

Гибридный подход для краткосрочного прогнозирования временных рядов на основе штрафного P -сплайна и эволюционной оптимизации

Е.А. Кочегурова¹, Е.Ю. Репина¹, О.Б. Цехан²

¹Национальный исследовательский Томский политехнический университет, Томск, Россия,

²Гродненский государственный университет имени Янки Купалы, Гродно, Беларусь

Аннотация

В работе предложена гибридная модель прогнозирования, включающая рекуррентную схему штрафного P -сплайна с адаптацией параметров на основе алгоритмов эволюционной оптимизации. В задачах краткосрочного прогнозирования, особенно в системах реального времени, актуальной является задача повышения скорости прогноза без ухудшения его качества. Высокая скорость прогнозирования в данном подходе достигается экономичной вычислительной схемой рекуррентного P -сплайна при малой глубине предыстории. А одновременная адаптация нескольких параметров P -сплайна позволяет управлять точностью прогноза.

Ключевые слова: временной ряд, штрафной P -сплайн, гибридная модель прогнозирования, многоцелевая оптимизация, эволюционные алгоритмы.

Цитирование: Кочегурова, Е.А. Гибридный подход для краткосрочного прогнозирования временных рядов на основе штрафного P -сплайна и эволюционной оптимизации / Е.А. Кочегурова, Е.Ю. Репина, О.Б. Цехан // Компьютерная оптика. – 2020. – Т. 44, № 5. – С. 821-829. – DOI: 10.18287/2412-6179-CO-667.

Citation: Kochegurova EA, Repina EY, Tsekhan OB. Hybrid approach for time series forecasting based on a penalty p-spline and evolutionary optimization. Computer Optics 2020; 44(5): 821-829. DOI: 10.18287/2412-6179-CO-667.

Введение

Прогнозирование временных рядов является актуальной задачей со множеством приложений в самых разнообразных прикладных областях [1–3].

Однако особенности прикладных задач прогнозирования последних лет во многом обусловлены большими объемами данных. Это связано с растущим использованием GPS, включая информацию от систем позиционирования транспорта и других объектов, с мобильных телефонов, фитнес-браслетов, от датчиков интернет вещей (IoT) и пр. Данные, собранные подобными системами, представляют собой временные ряды (ВР), т.е. наборы числовых последовательностей в хронологическом порядке. Данные ВР всегда рассматриваются как единое целое, а не как отдельные числовые поля [4]. Эти данные динамические, и их анализ позволяет оценить тенденции изменения процесса. На этом и основаны прогностические свойства ВР [5, 6].

В зависимости от числа прогнозируемых значений ВР (горизонт прогноза) выделяют краткосрочное, среднесрочное и долгосрочное прогнозирование. Горизонт краткосрочного прогнозирования включает 1–3 значения вперед [7] и используется во множестве прикладных областей. Краткосрочное прогнозирование актуально для предсказаний чрезвычайных ситуаций техногенного и природного характера; различных природных явлений; для развития возобновляемых источников энергии и предсказания нагрузок в традиционной энергетике; при расчете скоростей передачи данных оптоволоконных каналов связи; для прогноза спроса, цен и финансовых рынков; при выявлении эпидемических вспышек заболеваемости.

Особое место в задачах прогнозирования последнего десятилетия занимают задачи анализа и прогнозирования активности пользователей социальных сетей.

1. О методах краткосрочного прогнозирования

Существует несколько классификаций методов краткосрочного прогнозирования. Согласно одной из них все методы прогнозирования могут быть разбиты на интуитивные и формализованные методы прогнозирования ВР.

Формализованные методы (часто называемые моделями), в свою очередь, включают в себя статистические и структурные модели прогнозирования [8, 9]. Многие структурные модели основаны на методах искусственного интеллекта.

Согласно [10] методы прогнозирования могут быть сгруппированы в соответствии с параметрическим и непараметрическим подходом. Параметрический объединяет методы, использующие предшествующие знания о распределении данных (экспоненциальное сглаживание, основанные на авторегрессии и скользящих средних). Непараметрический подход не требует в явном виде информации о распределении данных.

По степени использования предшествующих ВР методы разделяются на глобальные или локальные. Локальные методы воспроизводят целые фрагменты ВР на основании их образов в прошлом. Часто используемый при этом математический инструментарий – выборки максимального правдоподобия и локальная аппроксимация. Основное требование этой группы методов состоит в превышении размера выборки над горизонтом прогноза. По разным оценкам

от 100 до 700 раз. И поэтому они не применимы в режиме реального времени (PPB) [9, 11].

Глобальные методы не отыскивают фрагменты квазипериодичности в прошлом, а на основании рекуррентной формулы скрыто используют всю информацию о наблюдаемом процессе. Глобальные методы не явно реализуют идею адаптивного прогнозирования в PPB.

Однако в последние годы появилось достаточно много гибридных методов, объединяющих различные подходы к временному анализу данных. Особенно часто гибридные подходы используются в реальных приложениях. И по данным [10], частота гибридных подходов превысила 21 % против 54 % использования непараметрических методов и 25 % параметрических методов. С учетом этого целесообразно группировать методы краткосрочного прогнозирования по трем категориям, приведенным на рис. 1.



Рис. 1. Классификация методов краткосрочного прогнозирования

Статистические методы основаны на информации о распределении данных для получения моделей прогноза. Это делает методы зависимыми от параметров, оптимизация которых значительно влияет на результаты прогнозирования. По математической сложности статистические методы разделены [4]:

- на модели экспоненциального сглаживания (SES – простое экспоненциальное сглаживание, HES – Хольта и HW – Хольта-Винтерса);
- регрессионные модели основаны на влиянии входных переменных на выходную. Способом реализации регрессионных моделей является преимущественно метод наименьших квадратов (МНК). Нередко используются рекуррентные модификации МНК с учетом весов предшествующих значений BP;
- авторегрессионные модели основаны на зависимости текущего значения прогнозируемой переменной только от прошлых значений этой же переменной. Наибольшее развитие получила модель ARIMA, включающая этапы авторегрессии, инте-

грации и оценки скользящего среднего. Модель ARIMA имеет несколько модификаций: ARIMAX – включает в себя некоторые экзогенные факторы, SARIMA – учет сезонности данных, VARIMA – многовекторные временные ряды.

Структурные модели прогнозирования нередко основаны на методах машинного обучения и не имеют ограничений на характер входных данных. Эти методы проще в настройке, в том числе и для сложных, и для сильно нелинейных BP.

Существует множество вариантов применения искусственных нейронных сетей (ANN) в краткосрочном прогнозировании. Нейронные сети могут отличаться архитектурой, количеством слоев и нейронов, использованием определенной функции активации или способом обучения. В прогнозировании используются следующие виды ANN:

- сети прямого распространения (feed-forward neural networks);
- рекуррентные нейронные сети (простые RNN, с долгосрочной памятью LSTM, с управляемыми нейронами GRU);
- сверточные нейронные сети.

Основным недостатком ИНН является ресурсоемкость обучения сети и сложность ее идентификации [11].

