DOI: 10.17323/2587-814X.2021.4.22.35

К вопросу о восстановлении символьных последовательностей, кодирующих зашумленные периодические функции

Г.Н. Жукова [a]
E-mail: galinanzhukova@gmail.com

М.В. Ульянов [b,c]
E-mail: muljanov@mail.ru

[a] Национальный исследовательский университет «Высшая школа экономики»
Адрес: 101000, г. Москва, ул. Мясницкая, д. 20

[b] Институт проблем управления им. В.А. Трапезникова Российской академии наук
Адрес: 117997, г. Москва, ул. Профсоюзная, д. 65

[c] Московский государственный университет им. М.В. Ломоносова
Адрес: 119991, г. Москва, Ленинские горы, д. 1

Аннотация

В бизнес-информатике, одним из предметов исследования которой является анализ данных о процессах в прикладных предметных областях, возникают задачи качественного анализа. Такого рода задачи возникают, например, при качественном исследовании лог-файлов бизнес-процессов, при анализе и прогнозировании временных рядов, и других процессов различной природы. Достаточно часто для представления информации об исследуемых процессах в методах качественного анализа используется символьное кодирование, позволяющее снять излишнюю детализацию числовых описаний. Актуальность данного исследования связана с тем, что при работе с исходными данными исследователи зачастую сталкиваются с наличием шумов и искажений в исходных данных, что существенно затрудняет решение задач качественного анализа. При работе с символьными представлениями исследуемых процессов, которые достаточно часто имеют периодический характер, мы наблюдаем шумы удаления, вставки и замены символов, усложняющие решение задачи определения и анализа периодичности. В статье рассматривается задача восстановления периодических символьных последовательностей, полученных кодированием по отсчетам периодических функций и искаженных шумами вставки, замены и удаления символов. В качестве конкретного примера синтетических данных временных рядов рассматриваются тригонометрические функции. Для кодирования тригонометрических функций используются алфавиты различных мощностей с различной детализацией интервалов отсчетов по модельному времени. В статье представлено экспериментальное исследование зависимости характеристик качества метода восстановления периода и периодически повторяющегося фрагмента, ранее предложенного авторами и усовершенствованного в данном исследовании. Для алфавитов разной мощности при фиксированных интервалах отсчетов по модельному времени приводятся доля последовательностей с удовлетворительно восстановленным

периодом и относительная погрешность определения длины периода. Качество восстановления периодически повторяющегося фрагмента оценивается отношением редакционного расстояния от восстановленной периодической последовательности до исходной последовательности, искаженной шумами.

Ключевые слова: символьная последовательность; мощность алфавита; периодическая последовательность; зашумленная последовательность; шум вставки; шум удаления; шум замены.

Цитирование: Жукова Г.Н., Ульянов М.В. К вопросу о восстановлении символьных последовательностей, кодирующих зашумленные периодические функции // Бизнес-информатика. 2021. Т. 15. № 4. С. 22–35. DOI: 10.17323/2587-814X.2021.4.22.35

Введение

Одним из предметов научных исследований в бизнес-информатике является анализ данных о процессах, регистрируемых в прикладных предметных областях [1]. Веер возникающих задач анализа данных включает в себя и задачи качественного анализа. Они возникают, например, при качественном исследовании лог-файлов бизнес-процессов [1] и связаны, например, с определением соответствия лог-файла модели процесса [2]. Одним из способов представления информации об исследуемых процессах в научных исследованиях является их представление в виде временных рядов. При этом значительная доля задач качественного анализа относится как раз к исследованиям в области временных рядов, и связана как с анализом их периодичности, так и в целом с проблематикой их комплексного анализа и прогнозирования [3–8]. Аналогичные качественные задачи возникают и при анализе других процессов различной природы, представленных временными рядами, например, при экологическом мониторинге и прогнозировании экологических изменений [9].

Наблюдаемые значения исследуемого процесса, являющиеся элементам временного ряда достаточно часто подвержены воздействию случайных искажений, вызванных внешними факторами. «Значения, сопоставляемые элементам получающегося ряда, содержат и ошибки измерения и, в общем случае, подвержены случайным внешним воздействиям. В дальнейшем такого рода ошибки измерений и результаты внешних воздействий трактуются, как шум» [10].

Укажем на некоторые публикации из области бизнес-информатики и менеджмента, в которых так или иначе обсуждается и/или нивелируется влияние шумов на результаты прогнозирования временных рядов. Так авторы отмечают, что при создании моделей и прогнозировании производства и потребления электроэнергии шумы в исходных данных оказывают влияние на прогностическую силу моделей и качество прогнозов [11–13]. Стохастическое прогнозирование рисков в бизнесе [14], в том числе риска операционной прибыли фирм, существенно опирается на предположение о неполноте и зашумленности данных. Авторы работы [15], указывают, что при исследовании поведения клиентов банка методом кластеризации временных рядов учитываются ошибки и разброс исходных данных. В работе [16] при оценке эффективности прогнозирования потоков пассажирских авиаперевозок с использованием множественных показателей ошибок, в частности показано, что шумы исходных данных влияют на качество прогнозирования. Авторы работы [17] пытаются повысить качество прогнозирования туристического спроса с помощью методов глубокого обучения совместно с обработкой изображений визуализации временных рядов, и снизить тем самым влияние ошибок исходных данных на результаты прогнозирования.

