбора значимых переменных моделей прогнозирования взаимосвязанных нестационарных ВР.

Основная идея отбора переменных заключается в исключении подмножества переменных, которые не только имеют незначительную, или вовсе не имеют, прогностическую информацию, но и те, которые сильно коррелируют между собой. Таким образом, задача состоит в выборе подмножества переменных с минимальной потерей или без потери точности моделирования.

3. На основе предложенной структурной схемы прогнозирования взаимосвязанных нестационарных ВР разработать гибридные разреженные нелинейные математические модели прогнозирования взаимосвязанных нестационарных ВР.

4. Гибридные математические модели и методы прогнозирования взаимосвязанных нестационарных временных рядов

Пусть имеется прогнозируемый ВР y_t, $t=\overline{1,n}$ и экзогенные ВР $x_t^{(i)}$, $t=\overline{1,n}$, $i=\overline{1,N}$. Необходимо определить прогнозную нелинейную функцию F от лаговых значений порождённых переменных (ПП) $p_t^{(i)}$, $i=\overline{1,\sum_{i=1}^{N+1} r_i}$ такую, что

$$y_t = F\left(p_{t-t_1}^{(1)}, p_{t-t_2}^{(1)}, \ldots, p_{t-t_{m_1}}^{(1)}, p_{t-t_1}^{(2)}, p_{t-t_2}^{(2)}, \ldots, p_{t-t_{m_2}}^{(2)}, \ldots, p_{t-t_1}^{\left(\sum_{i=1}^{N+1} r_i\right)}, p_{t-t_2}^{\left(\sum_{i=1}^{N+1} r_i\right)}, \ldots, p_{t-t_{m_{\sum_{i=1}^{N+1} r_i}}}^{\left(\sum_{i=1}^{N+1} r_i\right)} \right) + \varepsilon_t,$$

где $\varepsilon_t \sim N\left(0, \sigma_\varepsilon^2\right)$ – ошибки модели.

4. 1. Системный подход к синтезу математических моделей прогнозирования нестационарных взаимосвязанных временных рядов

Для того чтобы реализовать процесс прогнозирования, необходимо выявить его основные этапы и определить их содержание. При этом описание процесса прогнозирования следует формировать с учётом системного подхода, что требует построения системного описания как ВР, так и процедуры прогнозирования.

При системном исследовании описание элементов математической модели прогнозируемого случайного процесса проводится не само по себе, а лишь в связи и с учетом их места в целом. Элементы рассматриваются как относительно неделимые – только в рамках конкретной задачи и данной математической модели. Свойства прогнозной математической модели как целого определяются не только и не столько свойствами её отдельных элементов, сколько свойствами её структуры, особыми интегративными связями рассматриваемой математической модели. Сложность и многообразие элементов, связей математической модели обусловливают её иерархическое строение – упорядоченную последовательность ее различных компонентов и уровней взаимосвязи между ними. Говоря о системном подходе, имеют в виду также выработку средств соединения, синтеза в теоретическом знании отдельных представлений о сложных случайных взаимосвязанных процессах.

Рассмотрим модель сложной системы (процесса), представленную в форме триад (рис. 1) [30].

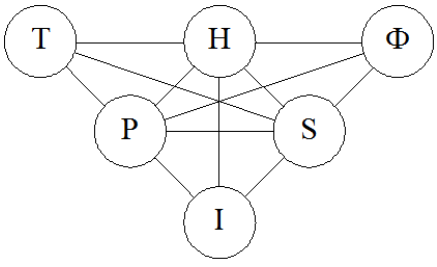

Рис. 1. Модель сложной системы (процесса) в форме триад

Опишем монады данной модели: I – интегративные свойства модели; T – цель функционирования модели; S – множество структур модели; P – множество параметров базовых элементов и связей между ними; Ф – множество базовых элементов (подсистем) модели; H – множество отношений между базовыми элементами модели. Общее число триад подлежащих рассмотрению в такой системе равно C_6^3. Остановимся подробнее на триадах «S–Ф–H» и «I–Ф–H», раскрывающих механизм образования различных структур из элементов множеств Ф и H.

В [8, 9] приведен обзор литературы, в которой предложены различные типы комбинированных, гибридных или декомпозиционных моделей прогнозирования ВР. Основные из этих моделей приведены на рис. 2.

На рис. 3 изображена структурная схема, подходящая для описания практически любой, известной на сегодняшний день, комбинированной, гибридной или декомпозиционной модели прогнозирования ВР.

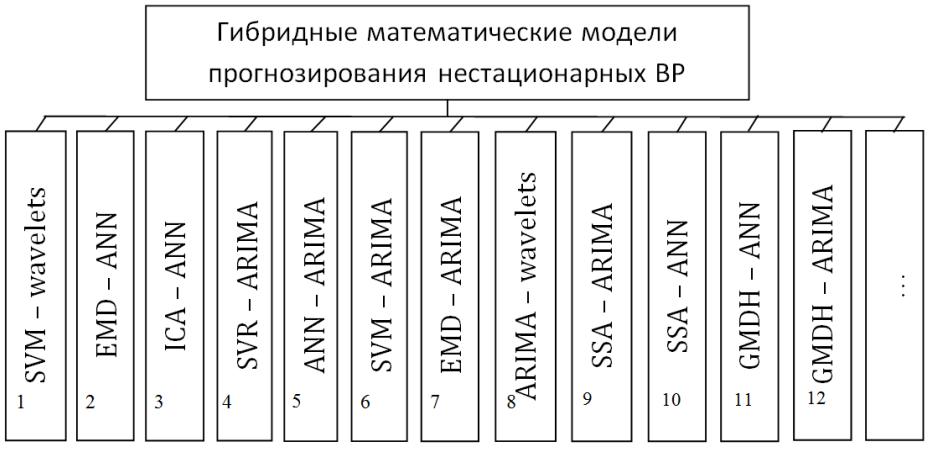

Рис. 2. Основные гибридные модели, используемые при прогнозировании нестационарных ВР

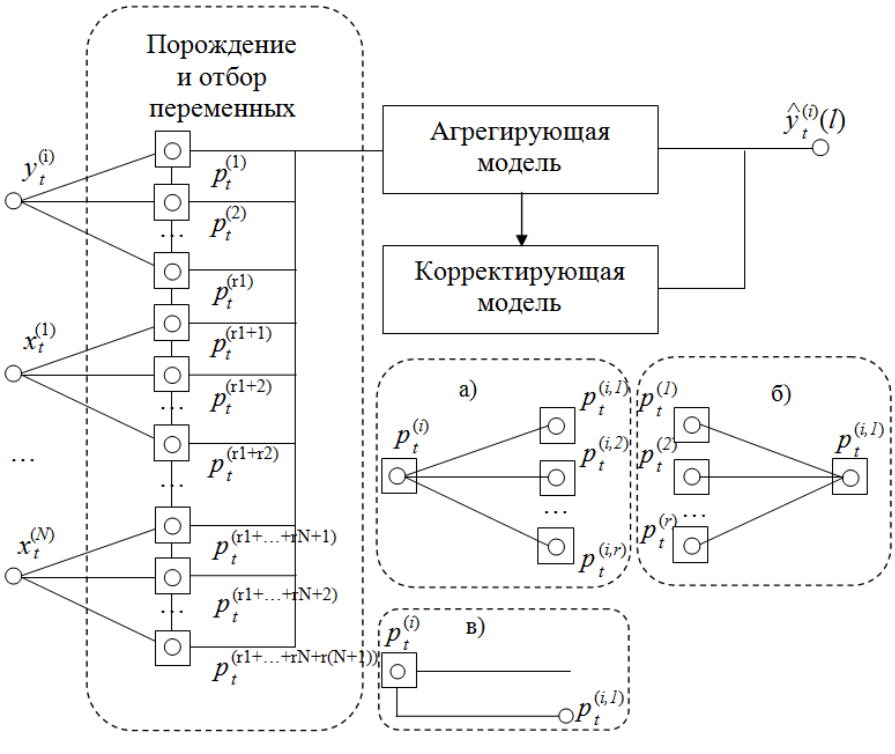

Рис. 3. Структурная схема моделей прогнозирования нестационарных взаимосвязанных временных рядов

Здесь $y_t^{(j)}$, $x_t^{(i)}$ – эндогенные и экзогенные ВР соответственно, $j=\overline{1,N^y}$; $p_t^{(i)}$ – компоненты разложения (КР), базисные функции или порождённые переменные. Четырёхугольниками обозначены математические модели.

В свою очередь КР могут быть разложены на более простые КР (EMD, SSA, SVM, разложение Фурье и пр.) (рис. 3, *а*) или, наоборот, сгруппированы в более интерпретируемые составляющие (рис. 3, *б*, SSA, GMDH и пр.). Также КР могут порождать ВР (рис. 3, *в*) (например, мгновенные амплитуды в методе Гильберта-Хуанга и пр.). Агрегирующей моделью часто выступают сумматор, искусственная нейронная сеть или полином Колмогорова-Габора и пр.; корректирующей – модель ARIMA и пр. Среди методов отбора переменных можно выделить: GA, FOS, LARS, LASSO, GMDH и пр.

Различают гомогенный состав иерархической модели, содержащей однотипные элементы (как, например, в методе МГУА, многослойном персептроне) и гетерогенный состав, элементы которой разнотипны. В общем случае состав, как правило, является смешанным. Однотипность не означает полной идентичности и определяет только близость основных свойств. Гомогенности, как правило, сопутствует избыточность и наличие скрытых, дополнительных, не использованных ресурсов или возможностей.