Метод опорных векторов (SVM) широко используется в регрессионном анализе и в задачах классификации [12]. В применении к задаче прогноза сложных и нелинейных данных метод SVM требует выбора не простых разделительных плоскостей, а сложных функций ядра. Эти ядра позволяют выполнить отображение входных данных в многомерное пространство признаков. SVM порождает несколько параметров, получаемых чаще всего при решении нелинейной задачи оптимизации с линейными ограничениями.

Название «гибридные методы прогнозирования» достаточно условно. На сегодня эта группа методов наиболее перспективна при создании прогнозных моделей для сложных процессов с переменной и неоднородной структурой [13].

Гибридные методы объединяют в себе особенности методов из различных групп. Обычно один метод выбран из группы статистических, а другой – из структурных. Например, концепции нечеткой логики и вейвлетов [14] или нейронных сетей [15] стали весьма привлекательными для комбинации алгоритмов. Нечеткая логика позволяет уменьшить сложность данных или их неопределенность [16], а вейвлет-анализ выделяет локальные фрагменты нестационарных сигналов [17]. На рис. 1 представлены лишь некоторые, наиболее часто используемые комбинации методов прогнозирования.

Часть гибридных методов является локальными и, как правило, двухэтапными. На начальном этапе временной ряд разбивают на куски структурно однородных фрагментов BP в соответствии с выделенной ха-

рактической или прикладной задачей анализа ряда, например, выделение тренда или сегментация. Далее формируется прогнозное значение на основе комбинации текущего значения ряда и характеристик выделенных фрагментов. Нередко данный подход включает алгоритмы машинного обучения, которые позволяют в процессе прогнозирования использовать различные фрагменты ВР для обучения.

Другая стратегия гибридных методов основана на методах поиска сходства. Алгоритм ближайшего соседа kNN, хорошо зарекомендовавший себя в задачах классификации и кластеризации [12, 18], эффективен и в составе гибридных методов прогнозирования. Вариации алгоритма kNN в комбинации с разными функциями приближения (например, взвешенное или локальное среднее) рядом авторов успешно использованы для нелинейных и сложных временных рядов [19, 20].

При неизвестной модели ВР начальный этап прогнозирования выполняет предобработку данных. И независимо от идей сегментации или кластеризации эффективно работает лишь в апостериорных режимах. Для дальнейшего предсказания данных в РРВ необходима аналитическая модель процесса, допускающая последовательную или рекуррентную реализацию.

При создании модели ВР целесообразно заменить традиционные сложные регрессионные соотношения простыми гладкими функциями аппроксимации. Перспективными с позиции описания локального поведения сложных ВР являются кусочно-полиномиальные сплайн-функции. Выбор узлов сплайна для получения его отдельных фрагментов является фактически способом создания паттерна для поиска сходства части ВР в будущем. Кроме того, некоторые виды сплайн-функций допускают реализацию, а соответственно, и прогноз в РРВ. Что значительно сокращает время получения прогнозных значений. Подобный подход к прогнозированию исследован в данной работе.

2. Описание метода

Сглаживающие сплайны – эффективный инструмент аппроксимации зашумленных данных. Наиболее известным примером сглаживающего сплайна является кубический сглаживающий сплайн [21], хорошо изученный в режиме апостериорного сглаживания.

Регрессионные сплайны используют определенный набор базисных функций с уменьшенным количеством узлов. Реализуются они часто на основе метода наименьших квадратов. Штрафы за негладкость в регрессионных сплайнах отсутствуют.

Эффективность штрафных P-сплайнов [22] и базисных B-сплайнов [23] значительно повышается при оптимальном выборе узлов. Именно выбор узлов можно рассматривать как способ параметрической адаптации сплайна к динамике ВР [24]. Другим инструментом адаптации является оценка штрафного

параметра с использованием регрессионных и вероятностных подходов [25].

Именно штрафные P-сплайны объединяют достоинства регрессионных и сглаживающих сплайнов: уменьшенное количество узлов и контроль гладкости сплайна.

В данной работе предлагается модификация P-сплайна, реализующая вариационный подход для отдельных фрагментов ВР (группы данных) [26, 27]. Объединение входных данных в группы, с одной стороны, решает проблему выбора узлов. С другой – выделяет фрагмент ВР (shapelet, паттерн) и, следовательно, реализует гибридный подход к прогнозированию на основе сходства. При этом экономичный алгоритм оценки параметров рекуррентной сплайн-функции исключает использование численных методов и тем самым позволяет реализовать сплайн в реальном времени.

Получение классического штрафного P-сплайна в апостериорном режиме основано на оптимизации специального вида функционала [24]

$$J(S) = \alpha \cdot \int_a^b [S''(t)]^2 dt + \sum_{i=0}^n [S(t_i) - y(t_i)]^2. \quad (1)$$

В (1) первое слагаемое вместе со сглаживающим параметром α определяет штраф кривизны; второе слагаемое, как и в методе наименьших квадратов, задает приближающие свойства. На основе функционала (1) возможно одновременно вычислить все звенья сплайна $S(t)$, но только после получения всех измерений $y(t_j)$, $j = 0, n$, на интервале наблюдения $[a, b]$.

Реализации прогноза в реальном времени потребовала модифицировать функционал (1) отдельно для каждого i -го звена сплайна, включающего группу из h входных отсчетов $y(t_j)$, $j = 0, h$ [26]. Шаг дискретизации Δt введен для регулирования размерности функционала

$$J(S) = (1 - \rho)(h\Delta t)^2 \int_{t_0^i}^{t_h^i} [S''(t)]^2 dt + \rho \sum_{j=0}^h [S(t_j^i) - y(t_j^i)]^2. \quad (2)$$

Для реализации штрафного P-сплайна для i -го звена в реальном времени

$$S_i(\tau, \bar{a}^i) = a_0^i + a_1^i \cdot \tau + a_2^i \cdot \tau^2 + a_3^i \cdot \tau^3, \quad (3)$$

$$-q \leq \tau \leq h - q,$$

в [27] получены коэффициенты такой формы P-сплайна в рекуррентном виде.

$$a_0^i = a_0^{i-1} + a_1^{i-1} + a_2^{i-1} + a_3^{i-1};$$

$$a_1^i = a_1^{i-1} + 2a_2^{i-1} + 3a_3^{i-1};$$

$$a_2^i = \frac{\rho(F_1^i C - F_2^i A)}{BC - A^2};$$

$$a_3^i = \frac{\rho(F_2^i B - F_1^i A)}{BC - A^2};$$

$$\begin{aligned}
 A &= 6(1-\rho)h^4 + \rho H_5; \\
 B &= 4(1-\rho)h^3 + \rho H_4; \\
 C &= 12(1-\rho)h^5 + \rho H_6; \\
 F_1^i &= \sum_{j=0}^h y(t_j^i)j^2 - a_0^i H_2 - a_1^i H_3; \\
 F_2^i &= \sum_{j=0}^h y(t_j^i)j^3 - a_0^i H_3 - a_1^i H_4; \\
 H_n &= \sum_{j=0}^h j^n, \quad n = \overline{2,6}.
 \end{aligned}
 \tag{4}$$

В (4) q – номер отсчета внутри i -го звена ($j = \overline{0, h}$; $q = \overline{0, h-1}$), в котором сопряжены непрерывные производные сплайна, $S_{i-1}^k(t_{q+1}^{i-1}, \bar{a}^{i-1})_+ = S_i^k(t_q^i, \bar{a}^i)_-$ $k = \overline{0, 1}$ для дефекта 2 и $k = 0$ для дефекта 1. Разрывные коэффициенты a_2^i, a_3^i найдены из условия минимизации функционала (2):

$$\frac{\partial J(S_i(\tau, \bar{a}^i))}{\partial a_2^i} = 0, \quad \frac{\partial J(S_i(\tau, \bar{a}^i))}{\partial a_3^i} = 0.$$

Формулы (4) локальны для входных измерений внутри звена сплайна и рекуррентны относительно коэффициентов для всего сплайна.

Соотношение моментов вычисления (τ) и сопряжения (q) для i -го звена сплайна порождает несколько вычислительных схем рекуррентного сплайна, подробно исследованных в [27]. Наиболее общим и приспособленным для прогнозирования ВР является текущий режим штрафного P -сплайна, рис. 2.

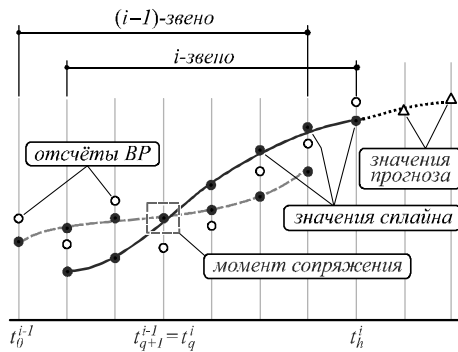


Рис. 2. Режим прогнозирования на основе штрафного P -сплайна

Модель фрагмента временного ряда в виде звена штрафного P -сплайна позволяет получать значения прогноза для $\tau > h$ в реальном времени. На рис. 2 изображен горизонт прогноза для двух значений $\tau = h + 1$ и $\tau = h + 1$.

Фактически группа h отсчетов i -го звена сплайна определяет фрагмент (сегмент) ВР, объединенный общими динамическими свойствами. Поэтому выбор размера группы отсчетов h является задачей адаптации и значительно повышает качество прогнозирования.

3. Адаптация параметров штрафного P -сплайна

Одной из основных трудностей, с которыми сталкиваются исследователи при прогнозировании ВР, является поиск наилучших параметров модели прогнозирования в соответствии с заданным набором данных.

Различные методы и алгоритмы оптимизации неоднократно использовались для обучения многих структурных моделей прогнозирования, в том числе и для искусственных нейронных сетей (ИНС) [28, 29]. Это могут быть локальные техники оптимизации, классические алгоритмы глобального поиска и смешанные технологии [30]. Также разработаны эффективные методы и алгоритмы оптимизации, основанные на оригинальных идеях. Среди них и алгоритмы на базе метода оврагов И.М. Гельфанда [31], и алгоритмы на основе метода неравномерных покрытий Ю.Г. Евтушенко [32] и др. Использование классических методов оптимизации малоэффективно для многоэкстремальных функций, к которым относится большинство реальных поверхностей уровня или модельных данных в условиях помех.

Группа метаэвристических методов хорошо зарекомендовала себя при отсутствии или неполной информации о свойствах целевой функции. Получаемое без доказательства оптимальности и корректности решение (часто субоптимальное) приемлемо и с позиции вычислительных затрат. В отличие от многих классических методов поиска эвристические методы реализуют алгоритмы случайно направленного перебора [33].

Идея метаэвристических алгоритмов исходит из поведения животных или физических явлений [30]. Среди метаэвристических группа эволюционных алгоритмов успешно использована во многих научных и технических задачах. Наиболее популярное приложение эволюционных алгоритмов – оптимизация многопараметрических функций и способность одновременной манипуляции несколькими параметрами.

Для оптимизации параметров штрафного P -сплайна по аналогии с ИНС были использованы генетический алгоритм (ГА) и алгоритм искусственных иммунных систем (ИА) [34].

В ИА используется способность естественной иммунной системы вырабатывать новые типы антител и отбирать наиболее подходящие из них для взаимодействия с попавшими в организм антигенами.

ГА обеспечивает выживание сильнейших генов из множества сгенерированных на основе эволюционных принципов наследственности, изменчивости и естественного отбора [3]. Использование ГА в задачах оптимизации требует меньших вычислительных ресурсов по сравнению с ИА. Но при этом ИА допускает реализацию параллельно распределенного поиска.

Структурные схемы генетических и иммунных алгоритмов весьма схожи и схематично представлены на объединенном рис. 3.

При адаптации параметров штрафного P -сплайна использовано основное достоинство эволюционных алгоритмов – одновременная оптимизация нескольких параметров. В данном случае были одновременно исследованы два основных параметра P -сплайна: сглаживающий множитель $\rho \in [0, 1]$ и длина звена $h \in [3, 20]$. Значение $h = 3$ соответствует использованию 4 отсчетов для построения полинома третьего порядка, а при $h > 20$ модель сложно использовать в системах реального времени.

Устойчивость штрафного P -сплайна подробно исследована в [35] и для выбранного текущего режима не оказывает влияния на области изменения параметров сплайна.

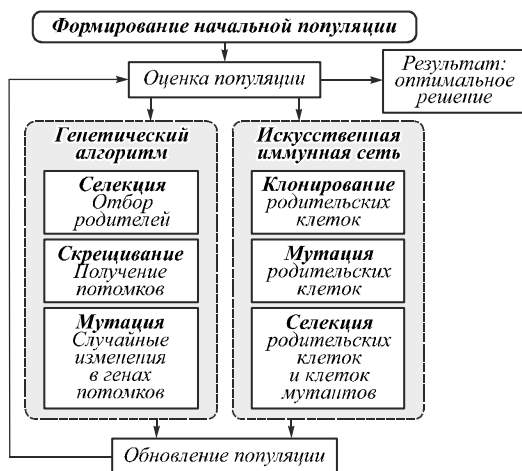


Рис. 3. Структурная схема генетического и иммунного алгоритмов

4. Результаты вычислительного эксперимента

Оптимизация параметров штрафного P -сплайна позволила построить эффективные прогностические модели с разными горизонтами прогнозирования. Качество прогноза традиционно оценивается совокуп-

ностью показателей [1, 10, 36], отражающих временные и точностные характеристики экстраполяции.

К основному показателю точности прогноза нами были предъявлены следующие требования:

- независимость от шкалы ВР;
- устойчивость к выбросам;
- симметричность оценки.

И в качестве основного показателя точности прогноза известная оценка RMSE (Root Mean Squared Error) была дополнена процентным нормированием RMSPE (Percentage Error)

$$RMSPE = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (\hat{y}_i - y_i)^2} \cdot \frac{100}{y_{\max} - y_{\min}}, [\%]. \quad (5)$$

В формуле (5) RMSPE-мера работает с разностью между фактическим наблюдаемым значением y_i и прогнозируемым значением \hat{y}_i .