4. 2. Краткое описание используемых математических моделей и методов прогнозирования нестационарных взаимосвязанных временных рядов

В статье предложены два метода структурной идентификации гибридных разреженных моделей. Первый метод основан на использовании методов «Гусеница»-SSA, быстрого ортогонального поиска и модели SARIMA. Второй – на использовании метода «Гусеница»-SSA, МГУА и модели SARIMA.

Рассмотрим структурную схему моделей прогнозирования (рис. 3). Для первой и второй из предлагаемых моделей в качестве порождающих переменных выбираем КР метода «Гусеница»-SSA. В качестве метода прореживания переменных для первого метода выбираем метод FOS, для второго – МГУА. В качестве агрегирующей модели в первом методе выбираем обычный сумматор, а во втором – редуцированный методом МГУА полином Колмогорова-Габора. В качестве корректирующей модели для обоих случаев выбираем модель SARIMA, построенную на остаточных ошибках агрегирующей модели.

Поэтому вначале рассмотрим модели и методы, составляющие предлагаемые гибридные.

4. 2. 1. Многомерный вариант метода «Гусеница»-SSA

Пусть требуется получить прогноз N+1 взаимосвязанных нестационарных ВР $Y=y_1, y_2, ..., y_{n_y}$ и $X^{(i)}=x_1^{(i)}, x_2^{(i)}, ..., x_{n_{x^{(i)}}}^{(i)}$, $i=\overline{1,N}$ произвольных длин n_y и $n_{x^{(i)}}$, $i=\overline{1,N}$ соответственно. Алгоритм анализа временных рядов многомерным вариантом метода «Гусеница»-SSA состоит из следующих этапов.

1. Вложение.

Выбираем длину окна L и строим траекторную матрицу

$$\mathbf{X}=\left[Y_1 \;\; Y_2 \;\; \ldots \;\; Y_{K_y} \;\; X_1^{(1)} \;\; X_2^{(1)} \;\; \ldots \;\; X_{K_{x^{(1)}}}^{(1)} \;\; X_1^{(2)} \;\; X_2^{(2)} \;\; \ldots \;\; X_{K_{x^{(2)}}}^{(2)} \;\; \ldots \;\; X_1^{(N)} \;\; X_2^{(N)} \;\; \ldots \;\; X_{K_{x^{(N)}}}^{(N)}\right]=\left[\mathbf{Y} \;\; \mathbf{X}^{(1)} \;\; \mathbf{X}^{(2)} \;\; \ldots \;\; \mathbf{X}^{(N)}\right]$$

из векторов вложения $X_j^{(i)}=\left(x_j^{(i)} \;\; x_{j-1}^{(i)} \;\; \ldots \;\; x_{j+L-1}^{(i)}\right)^T$, $1\le j\le K_{x^{(i)}}$, $K_{x^{(i)}}=n_{x^{(i)}}-L-1$, $i=\overline{1,N}$. Здесь $\mathbf{Y}$ – траекторная матрица ряда Y, $\mathbf{X}^{(i)}$, $i=\overline{1,N}$ – траекторная матрица ряда $X^{(i)}$, $i=\overline{1,N}$.

2. Сингулярное разложение.

Сформируем матрицу $S=\mathbf{X}\mathbf{X}^T$ и произведём сингулярное разложение траекторной матрицы $\mathbf{X}$ ВР.

Обозначим:

– $\lambda_1, \lambda_2, ..., \lambda_L$ – собственные числа матрицы S, взятые в порядке убывания ($\lambda_1\ge\lambda_2\ge...\ge\lambda_L\ge0$);

– $U_1, U_2, ..., U_L$ – ортонормированная система собственных векторов матрицы *S*, соответствующих этим собственным числам.

Произведём разложение траекторной матрицы

$$\mathbf{X}=\mathbf{X}_1+\ldots+\mathbf{X}_d, \qquad (1)$$

где

$$\mathbf{X}_j=\sqrt{\lambda_j}U_jV_j^T;\; V_j=\frac{1}{\sqrt{\lambda_j}}\mathbf{X}^TU_j;\; j=1, 2, ..., d;\; d=\max\left\{i\,|\,\lambda_i>0\right\}.$$

3. Группировка.

Разложение (1) в сгруппированном виде может быть записано следующим образом:

$$\mathbf{X}=\mathbf{X}_{I_1}+\mathbf{X}_{I_2}+\ldots+\mathbf{X}_{I_r}, \qquad (2)$$

где $\mathbf{X}_{I_j}=\mathbf{X}_{i_{j1}}+\mathbf{X}_{i_{j2}}+\ldots+\mathbf{X}_{i_{jp_j}}$; $I_j=\left\{i_{j1}, i_{j2}, ..., i_{jp_j}\right\}$, $j=\overline{1,r}$; I_j – непересекающиеся подмножества множества индексов {1, 2, ..., d}.

4. Диагональное усреднение.

Матрицы $\mathbf{X}_{I_j}$, $j=\overline{1,r}$ сгруппированного разложения переводятся в систему новых рядов длины n. Для этого они разбиваются следующим образом $\mathbf{X}_{I_j}=\left[\mathbf{Y}_{I_j} \;\; \mathbf{X}_{I_j}^{(1)} \;\; \mathbf{X}_{I_j}^{(2)} \;\; \ldots \;\; \mathbf{X}_{I_j}^{(N)}\right]$. Далее производится диагональное усреднение каждой из матриц $\mathbf{Y}_{I_j}$ и $\mathbf{X}_{I_j}^{(k)}$, k=1, 2, ..., N, $j=\overline{1,r}$, преобразуя их в ВР $\tilde{Y}_{I_j}$ и $\tilde{X}_{I_j}^{(k)}$, k=1, 2, ..., N, $j=\overline{1,r}$ соответственно. В результате каждая матрица $\mathbf{X}_{I_j}$ порождает многомерный ВР $\left(\tilde{Y}_{I_j} \;\; \tilde{X}_{I_j}^{(1)} \;\; \tilde{X}_{I_j}^{(2)} \;\; \ldots \;\; \tilde{X}_{I_j}^{(N)}\right)$, $j=\overline{1,r}$ – восстановленную аддитивную компоненту исходного ряда $\left(Y \;\; X^{(1)} \;\; X^{(2)} \;\; \ldots \;\; X^{(N)}\right)$. Переобозначим ВР $\tilde{Y}_{I_j}$ и $\tilde{X}_{I_j}^{(k)}$, k=1, 2, ..., N, $j=\overline{1,r}$ в ВР $\tilde{y}_t^{(j)}$, $\tilde{x}_t^{(i,j)}$, $i=\overline{1,N}$, $j=\overline{1,r}$.

4. 2. 2. Метод отбора переменных

Важным вопросом построения модели прогнозирования взаимосвязанных нестационарных ВР является определение экзогенных переменных и их лаговых значений, т. е. передаточной функции модели. При вы-

боре наилучшего подмножества регрессоров имеются два противоположных по характеру критерия. С одной стороны, для получения надёжных прогнозов ВР, в модель нужно включать как можно больше регрессоров. С другой стороны, с увеличением числа регрессоров возрастает дисперсия прогноза и увеличиваются затраты, связанные с получением информации о дополнительных регрессорах, поэтому желательно включать в уравнение как можно меньше регрессоров.

4. 2. 2. 1. Алгоритм быстрого ортогонального поиска

Среди алгоритмов, предложенных выше, для поиска вида передаточной функции модели предлагается использовать алгоритм быстрого ортогонального поиска, как одного из наиболее эффективных и быстрых. Метод быстрого ортогонального поиска (FOS, Fast Orthogonal Search) был предложен М. Коренбергом [31] для определения адекватной модели и её параметров. В этом алгоритме для определения существенных переменных используются следующие соображения.

Пусть требуется отобрать существенные переменные модели вида

$$\bar{y} = P^T \cdot \bar{g},$$

где $\bar{y} = (y_1, y_2, \ldots y_n)^T$, $\bar{g} = (g_1, g_2, \ldots g_n)^T$,

$$P = \begin{pmatrix} p_1^{(1)} & p_1^{(2)} & \vdots & p_1^{(M)} \\ p_2^{(1)} & p_2^{(2)} & \vdots & p_2^{(M)} \\ \ldots & \ldots & \ddots & \ldots \\ p_n^{(1)} & p_n^{(2)} & \vdots & p_n^{(M)} \end{pmatrix},$$

$\bar{p}^{(j)}$ – вектор представляющий *j*-ю переменную.