Говоря о быстродействии алгоритма, можно оценить временную сложность алгоритма, связанную с количеством вычислений для входного массива данных. В случае P -сплайна объем вычислений определяется формулами (4), реализуемыми в РРВ при получении очередного измерения. Алгоритм, реализующий (4), содержит единственный цикл размерности h (размер группы измерений). И в O -нотации временная сложность данного алгоритма составляет $O(h)$. При этом наилучшему случаю соответствует $O(h = 3)$. А наихудшему – $O(h = 20)$. Пространственная сложность (объем памяти) также невелика и составляет $(11+h)$.

Оценка прогностических свойств предлагаемого гибридного алгоритма была проведена с использованием нескольких тестовых (синтетических) и реальных ВР.

Для сопоставления точности прогноза RMSPE-ошибка была получена для случайных и оптимальных значений параметров P -сплайна. Для тестовых ВР результаты вычислительных экспериментов приведены в табл. 1.

Табл. 1. Оптимальные значения параметров и RMSPE-погрешность для тестовых ВР

$\sigma_{\xi}, \%$	0%		5%		10%		15%		20%	
Тест-1	$y_1(t) = 10 \cdot \sin\left(\frac{2\pi \cdot t}{100}\right)$									
h_{opt}	ГА	ИА	ГА	ИА	ГА	ИА	ГА	ИА	ГА	ИА
ρ_{opt}	0,99	0,99	0,61	0,72	0,49	0,68	0,42	0,37	0,35	0,3
RMSPE _{min}	0,006	0,006	0,69	0,65	1,17	1,18	1,66	1,58	2,01	2,06
Тест-2	$y_2(t) = \sin\left(\frac{\pi \cdot t}{20}\right) \cdot e^{0,02 \cdot t} + 3$									
h_{opt}	ГА	ИА	ГА	ИА	ГА	ИА	ГА	ИА	ГА	ИА
ρ_{opt}	0,99	0,99	0,97	0,83	0,62	0,58	0,54	0,48	0,45	0,43
RMSPE _{min}	0,025	0,025	1,4	1,32	2,38	2,47	3,39	3,25	4	3,84

Оптимальные значения параметров гибридного алгоритма прогнозирования h_{opt} и ρ_{opt} найдены на основе эволюционных алгоритмов. Характерно, что ГА и ИА приводят к близким, но все же различным значениям оптимальных параметров. Но при этом RMSPE-погрешность примерно одинакова при заданных значениях уровня шума входных данных σ_ξ . В качестве шума был выбран случайный сигнал $\xi(t)$: $M\{\xi(t)\} = 0$; $M\{\xi^2(t)\} = \sigma_\xi^2$.

Основная цель исследования прогноза на тестовых данных – оценка трендовой составляющей ВР в условиях шума. Графический анализ минимальных значений погрешности, приведенный на рис. 4а, б, показывает значительное сокращение шума входных данных при горизонте прогнозирования на 1 шаг вперед. Погрешность входных данных в среднем уменьшается в 5–10 раз в зависимости от вида тестовой функции. В плане вычислительной сложности используемые алгоритмы оптимизации не имеют преимущества друг перед другом.

Сравнивая производительность ГА и ИА, следует отметить, что время работы иммунного алгоритма значительно выше, чем генетического, в среднем в 1,5–1,75 раз. Поэтому для настройки параметров штрафного P-сплайна для реальных данных использован ГА.

В качестве реальных ВР были использованы данные о цене закрытия акций компаний ПАО «Мегафон» и АО «Сбербанк России» в 2018–2019 гг. Для отображения результатов исследования использованы диаграммы размаха, представленные на рис. 5 а, б. А соответствующие выбранным параметрам значения погрешности прогноза приведены в табл. 2.

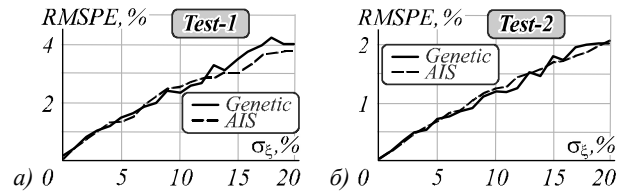


Рис. 4. RMSPE-погрешность при изменении уровня шума входных данных σ_ξ

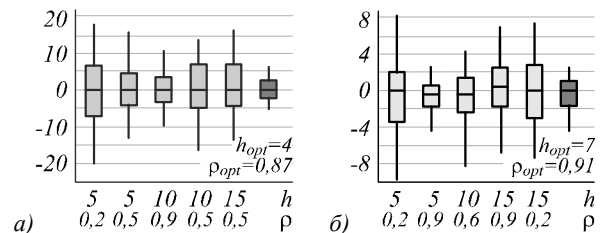


Рис. 5. Разброс RMSPE-погрешности для произвольных и оптимальных значений параметров P-сплайна

Табл. 2. RMSPE-погрешность для реальных временных рядов

Акции ПАО «Мегафон»						
Параметры P-сплайна	$h = 5;$ $\rho = 0,2$	$h = 5;$ $\rho = 0,5$	$h = 10;$ $\rho = 0,5$	$h = 10;$ $\rho = 0,9$	$h = 15;$ $\rho = 0,5$	$h_{opt} = 4;$ $\rho_{opt} = 0,87$
RMSPE, %	7,164	5,564	6,450	5,318	7,535	3,643
Акции АО «Сбербанк России»						
Параметры P-сплайна	$h = 5;$ $\rho = 0,2$	$h = 5;$ $\rho = 0,9$	$h = 10;$ $\rho = 0,6$	$h = 15;$ $\rho = 0,9$	$h = 15;$ $\rho = 0,2$	$h_{opt} = 7;$ $\rho_{opt} = 0,91$
MSPE, %	5,210	3,151	3,558	4,843	6,231	3,151

Графическое сопоставление реальных временных рядов и соответствующих прогнозных моделей, рис. 6, выявило ряд дополнительных свойств P-сплайна в составе гибридной модели прогнозирования. Исходя из RMSPE-оценки качество прогнози-

рования на основе штрафного сплайна находится в пределах 3–4%. Это является неплохим показателем в сравнении с известными методами прогнозирования, детально исследованными в [10].