Пусть V^TV – разложение Холецкого матрицы P^TP. При добавлении переменной-кандидата в матрицу P разложение Холецкого примет вид:

$$\begin{pmatrix} P^T \\ (\bar{p}^{(j)})^T \end{pmatrix} \cdot \begin{pmatrix} P & \bar{p}^{(j)} \end{pmatrix} = \begin{pmatrix} V^T & 0 \\ (\bar{v}^{(j)})^T & x^{(j)} \end{pmatrix} \cdot \begin{pmatrix} V & \bar{v}^{(j)} \\ 0 & x^{(j)} \end{pmatrix}, \quad (3)$$

где $(\bar{v}^{(j)} \quad x^{(j)})^T$ – новый столбец матрицы V, а для модифицированной матрицы P система нормальных уравнений $P^T \cdot \bar{y} = P^T \cdot P \cdot \bar{g}$ будет иметь вид

$$\begin{pmatrix} P^T \\ (\bar{p}^{(j)})^T \end{pmatrix} \cdot \begin{pmatrix} P & \bar{p}^{(j)} \end{pmatrix} \cdot \bar{g}^{(j)} = \begin{pmatrix} P^T \\ (\bar{p}^{(j)})^T \end{pmatrix} \cdot \bar{y}.$$

Из выражения (3) следует

$$\begin{pmatrix} V^T & 0 \\ (\bar{v}^{(j)})^T & x^{(j)} \end{pmatrix} \cdot \begin{pmatrix} V & \bar{v}^{(j)} \\ 0 & x^{(j)} \end{pmatrix} \cdot \bar{g}^{(j)} = \begin{pmatrix} P^T \cdot \bar{y} \\ (\bar{p}^{(j)})^T \cdot \bar{y} \end{pmatrix}, \quad (4)$$

а из (4) – $\bar{w}^{(j)} = \begin{pmatrix} \bar{w}^{(0)} \\ \frac{1}{x^{(j)}} \cdot \left((\bar{p}^{(j)})^T \cdot \bar{y} - (\bar{v}^{(j)})^T \cdot \bar{w}^{(0)} \right) \end{pmatrix}$,

где $\bar{w}^{(j)} = \begin{pmatrix} V & \bar{v}^{(j)} \\ 0 & x^{(j)} \end{pmatrix} \cdot \bar{g}^{(j)}$, $\bar{w}^{(0)} = (V^T)^{-1} \cdot P^T \cdot \bar{y}$ – решение для $\bar{w}$, полученное на предыдущем шаге и одинаковое для всех кандидатов $\bar{p}^{(j)}$. Таким образом, добавление разных переменных-кандидатов приводит к изменению только последнего элемента вектора $\bar{w}$.

Обозначив $\alpha^{(j)} = (\bar{p}^{(j)})^T \cdot \bar{y} - (\bar{v}^{(j)})^T \cdot \bar{w}^{(0)}$, оценки модели могут быть получены из выражения:

$$\begin{pmatrix} V & \bar{v}^{(j)} \\ 0 & x^{(j)} \end{pmatrix} \cdot \bar{g}^{(j)} = \begin{pmatrix} \bar{w}^{(0)} \\ \frac{\alpha^{(j)}}{x^{(j)}} \end{pmatrix}.$$

Можно определить изменения в весовых коэффициентах $\bar{g}^{(j)}$, $\Delta\bar{g}^{(j)}$, используя выражение $\bar{g}^{(j)} = \begin{pmatrix} \bar{g}^{(0)} \\ 0 \end{pmatrix} + \Delta\bar{g}^{(j)}$, где $\bar{g}^{(0)}$ – решение $P^T \cdot P \cdot \bar{g}^{(0)} = P^T \cdot \bar{w}^{(0)}$ до изменения матрицы P $\left(V \cdot \bar{g}^{(0)} = \bar{w}^{(0)} \right)$. Тогда

$$\begin{pmatrix} V & \bar{v}^{(j)} \\ 0 & x^{(j)} \end{pmatrix} \cdot \Delta\bar{g}^{(j)} = \begin{pmatrix} 0 \\ \frac{\alpha^{(j)}}{x^{(j)}} \end{pmatrix} \quad (5)$$

и новые оценки $\bar{y}$:

$$\hat{\bar{y}}^{(j)} = \begin{pmatrix} P & \bar{p}^{(j)} \end{pmatrix} \cdot \bar{g}^{(j)} = \hat{\bar{y}}^{(0)} + \begin{pmatrix} P & \bar{p}^{(j)} \end{pmatrix} \cdot \Delta g^{(j)},$$

где $\hat{\bar{y}}^{(0)} = P \cdot \bar{g}^{(0)}$ – оценка, полученная на предыдущем шаге, а $\Delta\bar{y}^{(j)} = \begin{pmatrix} P & \bar{p}^{(j)} \end{pmatrix} \cdot \Delta\bar{g}^{(j)}$. Наилучшая переменная из множества $\bar{p}^{(j)}$, $j = \overline{1,m}$ – признак, который дает наибольшее приращение $\Delta\bar{y}$. Таким образом,

$$(\Delta\bar{y})^2 = (\Delta\bar{y}^{(j)})^T \Delta\bar{y}^{(j)} = (\Delta\bar{g}^{(j)})^T \cdot \begin{pmatrix} P^T \\ (\bar{p}^{(j)})^T \end{pmatrix} \cdot \begin{pmatrix} P & \bar{p}^{(j)} \end{pmatrix} \cdot \Delta\bar{g}^{(j)}.$$

Подставляя преобразованное разложение Холецкого (3):

$$(\Delta\bar{y}^{(j)})^2 = (\Delta\bar{g}^{(j)})^T \cdot \begin{pmatrix} V^T & 0 \\ (\bar{v}^{(j)})^T & x^{(j)} \end{pmatrix} \cdot \begin{pmatrix} V & \bar{v}^{(j)} \\ 0 & x^{(j)} \end{pmatrix} \cdot \Delta\bar{g}^{(j)} a.$$

Применив выражение (5), получим

$$(\Delta\bar{y}^{(j)})^2 = \left(\begin{pmatrix} V & \bar{v}^{(j)} \\ 0 & x^{(j)} \end{pmatrix} \cdot \Delta\bar{g}^{(j)} \right)^2 = \begin{pmatrix} 0 \\ \frac{\alpha^{(j)}}{x^{(j)}} \end{pmatrix}^2 = \frac{(\alpha^{(j)})^2}{(x^{(j)})^2}.$$

Таким образом,

$$(\Delta\bar{y}^{(j)})^2 = \frac{(\alpha^{(j)})^2}{(x^{(j)})^2}.$$

По этому выражению можно судить о том, насколько сильно добавление j-й переменной влияет на моделируемый временной ряд, а, следовательно, позволяет отобрать существенные переменные из множества кандидатов.

4. 2. 3. Модель авторегрессии – проинтегрированного скользящего среднего

Сезонная модель авторегрессии – проинтегрированного скользящего среднего (SARIMA) в операторной форме может быть представлена в следующем виде [32]:

$$y_t = \frac{\theta^*_{q^*}(B)}{\Phi^+_{p^+}(B)} a_t . \qquad (6)$$

где y_t, $t = \overline{1,n}$ – исходный или преобразованный (нормированный или прологарифмированный) центрированный ВР; n – объём выборки; B – оператор сдвига по времени на одну единицу назад, такой что $B^i x_k = x_{k-i}$; $\Phi^+_{p^+}(B)$ – обобщенный оператор авторегрессии порядка

$$p^+ = p^* + \sum_{i=1}^{n_s} D_i S_i,\ p^* = \sum_{i=1}^{n_S} p_i S_i;$$

$$\Phi^+_{p^+}(B) = \Phi^*_{p^*}(B) \nabla^{D_1}_{S_1} \nabla^{D_2}_{S_2} \dots \nabla^{D_{n_S}}_{S_{n_S}};$$

D_i, $i = \overline{1,n_S}$ – порядок взятия разности S_i; S_i, $i = \overline{1,n_S}$ – период i-й периодической компоненты, причем $S_1 = 1$; n_S – количество периодических компонент; ∇_{S_i} и B^{S_i} – упрощающие операторы такие, что

$$\nabla_{S_i} x_t = \left(1 - B^{S_i}\right) x_t = x_t - x_{t-S_i}.$$

$\Phi^*_{p^*}(B)$ – обобщенный оператор авторегрессии порядка p^* вида

$$\Phi^*_{p^*}(B) = \prod_{i=1}^{n_S} \Phi^i_{p_i}\left(B^{S_i}\right);\ \Phi^i_{p_i}\left(B^{S_i}\right),\ i = \overline{1,n_S} -$$

полиномы от B^{S_i} степеней p_i соответственно, определяющие составляющие авторегрессии периодических компонент с периодами S_i соответственно; $\theta^*_{q^*}(B)$ – обобщенный оператор скользящего среднего порядка

$$q^* = \sum_{i=1}^{n_S} q_i S_i$$

вида

$$\theta^*_{q^*}(B) = \prod_{i=0}^{n_S} \theta^i_{q_i}\left(B^{S_i}\right);\ \theta^i_{q_i}\left(B^{S_i}\right),\ i = \overline{1,n_S} -$$

полиномы от B^{S_i} степеней q_i соответственно, определяющие составляющие скользящего среднего периодических компонент с периодами S_i соответственно; a_t – случайный процесс типа белый шум.

4. 2. 4. Метод группового учета аргументов

В наиболее общем виде функцию, аппроксимирующую зависимость одного ВР от N других можно представить следующим образом:

$$y_t = F\left(x_t^{(1)}, x_t^{(2)}, \dots, x_t^{(N)}\right).$$

В качестве такого аппроксиматора часто выступает полином Колмогорова-Габора [6]:

$$y_t = a_0 + \sum_{i=1}^{N} a_i \cdot x_t^{(i)} + \sum_{i=1}^{N}\sum_{j=1}^{N} a_{ij} \cdot x_t^{(i)} \cdot x_t^{(j)} + \\ + \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N} a_{ijk} \cdot x_t^{(i)} \cdot x_t^{(j)} \cdot x_t^{(k)} + \dots \qquad (7)$$

В МГУА эта сложная зависимость заменяется множеством простых функций, так называемых частичных описаний (ЧО):

$$y_t^{(1,1)} = f\left(x_t^{(1)}, x_t^{(2)}\right);\ y_t^{(2,1)} = f\left(x_t^{(1)}, x_t^{(3)}\right); \dots; \\ y_t^{(C,1)} = f\left(x_t^{(M-1)}, x_t^{(M)}\right), \qquad (8)$$

где $C = C_N^2$, причем функция f всюду одинакова.

Часто в качестве функции f выбираются простые зависимости вида:

$$y_t\left(x_t^{(i)}, x_t^{(j)}\right) = a_0 + a_1 \cdot x_t^{(i)} + a_2 \cdot x_t^{(j)} + a_3 \cdot x_t^{(i)} \cdot x_t^{(j)} \qquad (9)$$

или

$$y\left(x_t^{(i)}, x_t^{(j)}\right) = a_0 + a_1 \cdot x_t^{(i)} + a_2 \cdot x_t^{(j)} + \\ + a_3 \cdot x_t^{(i)} \cdot x_t^{(j)} + a_4 \cdot \left(x_t^{(i)}\right)^2 + a_5 \cdot \left(x_t^{(j)}\right)^2, \qquad (10)$$

связывающие только две переменные.

Модели (8) составляют первый ряд метода, из которых выбираются R наилучших моделей по комбинированному критерию эффективности и внешнего дополнения. На втором этапе алгоритма полученные и отобранные на обучающей выборке значения $y_t^{(i,1)}$ перенумеровываются по порядку и рассматриваются в качестве аргументов второго ряда:

$$y_t^{(1,2)} = f\left(y_t^{(1,1)}, y_t^{(2,1)}\right);\ y_t^{(2,2)} = f\left(y_t^{(1,1)}, y_t^{(3,1)}\right); \dots;$$

$$y_t^{\left(C_R^2,2\right)} = f\left(y_t^{(R-1,1)}, y_t^{(R,1)}\right).$$

Коэффициенты данных моделей находятся, используя данные той же обучающей последовательности, которая использовалась и на первом этапе алгоритма. Алгоритм построения рядов продолжается до тех пор, пока будет уменьшаться минимальная ошибка комбинированного критерия эффективности и внешнего дополнения наилучшей модели каждого следующего ряда или пока сложность модели не превысит информативные возможности обучающей выборки.

4. 3. Структурная идентификации гибридных нелинейных разреженных математических моделей временных рядов

В моделях, рассмотренных в [8, 9] использовались все лаговые переменные определённого порядка как значимые. Таким образом, это приводило к существенным временным затратам при идентификации таких моделей. Кроме того, оценивание большого количества

несущественных лаговых переменных приводило к неточностям в оценивании переменных.

В статье предложен метод идентификации передаточных функций моделей, представленных в [8, 9], использующий алгоритм быстрого ортогонального поиска, который отбирает лаговые переменные прогнозируемого и экзогенных временных рядов таким образом, что первыми отбираются переменные наиболее коррелируемые с прогнозируемой величиной.

Пусть задано множество прогнозируемой и экзогенных переменных

$$X=\left[y_t \quad x_t^{(1)} \quad x_t^{(2)} \quad \dots \quad x_t^{(N)}\right].$$

На этапе группировки методом «Гусеница»-SSA данное множество переменных преобразуется в сумму компонент разложения:

$$y_t=\sum_{i=1}^{r}\tilde{y}_t^{(i)},\ x_t^{(j)}=\sum_{i=1}^{r}\tilde{x}_t^{(j,i)},\ j=\overline{1,N}.$$

4. 3. 1. Гибридная разреженная модель на основе методов «Гусеница»-SSA, быстрого ортогонального поиска и модели SARIMA

$$X'=\left[\tilde{y}_t^{(1)} \quad \tilde{y}_t^{(2)} \quad \dots \quad \tilde{y}_t^{(r)} \quad \tilde{x}_t^{(1,1)} \quad \tilde{x}_t^{(1,2)} \quad \dots \quad \tilde{x}_t^{(1,r)} \quad \tilde{x}_t^{(2,1)} \dots \right.$$
$$\left. \dots \quad \tilde{x}_t^{(2,2)} \dots \quad \tilde{x}_t^{(2,r)} \quad \dots \quad \tilde{x}_t^{(N,1)} \quad \tilde{x}_t^{(N,2)} \quad \dots \quad \tilde{x}_t^{(N,r)}\right].$$

От множества X переходим к следующему множеству переменных:

Также вводим обозначения

$$w_t^{(i)}=\tilde{y}_t^{(i)},\ i=\overline{1,r},\ w_t^{(r+(j-1)r+i)}=\tilde{x}_t^{(j,i)},\ i=\overline{1,r},\ j=\overline{1,N}.$$

Выберем величины максимальных задержек по времени m_i для i-й КР соответственно, $i=\overline{1,(N+1)r}$. Тогда получим множество

$$X''=\left[w_t^{(1)} \quad w_{t-1}^{(1)} \quad \dots \quad w_{t-m_1}^{(1)} \quad w_t^{(2)} \quad w_{t-1}^{(2)} \quad \dots \quad w_{t-m_2}^{(2)} \quad \dots \right.$$
$$\left. \dots \quad w_t^{((N+1)r)} \quad w_{t-1}^{((N+1)r)} \quad \dots \quad w_{t-m_{((N+1)r)}}^{((N+1)r)}\right].$$

Матрицу X'' будем использовать в алгоритме FOS в качестве матрицы P.

В качестве модели, описывающей отношение между прогнозируемым ВР и порождёнными ВР $w_i^{(k)}$ используем модель регрессии, построенную на компонентах разложения метода "Гусеница"-SSA $w_t^{(i)}$, $i=\overline{1,(N+1)r}$ и их задержках $w_{t-j}^{(i)}$, $i=\overline{1,(N+1)r}$, $j=\overline{1,m_i}$:

$$y_t=\alpha_0+\sum_{i=1}^{(N+1)r}\sum_{j=0}^{m_i}\alpha_{ij}w_{t-j}^{(i)}+\varepsilon_t.$$

Обозначим через FOS(•) – оператор, прореживающий, подаваемую в качестве его аргумента, регрессию путём применения алгоритма FOS к переменным данной регрессии. Тогда предлагаемая модель примет вид:

$$y_t=\mathrm{FOS}\left(\alpha_0+\sum_{i=1}^{(N+1)r}\sum_{j=0}^{m_i}\alpha_{ij}w_{t-j}^{(i)}\right)+\varepsilon_t.$$

В качестве корректирующей модели выберем модель SARIMA. Тогда окончательный вид модели будет следующий:

$$y_t=\mathrm{FOS}\left(\alpha_0+\sum_{i=1}^{(N+1)r}\sum_{j=0}^{m_i}\alpha_{ij}w_{t-j}^{(i)}\right)+\frac{\theta_{q^*}^{*}(B)}{\Phi_{p^+}^{+}(B)}a_t. \qquad (11)$$

Прогнозные значения компонент $\hat{w}_t^{(j)}$ предлагается получать при помощи моделей, предложенных в [9].

4. 3. 1. 1. Метод структурной идентификации гибридной разреженной модели на основе методов «Гусеница»-SSA, быстрого ортогонального поиска и модели SARIMA

Перед выполнением алгоритма переменные необходимо нормировать.

1. Формируем множество прогнозируемого и экзогенных ВР:

$$X=\left[y_t \quad x_t^{(1)} \quad x_t^{(2)} \quad \dots \quad x_t^{(N)}\right].$$

Применяя метод «Гусеница»-SSA к данным ВР, переходим к множеству порождённых переменных

$$X''=\left[w_t^{(1)} \quad w_{t-1}^{(1)} \quad \dots \quad w_{t-m_1}^{(1)} \quad w_t^{(2)} \quad w_{t-1}^{(2)} \quad \dots \quad w_{t-m_2}^{(2)} \quad \dots \right.$$
$$\left. \dots \quad w_t^{((N+1)r)} \quad w_{t-1}^{((N+1)r)} \quad \dots \quad w_{t-m_{((N+1)r)}}^{((N+1)r)}\right].$$

1. Методом FOS определяются наиболее значимые переменные из множества X'' и определяются коэффициенты модели

$$y_t=\mathrm{FOS}\left(\alpha_0+\sum_{i=1}^{(N+1)r}\sum_{j=0}^{m_i}\alpha_{ij}w_{t-j}^{(i)}\right). \qquad (12)$$

2. Определяются остаточные ошибки модели

$$\varepsilon_t=y_t-\mathrm{FOS}\left(\alpha_0+\sum_{i=1}^{(N+1)r}\sum_{j=0}^{m_i}\alpha_{ij}w_{t-j}^{(i)}\right)$$

и строится для них модель SARIMA:

$$\varepsilon_t=\frac{\theta_{q^*}^{*}(B)}{\Phi_{p^+}^{+}(B)}a_t.$$

3. Строится модель (11).

4. Проверяется адекватность модели и вычисляются прогнозы, предварительно перейдя к разностной форме записи модели.

Для модели (11) дадим обозначение «Гусеница»-SSA – FOS – SARIMA.

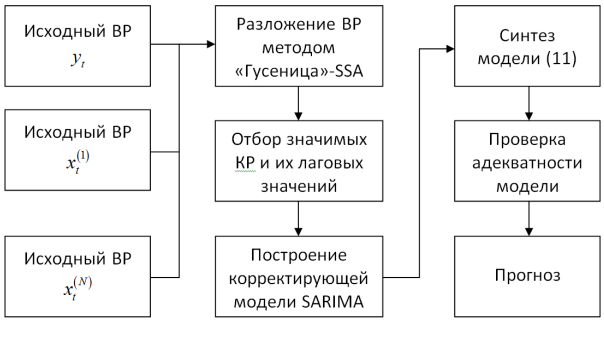

Рис. 4. Схема алгоритма структурной идентификации гибридной разреженной модели на основе методов «Гусеница»-SSA, FOS и модели SARIMA

4. 3. 2. Гибридная нелинейная разреженная модель на основе методов «Гусеница»-SSA, МГУА и модели SARIMA

Как было сказано выше, в качестве эффективного аппроксиматора выступает полином Колмогорова-Габора:

$$y_t = a_0 + \sum_{i=1}^{(N+1)r}\sum_{j=0}^{m_i} a_{ij} w_{t-j}^{(i)} + \sum_{i=1}^{(N+1)r}\sum_{j=1}^{(N+1)r}\sum_{k=0}^{m_i}\sum_{h=0}^{m_j} a_{ijkh} w_{t-k}^{(i)} w_{t-h}^{(j)} + \ldots + \varepsilon_t, \quad (13)$$

где (13) – полином Колмогорова-Габора, построенный на компонентах разложения метода "Гусеница"-SSA $w_t^{(i)}$, $i = \overline{1,(N+1)r}$ и их задержках $w_{t-j}^{(i)}$, $i = \overline{1,(N+1)r}$, $j = \overline{1,m_i}$; $a = (a_0, a_{10}, \ldots, a_{ij}, \ldots, a_{ijkh}, \ldots)$ – вектор коэффициентов модели, i, j, ... = 1, 2, ...