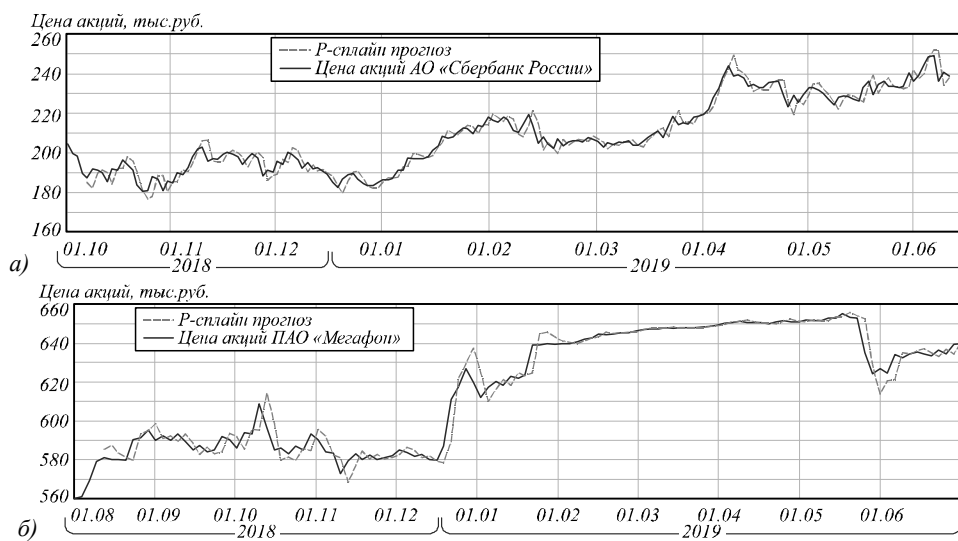


Рис. 6. Реальные временные ряды и прогнозные значения

В исследовании [10] одиннадцать наиболее известных алгоритмов прогнозирования сопоставлены по трем показателям: MSE-погрешность, TU-коэффициент Тейла, отражающий предсказательные возможности алгоритма по отношению к наивной модели прогнозирования, и ROCID-показатель сохранения тенденций.

Сравнение предлагаемого алгоритма на основе P-сплайна с исследуемыми в [10] методами было проведено на одинаковых наборах тестовых данных, включающих 17 детерминированных, 15 случайных и 8 хаотических ВР из репозитория [37].

Полученные в ходе исследования MSE-погрешности для всех двенадцати алгоритмов прогнозирования находятся в широком диапазоне [10⁻⁹ – 10⁹]. Поэтому для наглядного отображения результатов на рис. 7 приведены средние значения MSE в логарифмической шкале для каждой группы ВР. Можно отметить, что минимальная MSE-погрешность для алгоритма SARIMA близка к нулю и на рис. 7 находится в отрицательной плоскости.

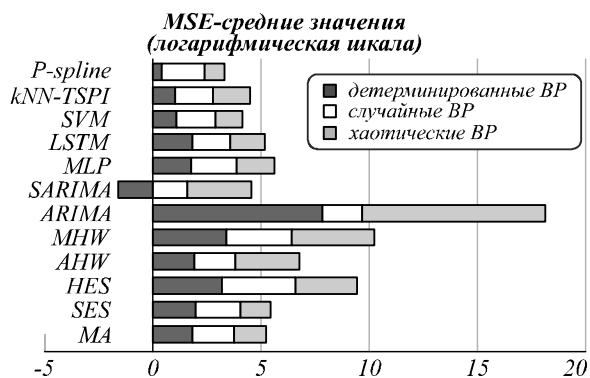


Рис. 7. Диаграмма средних значений MSE-погрешности

TU-коэффициент отражает эффективность алгоритма в сравнении с тривиальным повторением предыдущих значений ВР (наивная модель прогнозирования). В [10] использована следующая интерпретация этого коэффициента: если TU=1, то эффективность алгоритма такая же, как у наивного алгоритма; TU>1 – эффективность ниже; TU<1 – выше; а при TU≤0,55 алгоритм способен качественно выполнять прогноз. И для 17 детерминированных ВР на рис. 8 приведены соответствующие значения TU в каждом из трех указанных диапазонов.

Весьма важным при прогнозировании ВР является сохранение тенденции процесса. Значение показателя ROCID изменяется от 0 до 100: 100 – полное совпадение, 0 – отсутствие совпадения тенденции исходного и прогнозируемого ВР. И для всех типов детерминированных ВР показатели ROCID приведены для исследуемых двенадцати алгоритмов прогнозирования.

Процентный ROCID-параметр довольно нагляден для отображения тенденции. И в зависимости от прикладной задачи может быть установлено его пороговое значение.

Что касается предлагаемого в данной работе алгоритма прогнозирования на базе P-сплайна, то следует отметить его эффективность по MSE- и TU-показателям и более низкую эффективность в плане показателя тенденций ROCID.

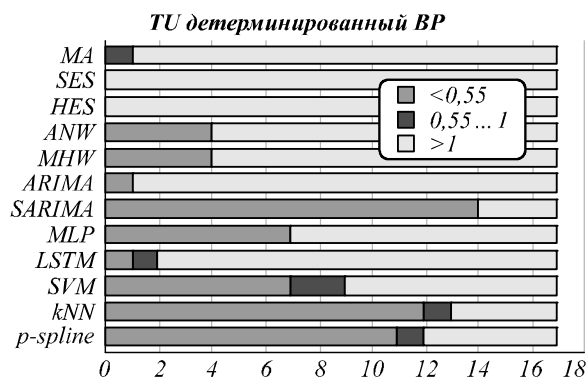


Рис. 8. Диаграмма распределения TU-коэффициента по диапазонам

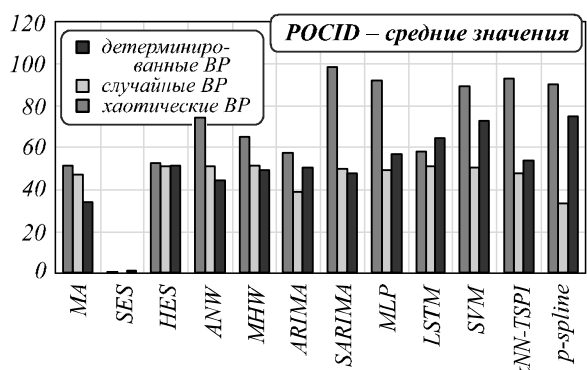


Рис. 9. Диаграмма средних значений ROCID-параметра

Что касается гибридных методов оптимизации, то можно указать весьма эффективные комбинации. Например, комбинация нейронные сети и регрессионные методы (ARIMA) дает ошибку прогноза около 1% [38]. А гибридный метод на основе kNN при использовании многомерных временных рядов имеет ошибку прогноза менее 1% [1]. Однако в последнем исследовании [1] глубина предыстории составляет 49900 точек ВР. А прогноз осуществляется на 1 шаг вперед по последним 100 точкам данных ВР. Подобный подход к прогнозированию делает его невозможным в масштабе реального времени. Предлагаемый нами гибридный метод, основанный на штрафных P-сплайнах, обеспечивает приемлемое соотношение качества прогнозирования и скорости прогнозирования. При этом тип и природа входных данных не анализируется и не требует предварительной обработки.

Заключение

В данной работе рассмотрен гибридный подход к прогнозированию ВР, воспроизводящий динамику процесса в форме рекуррентной модели штрафного P-сплайна.

Предложенная модель была протестирована на ВР-стоимости акций нескольких известных компаний. Сопоставление с результатами других гибридных методов (на основе ИНС и kNN-классификатора) показало, что точность прогноза предложенного метода хотя и несколько ниже (в среднем в 2 раза), но вполне допустима для использования в инженерных приложениях. Полученные исследования позволили сделать заключение, что для получения эффективных прогнозов необходимо производить адаптацию параметров модели. При использовании для адаптации генетического и иммунного алгоритмов оптимизации предпочтение отдано генетическому, скорость прогноза с его использованием в среднем выше в 1,5 раза при сопоставимой точности.