Поэтому в данном разделе в качестве порождённых переменных предлагается использовать не только компоненты разложения метода «Гусеница»-SSA и их лаговые значения, но и их степени и сочетания.

Запишем полином (13) в виде линейной комбинации порождённых переменных,

$$y_t = \sum_{i\in I} a_i p_t^{(i)}, \ |I| = M, \quad (14)$$

где индекс i – номер члена линейной комбинации, $p_t^{(i)}$ – мономы полинома Колмогорова-Габора. Переменные $p_t^{(i)}$ поставлены в однозначное соответствие мономам полинома (13).

В работе предложен метод идентификации разреженных нелинейных математических моделей ВР, основанных на полиномиальных функциях высоких порядков, способных выполнять сложные нелинейные отображения. Точность моделирования требует большого количества базисных функций их задержек, степеней и сочетаний. Поэтому алгоритм FOS для редукции линейной комбинации (14) является непригодным при большом количестве компонент разложения. В этом случае для редуцирования полинома (13) целесообразнее использовать МГУА.

Обозначим через GMDH(•) – оператор прореживающий, подаваемый в качестве его аргумента, полином Колмогорова-Габора путём применения МГУА к переменным данного полинома. Тогда предлагаемая гибридная прореженная нелинейная модель на основе методов «Гусеница»-SSA, МГУА и модели SARIMA примет следующий вид:

$$y_t = \mathrm{GMDH}\left(\alpha_0 + \sum_{i=1}^{(N+1)r}\sum_{j=0}^{m_i}\alpha_{ij} w_{t-j}^{(i)} + \sum_{i=1}^{(N+1)r}\sum_{j=1}^{(N+1)r}\sum_{k=0}^{m_i}\sum_{h=0}^{m_j}\alpha_{ijkh} w_{t-k}^{(i)} w_{t-h}^{(j)} + \ldots\right) + \frac{\theta_{q^*}^*(B)}{\Phi_{p^+}^+(B)} a_t. \quad (15)$$

Прогнозные значения компонент $\hat{w}_t^{(j)}$ предлагается получать при помощи моделей, предложенных в [9].

4. 3. 2. 1. Метод структурной идентификации гибридной разреженной нелинейной модели на основе методов «Гусеница»-SSA, МГУА и модели SARIMA

Перед выполнением алгоритма переменные необходимо нормировать.

1. Формируем множество прогнозируемого и экзогенных ВР:

$$X = \left[y_t \quad x_t^{(1)} \quad x_t^{(2)} \quad \ldots \quad x_t^{(N)}\right].$$

Применяя метод «Гусеница»-SSA к данным ВР, переходим к множеству признаков

$$X'' = \left[w_t^{(1)} \quad w_{t-1}^{(1)} \quad \ldots \quad w_{t-m_1}^{(1)} \quad w_t^{(2)} \quad w_{t-1}^{(2)} \quad \ldots \quad w_{t-m_2}^{(2)} \quad \ldots \quad \ldots \quad w_t^{((N+1)r)} \quad w_{t-1}^{((N+1)r)} \quad \ldots \quad w_{t-m_{((N+1)r)}}^{((N+1)r)}\right].$$

2. Методом *GMDH* формируется нелинейная модель

$$y_t = \mathrm{GMDH}\left(\alpha_0 + \sum_{i=1}^{(N+1)r}\sum_{j=0}^{m_i}\alpha_{ij} w_{t-j}^{(i)} + \sum_{i=1}^{(N+1)r}\sum_{j=1}^{(N+1)r}\sum_{k=0}^{m_i}\sum_{h=0}^{m_j}\alpha_{ijkh} w_{t-k}^{(i)} w_{t-h}^{(j)} + \ldots\right). \quad (16)$$

3. Определяются остаточные ошибки модели

$$\varepsilon_t = y_t - \mathrm{GMDH}\left(\alpha_0 + \sum_{i=1}^{(N+1)r}\sum_{j=0}^{m_i}\alpha_{ij} w_{t-j}^{(i)} + \sum_{i=1}^{(N+1)r}\sum_{j=1}^{(N+1)r}\sum_{k=0}^{m_i}\sum_{h=0}^{m_j}\alpha_{ijkh} w_{t-k}^{(i)} w_{t-h}^{(j)} + \ldots\right)$$

и строится для них модель SARIMA:

$$\varepsilon_t = \frac{\theta_{q^*}^*(B)}{\Phi_{p^+}^+(B)} a_t.$$

4. Строится модель (15).

5. Проверяется адекватность синтезированной модели и вычисляются прогнозы, предварительно перейдя к разностной форме записи модели.

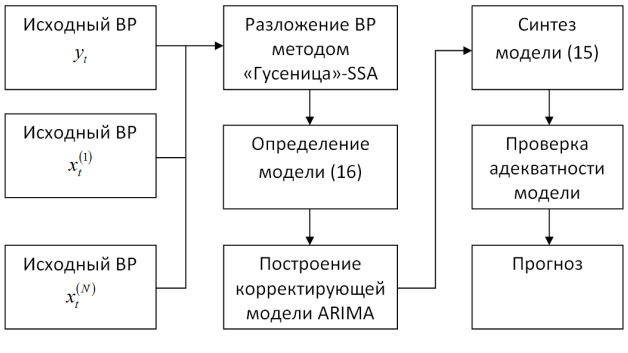

Рис. 5. Схема алгоритма структурной идентификации гибридной разреженной нелинейной модели на основе методов «Гусеница»-SSA, МГУА и модели SARIMA

Для модели (15) дадим обозначение «Гусеница»-SSA – МГУА – SARIMA.

5. Способы обнаружения мультиколлинеарности

Важным условием, необходимым для получения состоятельных оценок модели, является отсутствие мультиколлинеарности. При наличии мультиколлинеарности определитель матрицы $X''^{T}X''$ системы нормальных уравнений равен или близок нулю и, следовательно, матрица вырождена. Поэтому решения системы нормальных уравнений не существует.

Для оценки мультиколлинеарности в работе предлагается использовать следующий критерий:

$$\chi^2 = -\left(N-1-\frac{1}{6}(2n+5)\right)\lg\left(\left|\tilde{X}^{T}\tilde{X}\right|\right), \quad (17)$$

где $\left|\tilde{X}^{T}\tilde{X}\right|$ – определитель матрицы $\left[\tilde{X}^{T}\tilde{X}\right]$, имеющий асимптотическое распределение Пирсона χ^2 с $1/2n(n-1)$ степенями свободы; n – число наблюдений; M – число независимых переменных; матрица $\left[\tilde{X}^{T}\tilde{X}\right]$ составлена из значений независимых переменных, преобразованных по формуле

$$\tilde{x}_{ik} = \frac{x''_{ik} - \overline{x}''_i}{\sigma_i\sqrt{n}},$$

где $\overline{x}''_i$, σ_i – соответственно среднее значение и среднеквадратическое отклонение для *i*-ой независимой переменной.

Считается, что мультиколлинеарность отсутствует, если выполняется условие

$$\chi^2 \geq \chi^2_{табл},$$

где χ^2 – расчётное значение критерия χ^2, определяемое по формуле (17), $\chi^2_{табл}$ – табличное значение критерия χ^2 с 1/2M(M-1) степенями свободы и выбранным уровнем надёжности.

В противном случае для каждой *i*-й переменной определяются величины

$$d_i = \frac{\left(\tilde{x}^{T}\tilde{x}\right)_{ii}}{\left|\tilde{X}^{T}\tilde{X}\right|},$$

где $\left(\tilde{x}^{T}\tilde{x}\right)_{ii}$ – i-й диагональный элемент матрицы $\left[\tilde{X}^{T}\tilde{X}\right]$.

При отсутствии мультиколлинеарности величина d_i близка к единице, при наличии мультиколлинеарности – стремится к бесконечности.

Знание величины d_i даёт основание оставить или отбросить показатель x''_i. Надёжность принимаемого решения относительно независимой переменной x''_i определяется величиной

$$W_i = (d_i - 1)\frac{n-M}{M-1},$$

которая имеет распределение Фишера с $v_1 = n-M$ и $v_1 = M-1$ степенями свободы.

Если выполняется условие $W_i \geq F_{p,v_1,v_2}$, то принимается решение о том, что независимая переменная x''_i должна оставаться в модели. В целях устранения или уменьшения мультиколлинеарности можно переходить к разностям для исходной информации или использовать методы факторного анализа, метод главных компонент или к КР метода «Гусеница»-SSA, что и делается в данной работе.

Для непосредственной оценки порядка выбираемых моделей с учётом требований точности и надёжности результатов исследуются показатели относительной надежности оценок параметров (коэффициенты вариации):

$$V_i = \frac{\sigma_i(m)}{|\tilde{a}_i|},\ i=\pm1, ...,\pm m.$$

Коэффициент вариации используется как мера рассеивания коэффициентов модели. Для проведения конкретных расчётов могут задаваться специальные ограничения, определяющие надёжность полученных коэффициентов модели, например

$$\frac{\sigma_i(m)}{|\tilde{a}_i|} < 0{,}5 \ \ (i = \overline{0,m}).$$

Подобное ограничение должно задаваться для каждого отдельного случая. Действительно, с увеличением порядка модели уменьшается величина $\sigma_i(m)$, но и величины $|\hat{a}_i|$ при больших порядках модели быстро уменьшаются. В этой связи надёжность (достоверность) получаемых оценок коэффициентов существенно падает. Таким образом, необходимо достижение определённого компромисса между некоторым увеличением точности и уменьшением надёжности при увеличении порядка модели. Для используемых моделей выбирается такой порядок, для которого реализуется условие

$$\frac{\sigma_i(m)}{|\hat{a}_{m_i}|} < l,$$

где величина l задаётся из некоторых условий, например по критерию Стьюдента. Но можно указать и область значений порядка модели, для которого надёжность вычисляемых оценок коэффициентов мала. Эта область задаётся условием

$$\frac{\sigma_i(m)}{|\hat{a}_{m_i}|} < t_{p,v},$$

где $t_{p,v}$ – значение критерия Стьюдента с v степенями свободы и уровнем надёжности *p*. В этом случае доверительный интервал для коэффициентов модели имеет вид

$$\hat{a}_{m_i} - t_{p,v}\sigma_i(m) < a_{m_i} < \hat{a}_{m_i} + t_{p,v}\sigma_i(m)$$

и включает нулевые значения коэффициентов a_{m_i}, т. е. гипотеза о нулевом математическом ожидании оценок коэффициентов не отвергается [21].

6. Результаты исследования исследований эффективности предложенных моделей прогнозирования временных рядов

Исследование предлагаемых гибридных моделей на основе многомерного варианта метода «Гусе-

ница»-SSA, МГУА, FOS и моделей SARIMAX осуществим, сопоставляя их результаты прогнозов с результатами прогнозов, полученных гибридными моделями на основе методов «Гусеница»-SSA и моделей SARIMA [9]. Реализация рассмотренных моделей производилась в математическом пакете MATLAB R2014a.

Тестирование будем проводить на ВР часовых значений потребления электроэнергии объёма 1008 значений, что соответствует 6 неделям (рис. 6), с учётом изменения температуры воздуха (рис. 7). Обучение моделей будет производиться на выборках данных за 5 недель (840 значений), а тестирование – на данных последней недели. Прогноз будет выполняться одношаговый и производиться скольжение окна до последнего значения ВР.

Сравнительный анализ эффективности прогнозирования рассмотренными моделями будем осуществлять при помощи статистики RMSE (Root Mean Squared Error):

$$RMSE = \sqrt{\frac{1}{n_1}\sum_{t=1}^{n_1}\left(y_t - \hat{y}_t\right)^2},$$

где n_1 – количество вычисленных прогнозов, y_t – фактические значения ВР, $\hat{y}_t$ – прогнозные значения ВР.

Для упрощения эксперимента выберем длину окна L=24. В модель включим все 24 КР прогнозируемого ВР и 24 КР экзогенного ВР. Выберем максимальную величину задержки для каждой из КР равной m_i=10, $i = \overline{1,2 \cdot 24}$.

На рис. 8 приведена столбчатая диаграмма зависимости среднеквадратической ошибки моделирования ВР от включения в модель i-ой переменной, отобранной методом FOS. Для наглядности отобразим зависимость только для 30-и первых отобранных переменных.

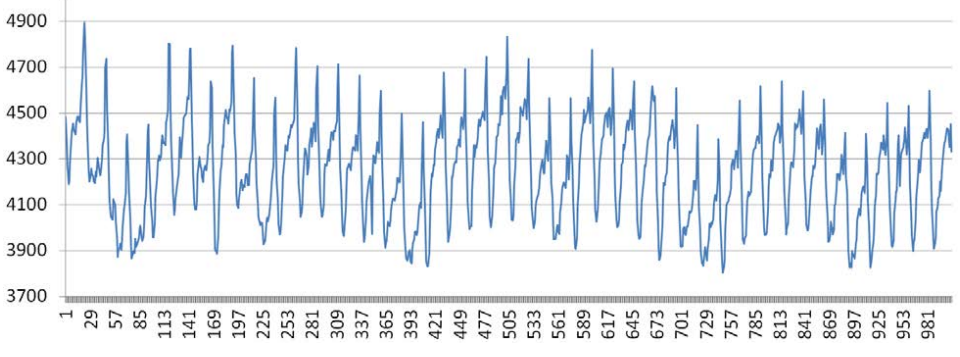

Рис. 6. График ВР часовых значений потребления электроэнергии за 6 недель (в МВт)

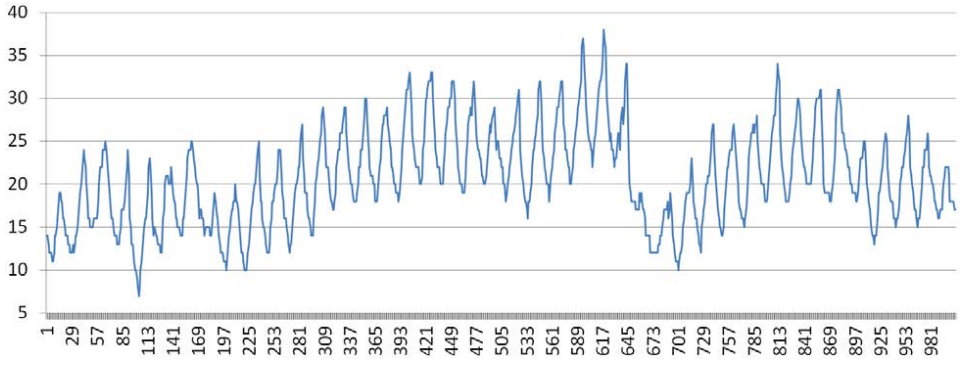

Рис. 7. График ВР часовых значений изменения температуры воздуха за 6 недель (в С°)

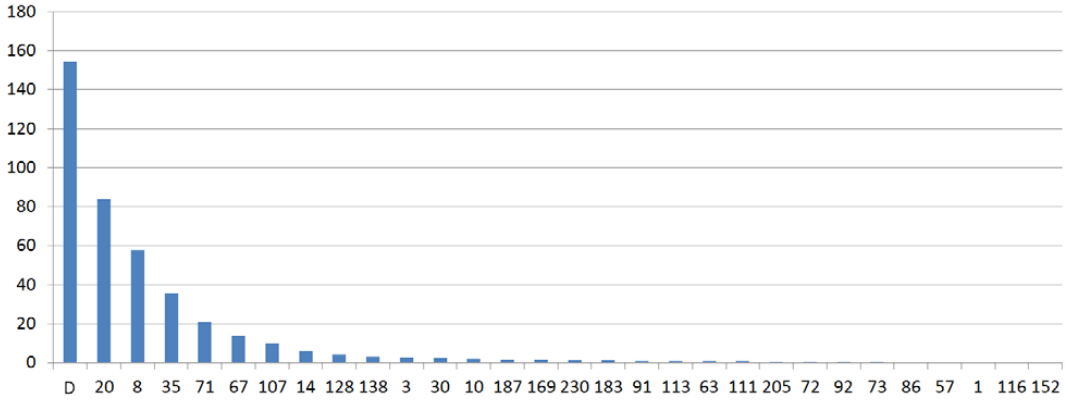

Рис. 8. Зависимость среднеквадратической ошибки аппроксимации ВР методом FOS от включения в модель *i*-й переменной

Тестирование будет проводиться следующим образом. Длина окна метода «Гусеница»-SSA будет фиксированной, но модели одного класса будут отличаться количеством переменных, отобранных алгоритмом FOS и корректирующей моделью. На рис. 8 изображена столбчатая диаграмма зависимости среднеквадратической ошибки прогнозирования ВР методом «Гусеница»-SSA – FOS – SARIMA от включения в модель i-го регрессора, отобранного по алгоритму FOS. Для наглядности отобразим зависимость только для 80-и первых отобранных переменных.

В результате обучения модели «Гусеница»-SSA – FOS – SARIMA, для наиболее эффективной модели метод FOS отобрал 60 переменных (рис. 9). С ростом числа оцениваемых параметров и связанным с этим улучшением точности происходит уменьшение статистической надёжности оценок параметров. Это объясняется тем, что при переходе от быстрого убывания величин дисперсий ошибки модели к медленному, векторы переменных модели становятся практически линейно зависимыми, а матрицы соответствующих систем нормальных уравнений – плохо обусловленными.

На рис. 10 приведены прогнозы модели «Гусеница»-SSA – FOS – SARIMA.

Ошибка RMSE модели «Гусеница»-SSA – FOS – SARIMA составила 5,791.

Протестируем модель «Гусеница»-SSA – МГУА – SARIMA на тех же данных. В качестве ЧО выберем (10). Параметры метода «Гусеница»-SSA оставляем теми же. Для МГУА, для упрощения, выберем количество слоёв не более 15-и и количество отбираемых моделей на каждом слое – 40.

Точки предыстории, по которым осуществляется выбор модели тренда оптимальной сложности, разбиваем на обучающую и проверочную последовательности. Объём проверочной выборки составляет 30 % от общего объёма данных. Разбиение целесообразно осуществлять по величине их дисперсии относительно среднего значения. При этом способе разбиения предыстории для определения коэффициентов модели используем более удалённые точки от среднего значения, а проверочную последовательность составим из точек, имеющих меньшую дисперсию.