Также необходимо отметить важное свойство предлагаемой модели прогнозирования – малая глубина предыстории (порядка 10 точек ВР). А это вносит основной вклад в скорость прогноза и делает данную модель пригодной для использования в реальном времени.

Благодарности

Работа выполнена при поддержке гранта РФФИ (№ 18-07-01007).

Литература

1. **Yin, Y.** Forecasting traffic time series with multivariate predicting method / Y. Yin, P. Shang // *Applied Mathematics and Computation*. – 2016. – Vol. 291, Issue 1. – P. 266-278.
2. **Агафонов, А.А.** Анализ больших данных в геоинформационной задаче краткосрочного прогнозирования параметров транспортного потока на базе метода k ближайших соседей / А.А. Агафонов, А.С. Юмаганов, В.В. Мясников // *Компьютерная оптика*. – 2018. – Т. 42, № 6. – С. 1101-1111. – DOI: 10.18287/2412-6179-2018-42-6-1101-1111.
3. **Sbrana, G.** Short-term inflation forecasting: The M.E.T.A. approach / G. Sbrana, A. Silvestrini, F. Venditti // *International Journal of Forecasting*. – 2017. – Vol. 33. – P. 1065-1081.
4. **Montgomery, D.C.** Introduction to time series analysis and forecasting / D.C. Montgomery, C.L. Jennings, M. Kulahci. – Hoboken, New Jersey: John Wiley and Sons, Inc., 2015. – 443 p.
5. **Wang, H.** Time series feature learning with labeled and unlabeled data / H. Wang, Q. Zhangc, J. Wud, S. Panf, Y. Chene // *Pattern Recognition*. – 2019. – Vol. 89. – P. 55-66.
6. **Бокс, Дж.** Анализ временных рядов. Прогноз и управление / Дж. Бокс, Г. Дженкинс. – М.: Мир, 1974. – 406 с.
7. **Астахова, Н.Н.** Применение многоцелевой оптимизации для прогнозирования групп временных рядов / Н.Н. Астахова, Л.А. Демидова, Е.В. Никульчев // *Кибернетика и программирование*. – 2016. – № 5. – С. 175-190.
8. **Parmezan, A.** Metalearning for choosing feature selection algorithms in data mining: Proposal of a new framework / A. Parmezan, H. Lee, F. Wu // *Expert Systems with Applications*. – 2017. – Vol. 75. – P. 1-24.
9. **Чучуева, И.А.** Модель прогнозирования временных рядов по выборке максимального подобия : дис. ... канд. техн. наук : 05.13.18 / Чучуева Ирина Александровна. – М., 2012. – 155 с.
10. **Parmezan, A.** Evaluation of statistical and machine learning models for time series prediction: Identifying the state-

of-the-art and the best conditions for the use of each model / A. Parmezan, V. Souza, G. Batistaa // *Information Sciences*. – 2019. – Vol. 484. – P. 302-337.

11. **Запорожцев, И.Ф.** Краткосрочное прогнозирование пространственно-временной изменчивости океанографических характеристик методами анализа многомерных временных рядов : дис. ... канд. техн. наук : 05.13.18 / Запорожцев Иван Федорович. – Мурманск: 2016. – 183 с.
12. **Демидова, Л.А.** Классификация данных на основе SVM-алгоритма и алгоритма k-ближайших соседей / Л.А. Демидова, Ю.С. Соколова // *Вестник Рязанского государственного радиотехнического университета*. – 2017. – Т. 62. – С. 119-132.
13. **Hajirahimi, Z.** Hybrid structures in time series modeling and forecasting A review / Z. Hajirahimi, M. Khashei // *Engineering Applications of Artificial Intelligence*. – 2019. – Vol. 86. – P. 83-106.
14. **Lu, C.** Wavelet fuzzy neural networks for identification and predictive control of dynamic systems / C. Lu // *IEEE Transactions on Industrial Electronics*. – 2011. – Vol. 58, Issue 7. – P. 3046-3058.
15. **Аверкин, А.Н.** Гибридный подход для прогнозирования временных рядов на основании нейросети ANFIS и нечетких когнитивных карт / А.Н. Аверкин, С.А. Ярушев // *Международная конференция по мягким вычислениям и измерениям*. – 2017. – Т. 1. – С. 467-470.
16. **Chen, M.-Y.** A hybrid fuzzy time series model based on granular computing for stock price forecasting / M.-Y. Chen, B.-T. Chen // *Information Sciences*. – 2015. – Vol. 294. – P. 227-241.
17. **Rafiei, M.** Probabilistic forecasting of hourly electricity price by generalization of ELM for usage in improved wavelet neural network / M. Rafiei, T. Niknam, M.-H. Khooban // *IEEE Transactions on Industrial Informatics*. – 2017. – Vol. 13, Issue 1. – P. 71-79.
18. **Zhang, M.L.** A k-nearest neighbor based algorithm for multi-label classification / M.L. Zhang, Z.H. Zhou // *Proceedings of the 1st IEEE International Conference on Granular Computing*. – 2005. – P. 718-721.
19. **Chernoff, K.** Weighting of the k-nearest-neighbors / K. Chernoff, M. Nielsen // *Proceedings of the 20th IEEE International Conference on Pattern Recognition (ICPR)*. – 2010. – P. 666-669.
20. **Liu, H.** Noisy data elimination using mutual k-nearest neighbor for classification mining / H. Liu, S. Zhang // *Journal of Systems and Software*. – 2012. – Vol. 85, Issue 5. – P. 1067-1074.
21. **de Boor, C.** A practical guide to splines / C.A. de Boor. – New York: Springer-Verlag, 2001. – 348 p.
22. **Sharif, S.** Comparison of significant approaches of penalized spline regression (P-splines) / S. Sharif, S. Kamal // *Pakistan Journal of Statistics and Operation Research*. – 2018. – Vol. 14, Issue 2. – P. 289-303.
23. **Budakçi, G.** Extending fundamental formulas from classical B-splines to quantum B-splines / G. Budakçi, Ç. Dişibüyük, R. Goldman, H. Oruç // *Journal of Computational and Applied Mathematics*. – 2015. – Vol. 282. – P. 17-33.
24. **Eilers, P.** Twenty years of P-splines / P. Eilers, B. Marx, M. Durbán // *Statistics and Operations Research Transactions*. – 2015. – Vol. 39, Issue 2. – P. 149-186.
25. **Yang, L.** Adaptive penalized splines for data smoothing / L. Yang, Y. Hong // *Computational Statistics and Data Analysis*. – 2017. – Vol. 108. – P. 70-83.
26. **Kochegurova, E.A.** Current estimation of the derivative of a non-stationary process based on a recurrent smoothing

- spline / E.A. Kochegurova, E.S. Gorokhova // *Optoelectronics, Instrumentation and Data Processing*. – 2016. – Vol. 52, Issue 3. – P. 280-285.
27. **Кочегурова, Е.А.** Frequency analysis of recurrence variational P-splines / E.A. Kochegurova, A.I. Kochegurov, N.E. Rozhkova // *Optoelectronics, Instrumentation and Data Processing*. – 2017. – Vol. 53, Issue 6. – P. 591-598.
 28. **Martín, A.** EvoDeep: a new evolutionary approach for automatic Deep Neural Networks parametrization / A. Martín, R. Lara-Cabrera, F. Fuentes-Hurtado, V. Naranjo, D. Camacho // *Journal of Parallel and Distributed Computing*. – 2018. – Vol. 117. – P. 180-191.
 29. **Zhang, K.Q.** Research on a combined model based on linear and nonlinear features – A case study of wind speed forecasting / K.Q. Zhang, Z.X. Qu, Y.X. Dong, H.Y. Lu, W.N. Leng, J.Z. Wang, W.Y. Zhang // *Renewable Energy*. – 2019. – Vol. 130. – P. 814-830.
 30. **Пантелеев, А.В.** Методы глобальной оптимизации. Метаэвристические стратегии и алгоритмы / А.В. Пантелеев, Д.В. Метлицкая, Е.А. Алешина. – М.: Вузовская книга, 2013. – 244 с.
 31. **Гельфанд, И.М.** Принцип нелокального поиска в задачах автоматической оптимизации / И.М. Гельфанд, М.Л. Цетлин // *ДАН СССР*. – 1961. – Т. 137, № 2. – С. 295-298.
 32. **Коварцев, А.Н.** К вопросу об эффективности параллельных алгоритмов глобальной оптимизации функций многих переменных / А.Н. Коварцев, Д.А. Попова-Коварцева // *Компьютерная оптика*. – 2011. – Т. 35, № 2. – С. 256-261.
 33. **Bergstra, J.** Algorithms for hyper-parameter optimization / J. Bergstra, R. Bardenet, Y. Bengio, B. Kégl // *Proceedings of the 25th Annual Conference on Neural Information Processing Systems (NIPS)*. – 2011. – P. 1-9.
 34. **Меняйлов, Е.С.** Обзор и анализ существующих модификаций генетических алгоритмов // *Открытые информационные и компьютерные интегрированные технологии*. – 2015. – № 70. – С. 244-254.
 35. **Кочегурова, Е.А.** Some results of designing an IIR smoothing filter with p-splines / E.A. Kochegurova, I.Y. Khozhaev, S.V. Rybushkina // *International Review of Automatic Control*. – 2019. – Vol. 12, Issue 4. – P. 200-209.
 36. **Shcherbakov, M.V.** A survey of forecast error measures / M.V. Shcherbakov, A. Brebels, N.L. Shcherbakova, A.P. Tyukov, T.A. Janovsky, V.A. Kamaev // *World Applied Sciences Journal*. – 2013. – Vol. 24, Issue 24. – P. 171-176.
 37. **Parmezan, A.** ICMC-USP time series prediction repository [Electronical Resource] / A. Parmezan, G. Batista. – 2014. – URL: http://sites.labc.icmc.usp.br/aparmezan/publications/pdf/Repository_Parmezan_USP_2014_TSPR.pdf (request date 06.02.2020).
 38. **Babu, C.** A moving-average filter based hybrid ARIMA-ANN model for forecasting time series data / C. Babu, B. Reddy // *Applied Soft Computing*. – 2014. – Vol. 23. – P. 27-38.

Сведения об авторах

Кочегурова Елена Алексеевна, окончила Томский политехнический институт по специальности «Прикладная математика», доцент кафедры автоматизации и компьютерных систем. Область научных интересов: обработка информации в реальном времени, идентификация клавиатурного почерка, биоинспирированные методы. E-mail: kocheg@tpu.ru.

Репина Елизавета Юрьевна, студентка Томского политехнического университета по направлению «Информационные системы и технологии». Область научных интересов: прикладная теория фильтрации, биоинспирированные алгоритмы оптимизации, разработка web-приложений. Email: eyr4@tpu.ru.

Цехан Ольга Борисовна, окончила Белорусский государственный университет имени В.И. Ленина по специальности «Прикладная математика», заведующий кафедрой математического и информационного обеспечения экономических систем. Область научных интересов: математическая теория управления, методы прикладной математики, математические методы принятия решений. E-mail: tsekhan@grsu.by.

ГРНТИ: 27.41.17 28.23.19

Поступила в редакцию 13 ноября 2019 г. Окончательный вариант – 17 февраля 2020 г.

Hybrid approach for time series forecasting based on a penalty p-spline and evolutionary optimization.

E.A. Kochegurova¹, E.Yu. Repina¹, O.B. Tsekhan²

¹Tomsk Polytechnic University, Tomsk, Russia,

²Yanka Kupala State University of Grodno, Grodno, Belarus

Abstract

In this work, a hybrid-forecasting model is proposed. The model includes a recursive penalty P-spline with parameters adaptation based on evolutionary optimization algorithms. In short-term forecasting, especially in real-time systems, the urgent task is to increase the forecast speed without compromising its quality. High forecasting speed has been achieved by an economical computational scheme of a recurrent P-spline with a shallow depth of prehistory. When combined with the adaptation of some parameters of the P-spline, such an approach allows you to control the forecast accuracy.

Keywords: penalized spline, smoothing spline, digital filter, impulse infinite response (IIR filter), instrumental function, amplitude and phase-frequency response.

Citation: Kochegurova EA, RepinaEY, Tsekhan OB. Hybrid approach for time series forecasting based on a penalty p-spline and evolutionary optimization. *Computer Optics*. 2020; 44(5): 821-829. DOI: 10.18287/2412-6179-CO-667.

Acknowledgements: The work was funded by the Russian Foundation for Basic Research under grant #18-07-01007.