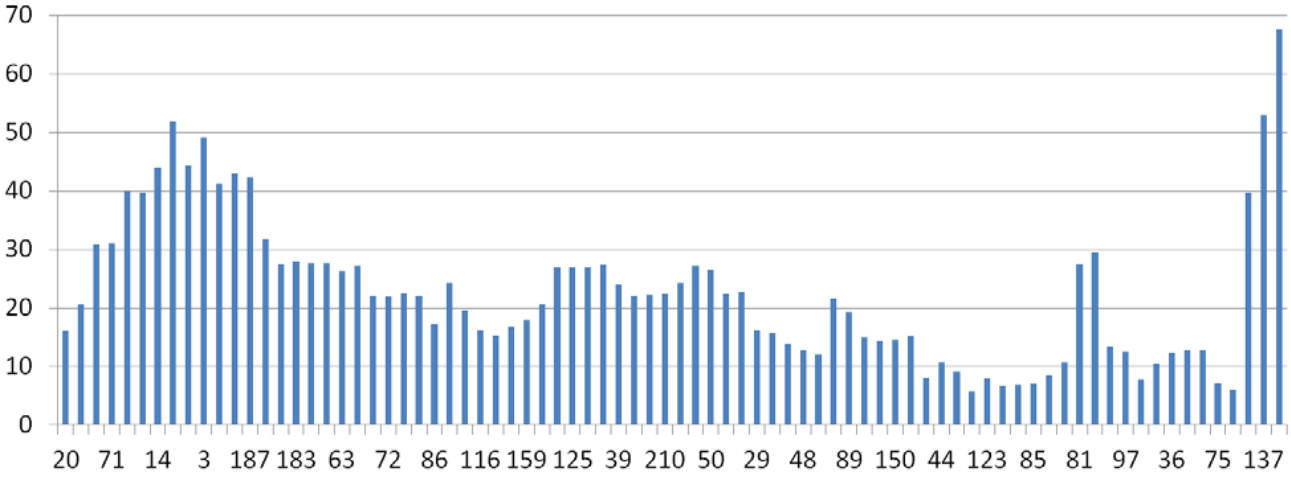

Рис. 9. Зависимость среднеквадратической ошибки прогнозирования ВР методом «Гусеница»-SSA – FOS – SARIMA от включения в модель i-го регрессора

Рис. 10. График потребления электроэнергии и прогнозы, полученные моделью «Гусеница»-SSA – FOS – SARIMA

В результате обучения модели «Гусеница»-SSA – МГУА – SARIMA была получена модель МГУА со следующей структурой (табл. 1).

Таблица 1

Структура модели, полученной по МГУА

№ слоя	Индексы переменных
15	6
14	7, 120
13	211, 7, 214, 196
12	492, 283, 65, 113, 314, 548
11	156, 76, 25, 72, 2, 1, 29, 155
10	164, 572, 0, 418, 157, 374, 49, 411, 413, 191, 574
9	92, 42, 527, 17, 128, 2, 550, 433, 482, 293
8	32, 588, 5, 17, 31, 7, 15, 35, 25, 24, 21
7	42, 410, 65, 209, 59, 201, 419, 64, 206, 28, 302, 417, 297
6	7, 15, 19, 25, 29, 32, 27, 30, 2, 37, 22
5	34, 11, 17, 23, 26, 36, 37, 28, 35, 6, 38, 22
4	270, 78, 32, 12, 37, 15, 22, 28, 26, 29, 24, 31, 16
3	40, 2, 42, 8, 4, 47, 12, 57, 24, 28, 29, 17, 22, 46, 32, 13
2	639, 642, 638, 700, 59, 63, 201, 100, 234, 297, 301, 270, 356, 212, 205
1	2114, 935, 824, 1045, 2104, 1054, 833, 934, 711, 598, 721, 608, 2109
переменные	20, 45, 9, 21, 8, 10, 35, 30, 7, 6, 40

На рис. 11 приведены прогнозы модели «Гусеница»-SSA – МГУА – SARIMA.

Ошибка RMSE модели «Гусеница»-SSA – МГУА – SARIMA составила 3,639.

На рис. 12 приведены ошибки прогнозирования для рассмотренных классов моделей.

Из результатов видно существенное снижение ошибки прогнозирования предложенными нелинейными разреженными моделями в сравнении с гибридными моделями на основе методов «Гусеница»-SSA и моделей SARIMA.

7. Выводы

В работе приведена структурная схема, подходящая для описания практически любой известной на сегодняшний день комбинированной, гибридной или декомпозиционной модели прогнозирования временных рядов. На основе данной схемы предложены методы структурной идентификации разреженных нелинейных моделей взаимосвязанных нестационарных временных рядов на основе методов «Гусеница»-SSA, быстрого ортогонального поиска, метода группового учёта аргументов и моделей SARIMA.

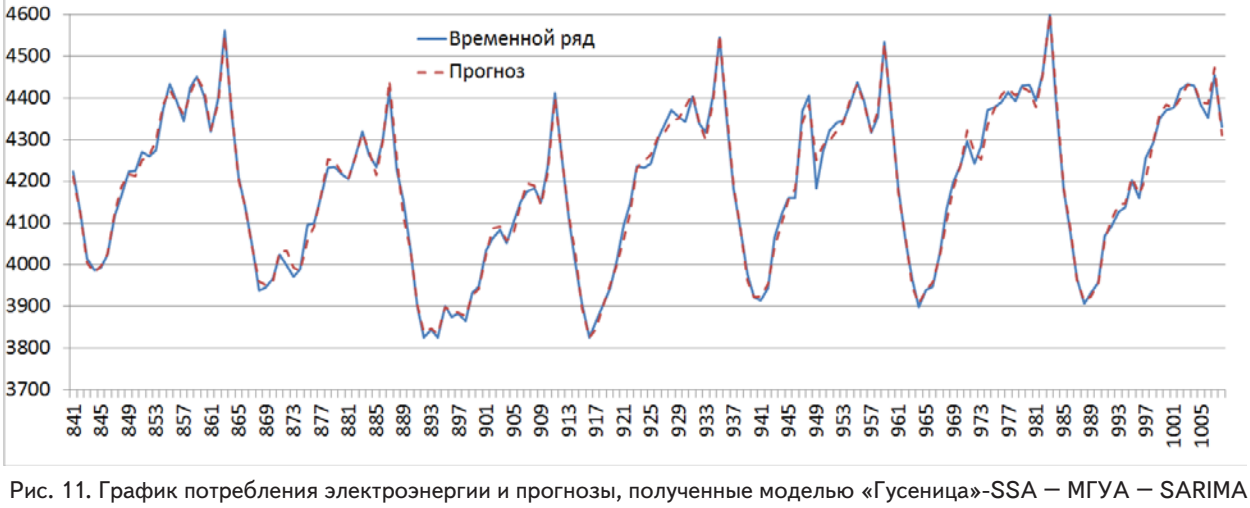

Рис. 11. График потребления электроэнергии и прогнозы, полученные моделью «Гусеница»-SSA – МГУА – SARIMA

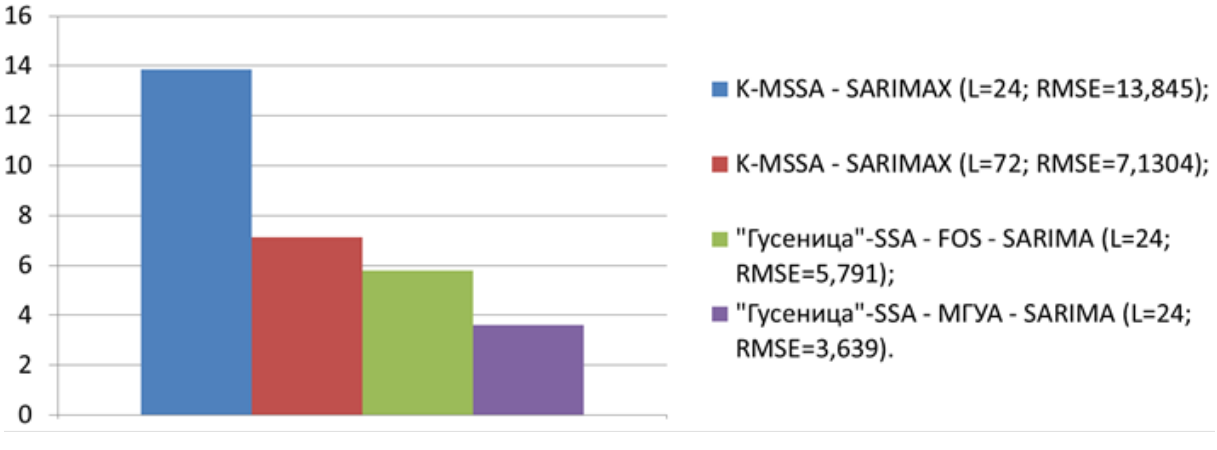

Рис. 12. Диаграмма ошибок прогнозирования для рассмотренных классов моделей

Метод «Гусеница»-SSA в работе применяется для порождения переменных. Метод быстрого ортогонального поиска, в одной из предложенных моделей, применяется для оптимального прореживания. В другой – метод группового учёта аргументов применяется для прореживания полинома Колмогорова-Габора, построенного на компонентах разложения метода «Гусеница»-SSA, примененного к прогнозируемому и экзогенным временным рядам. Для коррекции прогнозов в обеих моделях использовалась сезонная модель авторегрессии – проинтегрированного скользящего среднего.

Выбор соответствующего набора лаговых значений переменных повышает эффективность модели прогнозирования, снижает затраты на параметрическую идентификацию модели и облегчает интерпретацию прогнозируемого ВР.