References

- [1] Yin Y, Shang P. Forecasting traffic time series with multivariate predicting method. *Appl Math Comput* 2016; 291(1): 266-278.
 - [2] Agafonov AA, Yumaganov AS, Myasnikov VV. Big data analysis in a geoinformatic problem of short-term traffic flow forecasting based on a K nearest neighbors method [In Russian]. *Computer Optics* 2018; 42(6): 1101-1111. DOI: 10.18287/2412-6179-2018-42-6-1101-1111.
 - [3] Sbrana G, Silvestrini A, Venditti F. Short-term inflation forecasting: The M.E.T.A. approach. *Int J Forecast* 2017; 33: 1065-1081.
 - [4] Montgomery DC., Jennings CL., Kulahci M. *Introduction to time series analysis and forecasting*. Hoboken, New Jersey: John Wiley and Sons Inc; 2015.
 - [5] Wang H, Zhang Q, Wud J, Panf S, Chene Y. Time series feature learning with labeled and unlabeled data. *Pattern Recognit* 2019; 89: 55-66.
 - [6] Box J, Jenkins G. *Time Series Analysis. Forecast and management*. San Francisco: CA Holden-Day; 1976.
 - [7] Astakhova NN, Demidova LA, Nikulchev EV. Application of multi-purpose optimization for forecasting time series groups [In Russian]. *Cybernetics and Programming* 2016; 5: 175-190.
 - [8] Parmezan A, Lee H, Wu F. Metalearning for choosing feature selection algorithms in data mining: Proposal of a new framework. *Expert Syst Appl* 2017; 75: 1-24.
 - [9] Chuchueva AI. A model for predicting time series for a sample of maximum similarity [In Russian]. The thesis for the Candidate's degree in Technical Sciences. Moscow: 2012.
 - [10] Parmezan A, Souza V, Batistaa G. Evaluation of statistical and machine learning models for time series prediction: Identifying the state-of-the-art and the best conditions for the use of each model. *Inf Sci* 2019; 484: 302-337.
 - [11] Zaporozhtsev, I.F. Short-term forecasting of spatio-temporal variability of oceanographic characteristics by methods multidimensional time series analysis [In Russian]. The thesis for the Candidate's degree in Technical Sciences. Murmansk: 2016.
 - [12] Demidova LA, Sokolova YS. Data classification based on the SVM algorithm and the k-nearest neighbor algorithm [In Russian]. *Bulletin of the Ryazan State Radio Engineering University* 2017; 62: 119-132.
 - [13] Hajirahimi Z, Khashei M. Hybrid structures in time series modeling and forecasting: A review. *Eng Appl Artif Intell* 2019; 86: 83-106.
 - [14] Lu C. Wavelet fuzzy neural networks for identification and predictive control of dynamic systems. *IEEE Trans Ind Electron* 2011; 58(7): 3046-3058.
 - [15] Averkin AN, Yarushev S. Hybrid approach for time series forecasting based on ANFIS and Fuzzy Cognitive Maps. *Proc XXth IEEE Int Conf Soft Computing and Measurements (SCM 2017)* 2017: 379-381.
 - [16] Chen MY, Chen BT. A hybrid fuzzy time series model based on granular computing for stock price forecasting. *Inf Sci* 2015; 294: 227-241.
 - [17] Raffie M, Niknam T, Khooban MH. Probabilistic forecasting of hourly electricity price by generalization of ELM for usage in improved wavelet neural network. *IEEE Trans Ind Inf* 2017; 13(1): 71-79.
 - [18] Zhang ML, Zhou ZH. A k-nearest neighbor based algorithm for multi-label classification. *Proc 1st IEEE Int Conf on Granular Computing* 2005: 718-721.
 - [19] Chernoff K, Nielsen M. Weighting of the k-Nearest-Neighbors. *Proc 20th IEEE Int Conf Pattern Recognit (ICPR)* 2010: 666-669.
 - [20] Liu H, Zhang S. Noisy data elimination using mutual k-nearest neighbor for classification mining. *J Syst Softw* 2012; 85(5): 1067-1074.
 - [21] de Boor C. *A practical guide to splines*. New York: Springer-Verlag; 2001.
 - [22] Sharif S, Kamal S. Comparison of significant approaches of penalized spline regression (P-splines). *Pakistan J Stat Oper Res* 2018; 14(2): 289-303.
 - [23] Budakçı G, Dişibüyük Ç, Goldman R, Oruç H. Extending fundamental formulas from classical B-
-

-
- splines to quantum B-splines. *J Comput Appl Math* 2015; 282: 17-33.
- [24] Eilers P, Marx B, Durbán M, Durbán M. Twenty years of P-splines. *Statistics and Operations Research Transactions* 2015; 39(2): 149-186.
- [25] Yang L, Hong Y. Adaptive penalized splines for data smoothing. *Comput Stat Data Anal* 2017; 108: 70-83.
- [26] Kochegurova EA, Gorokhova ES. Current Estimation of the Derivative of a Non-stationary Process Based on a Recurrent Smoothing Spline. *Optoelectronics, Instrumentation and Data Processing* 2016; 52(3): 280-285.
- [27] Kochegurova EA, Kochegurov AI, Rozhkova NE. Frequency analysis of recurrence variational P-splines. *Optoelectronics, Instrumentation and Data Processing* 2017; 53(6): 591-598.
- [28] Martín A, Lara-Cabrera R, Fuentes-Hurtado F, Naranjo V, Camacho D. EvoDeep: a new evolutionary approach for automatic Deep Neural Networks parametrization. *J Parallel Distrib Comput* 2018; 117: 180-191.
- [29] Zhang KQ, Qu ZX, Dong YX, Lu HY, Leng WN, Wang JZ, Zhang WY. Research on a combined model based on linear and nonlinear features – A case study of wind speed forecasting. *Renew Energy* 2019; 130: 814-830.
- [30] Panteleev AV, Metlitskaya DV, Aleshina EA. Methods of global optimization. Metaheuristic strategies and algorithms [In Russian]. Moscow: "VUZovskaya Kniga" Publisher; 2013.
- [31] Gelfand IM, Tsetlin ML. The principle of nonlocal search in automatic optimization problems. *Doklady Akademii Nauk SSSR* 1961; 137(2): 295-298.
- [32] Kovartsev AN, Popova-Kovartseva DA. On efficiently of parallel algorithms for global optimization of functions of several variables. *Computer Optics* 2011; 35(2): 256-261.
- [33] Bergstra J, Bardenet R, Bengio Y, Kégl B. Algorithms for hyper-parameter optimization. *Proc 25th Annual Conf on Neural Information Processing Systems (NIPS)* 2011: 1-9.
- [34] Menyailov ES. Review and analysis of existing modifications of genetic algorithms [In Russian]. *Open Information and Computer Integrated Technologies* 2015; 70: 244-254.
- [35] Kochegurova EA, Khozhaev IY, Rybushkina SV. Some results of designing an IIR smoothing filter with p-splines. *Int Rev Autom Control* 2019; 12(4): 200-209.
- [36] Shcherbakov MV, Brebels A, Shcherbakova NL, Tyukov AP, Janovsky TA, Kamaev VA. A survey of forecast error measures. *World Appl Sci J* 2013; 24(24): 171-176.
- [37] Parmezan A, Batista G. ICMC-USP time series prediction repository. 2014. Source: (http://sites.labc.icmc.usp.br/aparmezan/publications/pdf/Repository_Parmezan_USP_2014_TSPR.pdf).
- [38] Babu C, Reddy BA moving-average filter based hybrid ARIMA-ANN model for forecasting time series data. *Appl Soft Comput* 2014; 23: 27-38.
-

Authors' information

Elena Alekseevna Kochegurova, Candidate of Technical Sciences, associate professor of Information Technology department, Tomsk Polytechnic University. Research interests are information processing in real time, keyboard dynamics, bio-inspected techniques. E-mail: kocheg@mail.ru.

Elizaveta Yuryevna Repina, a student of Tomsk Polytechnic University (bachelor program Information systems) Her research interests include filter theory, bio-operated techniques, web application developmen. Email: evr4@tpu.ru.

Olga Borisovna Tsekhan, Candidate of Physics and Mathematics, head of Mathematical and Information Support of Economic Systems department, Grodno State University. Research interests: mathematical control theory, applied mathematics methods, mathematical decision-making methods. E-mail: tsekhan@grsu.by.

Received November 13, 2019. The final version – February 17, 2020.