Предложенный алгоритм структурной идентификации на основе алгоритма *FOS* использует тот факт, что базисные функции, наиболее коррелируемые с прогнозируемым ВР являются наиболее значимыми переменными модели. Алгоритм сортирует все возможные переменные-кандидаты в порядке убывания их корреляции с прогнозируемым ВР. Показано, что такой порядок отбора гарантирует, что наиболее существенные переменные будут отобраны первыми. Используемый в работе алгоритм быстрого ортогонального поиска применяется, чтобы отобрать наиболее значимые переменные и вычислить связанные с ними весовые коэффициенты с помощью ортогонального поиска и разложения Холецкого.

Экспериментальные результаты показывают высокую эффективность предложенных моделей прогнозирования в сравнении с гибридными моделями на основе методов «Гусеница»-SSA и моделей SARIMAX.

Таким образом, решение проблемы синтеза класса математических моделей прогнозирования взаимосвязанных нестационарных ВР является сложным, требующим перебора большого количества вариантов. Субъектом структурной идентификации (СИ) таких моделей должен быть коллектив специалистов, а сама СИ должна являться системным объектом. Используя приведенные схемы и с появлением новых методов разложения, отбора, порождения переменных и прогнозирования временных рядов, можно синтезировать новые классы гибридных моделей.

Литература

1. Давыдов, В. А. Очистка геофизических данных от шумов с использованием преобразования Гильберта-Хуанга [Текст] / В. А. Давыдов, А. В. Давыдов // Электронное научное издание "Актуальные инновационные исследования: наука и практика". – 2010. – № 1.
2. Городеций, А. Е. Нечеткое математическое моделирование плохо формализуемых процессов и систем [Текст] / А. Е. Городеций, И. JI. Тарасова. – СПб.: Изд-во Политехи, ун-та, 2010. – 336 с.
3. Стрижов, В. В. Методы выбора регрессионных моделей [Текст] / В. В. Стрижов, Е. А. Крымова. – М.: Вычислительный центр им. А. А. Дородницына, 2010. – 60 с.
4. Страгович, В. Г. Адаптивное управление [Текст] / В. Г. Страгович. – М.: Наука, 1981. – 381 с.
5. Смоляк, С. А. Устойчивые методы оценивания [Текст] / С. А. Смоляк, Б. И. Титаренко. – М.: Статистика, 1980. – 208 с.
6. Седов, А. В. Моделирование объектов с дискретно-распределенными параметрами: декомпозиционный подход [Текст] / А. В. Седов. – М.: Наука, 2010. – 438 с.
7. Гинсберг, К. С. Проблема структурной идентификации для цели проектирования системы автоматического управления [Текст]: труды X междун. конф. / К. С. Гинсберг // Идентификация систем и задачи управления. – Институт проблем управления им. В. А. Трапезникова РАН. – М.: , 2015. – С. 43–80.
8. Щелкалин, В. Н. Гибридные модели и методы прогнозирования временных рядов на основе методов «Гусеница»-SSA и Бокса-Дженкинса [Текст] / В. Н. Щелкалин // Восточно-Европейский журнал передовых технологий. – 2014. – Т. 5, № 4 (71). – С. 43–62. doi: 10.15587/1729-4061.2014.28172
9. Щелкалин, В. Н. Гибридные математические модели и методы прогнозирования временных рядов с учётом внешних факторов [Текст] / В. Н. Щелкалин // Восточно-Европейский журнал передовых технологий. – 2014. – Т. 6, № 4 (72). – С. 38–58. doi: 10.15587/1729-4061.2014.31729
10. Щелкалин, В. Н. Гибридные математические модели и методы прогнозирования взаимосвязанных нестационарных временных рядов [Текст] / В. Н. Щелкалин // Восточно-Европейский журнал передовых технологий. – 2015. – Т. 1, No 4 (73). – С. 42–58. doi: 10.15587/1729-4061.2015.37317
11. Zhang, G. P. Time Series Forecasting Using a Hybrid ARIMA and Neural Network Model [Text] / G. P. Zhang // Neurocomputing. – 2003. – Vol. 50. – P. 159–175. doi: 10.1016/s0925-2312(01)00702-0
12. Jain, A. An evaluation of artificial neural network technique for the determination of infiltration model parameters [Text] / A. Jain, A. Kumar // Applied Soft Computing. – 2006. – Vol. 6, Issue 3. – P. 272–282. doi: 10.1016/j.asoc.2004.12.007
13. Su, C. T. Combination of time series and neural network for reliability forecasting modeling [Text] / C. T. Su, L. I. Tong, C. M. Leou // Journal of Chinese Industrial Engineering. – 1997. – Vol. 14. – P. 419–429.
14. Wang, W. Improving daily stream flow forecasts by combining ARMA and ANN models [Text] / W. Wang, P. V. Gelder, J. K. Vrijling // International Conference on Innovation Advances and Implementation of Flood Forecasting Technology, 2005.
15. Onwubolu, G. C. Design of hybrid differential evolution and group method of data handling networks for modeling and prediction [Text] / G. C. Onwubolu // Information Sciences. – 2008. – Vol. 178, Issue 18. – P. 3616–3634. doi: 10.1016/j.ins.2008.05.013

16. Samsudin, R. A hybrid GMDH and least squares support vector machines in time series forecasting [Text] / R. Samsudin, P. Saad, A. Shabri // Neural Network World. – 2011. – Vol. 21, Issue 3. – P. 251–268. doi: 10.14311/nnw.2011.21.015
17. Бэнн, Д. В. Сравнительные модели прогнозирования электрической нагрузки [Текст] / Д. В. Бэнн, Е. Д. Фармер; пер. с англ. – М.: Энергоатомиздат, 1987. – 200 с.
18. Тутубалин, В. Н. Теория вероятностей и случайных процессов [Текст]: учеб. пособие / В. Н. Тутубалин. – М.: Изд-во МГУ, 1992. – 400 с.
19. Прангишвили, И. В. Идентификация систем и задачи управления: на пути к современным системным методологиям [Текст] / И. В. Прангишвили, В. А. Лотоцкий, К. С. Гинсберг, В. В. Смолянинов // Проблемы управления. – 2004. – № 4. – С. 2–15.
20. Щелкалин, В. Н. Системный поход к синтезу класса моде- лей для прогнозирования взаимосвязанных нестационарных временных рядов [Текст] / В. Н. Щелкалин // Материалы 15-й Международной научно-технической конференции SAIT, 2013. – УНК «ИПСА» НТУУ «КПИ», 2013. – С. 338–339.
21. Горелова, В. Л. Основы прогнозирования систем [Текст]: учеб. пособ. / В. Л. Горелова, Е. Н. Мельникова. – М.: Высш. шк., 1986. – 287 с.
22. Гребенюк, Е. А. Проблемы субъективности в решении задач управления и прогноза, связанных с анализом временных рядов [Текст] / Е. А. Гребенюк, М. Г. Логунов, О. А. Мамиконова, Л. А. Панкова. – Человеческий фактор в управлении, 2006. – С. 156–178.
23. Valenca, I. Hybrid Systems to Select Variables for Time Series Forecasting Using MLP and Search Algorithms [Text] / I. Valenca, T. Ludermir, M. Valenca // Eleventh Brazilizn Symposium on Neural Networks, 2010, p. 247 – 252. doi: 10.1109/sbrn.2010.50
24. Yao, L. Genetic algorithm based identification of nonlinear systems by sparse volterra filters [Text] / L. Yao // IEEE Transactions on Signal Processing. – 1999. – Vol. 47, Issue 12. – P. 3433–3435. doi: 10.1109/78.806093
25. Abbas, H. Volterra-system identification using adaptive real-coded genetic algorithm [Text] / H. Abbas, M. Bayoumi // IEEE Transactions on Systems, Man and Cybernetics, Part A. – 2006. – Vol. 36, Issue 4. – P. 671–684. doi: 10.1109/tsmca.2005.853495
26. Chen, S. Orthogonal least squares learning for radial basis function networks [Text] / S. Chen, C. Cowan, P. Grant // IEEE Transactions on Neural Networks. – 1991. – Vol. 2, Issue 2. – P. 302–309. doi: 10.1109/72.80341
27. Ivakheneko, A. G. A review of problems solved by algorithms of the GMDH [Text] / A. G. Ivakheneko, G. A. Ivakheneko // Pattern Recognition and Image Analysis. – 1995. – Vol. 5, Issue 4. – P. 527–535.
28. Guyon, I. An introduction to variable and feature selection [Text] / I. Guyon and A. Elisseeff // J. Mach. Learn. Res. – 2003. – Vol. 3. – P. 1157–1182.
29. Blum, A. L. Selection of relevant features and examples in machine learning [Text] / A. L. Blum, P. Langley // Artificial Intelligence. – 1997. – Vol. 97, Issue 1-2. – P. 245–271. doi: 10.1016/s0004-3702(97)00063-5
30. Гузаиров, М. Б. Системный подход к анализу сложных систем и процессов на основе триад [Текст] / М. Б. Гузаиров, Б. Г. Ильясов, И. Б. Герасимова // Проблемы управления. – 2007. – № 5. – С. 32–38.
31. Korenberg, M. J. A robust orthogonal algorithm for system identification and time-series analysis [Text] / M. J. Korenberg // Biological Cybernetics. – 1989. – Vol. 60, Issue 4. – P. 267–276. doi: 10.1007/bf00204124
32. Евдокимов, А. Г. Оперативное управление потокораспределением в инженерных сетях [Текст] / А. Г. Евдокимов, А. Д. Тевяшев. – Х.: Вища школа, 1980. – 144 с.