Из других предметных областей отметим, например, микроэлектронику, где размытости изображений, связанные с шумами, сильно влияют на качество прогнозирования [18], и наблюдения за биосферой, где авторы отмечают, что биометрические данные (в широком понимании, как данные наблюдения за биосферой) часто имеют пропуски наблюдений, выбросы и разрывы [19]. Для еще одной предметной области – дистанционного зондирования Земли из космоса, приведем цитату из [20]: «Использование временных рядов спутниковых данных для мониторинга земной поверхности

сопряжено с проблемой учета разного рода мешающих факторов, приводящих к частичной потере или искажению информации о динамике спектрально-отражательных характеристик объектов наблюдения. К таким факторам относятся непрозрачные в видимом и ближнем ИК диапазоне атмосферная дымка и облачность, тени от нее, а также шумы прибора».

В целях решения задач качественного анализа, данные временных рядов подвергаются символьному кодированию, позволяющему снять излишнюю детализацию числовых описаний [21–23]. При этом описание элементов временного ряда или шагов бизнес-процесса кодируется словом над конечным алфавитом, которое и является объектом дальнейшего исследования. Более того, такое кодирование является актуальным при качественном анализе больших данных. Это связано с тем, что высокая точность числовых представлений элементов временных рядов приводит не только к неоправданно большим объемам информации, но и трудоемким вычислениям, не улучающим качество получаемых результатов [22, 23].

Очевидно, что при исследовании временных рядов работа с зашумленными данными вызывает значительные трудности. Это приводит к формулировке задачи устранения шума. Для шумоподавления в числовых данных временных рядов используются различные методы сглаживания, такие как методы скользящего среднего, экспоненциального сглаживания и др. [10]. Однако эти методы не применимы при работе с зашумленными символьными последовательностями.

В аспекте символьного кодирования возникающие ошибки, трактуемые как шум, приводят к тому, что в символьных последовательностях возникают шумы вставки, удаления и замены символов. Так, например, ошибки, связанные с настройкой измерительных приборов, неточности при ручном вводе данных и случайные описки или преднамеренное искажение значений отдельных показателей [19], приводят к шуму замены. Ошибки регистрации являются источником шума удаления и шума вставки, эти шумы могут так же возникать и при подготовке данных. Отметим также, что методы выявления периодичности, применяемые для числовых последовательностей [24], неприменимы при работе с символьными представлениями.

Актуальность данного исследования связана с тем, что шумы снижают эффективность анализа временных рядов. В связи с этим в статье рассматривается задача восстановления периодических символьных последовательностей, полученных кодированием по отсчетам периодических функций и искаженных шумами вставки, замены и удаления символов. В качестве конкретного примера синтетических данных временных рядов рассматриваются тригонометрические функции. Для кодирования тригонометрических функций используются алфавиты различных мощностей с различной детализацией интервалов отсчетов по модельному времени.

В статье представлено экспериментальное исследование зависимости характеристик качества метода восстановления периода и периодически повторяющегося фрагмента, ранее предложенного авторами работы [25] и усовершенствованного в данном исследовании. Для алфавитов разной мощности при фиксированных интервалах отсчетов по модельному времени приводятся доля последовательностей с удовлетворительно восстановленным значением длины периода и относительная погрешность определения длины периода. Качество восстановления периодически повторяющегося фрагмента оценивается отношением редакционного расстояния от восстановленной периодической последовательности до исходной последовательности, искаженной шумами.

1. Терминология и обозначения

Далее мы будем использовать обозначения, введенные нами в статье [25], описывающей метод восстановления периодической символьной последовательности.

Обозначим $\sum^{\sigma}$ алфавит мощности $\sigma = \left|\sum\right| \geq 2$. Будем называть словом длины n символьную последовательность $q^{\sigma} = s_1, s_2, ..., s_n$ над конечным алфавитом $\sum^{\sigma}$, где s –произвольный символ из $\sum^{\sigma}$, а подсловом или фрагментом слова q^{σ} назовем любую последовательность символов $s_k, s_{k+1}, ..., s_{l-1}, s_l$, $1 \leq k \leq l \leq n$.

Мы рассматриваем периодические символьные последовательности с периодом p. Во избежание неоднозначности понимания, будем называть периодом p длину повторяющегося подслова (фрагмента), а часть периодической последовательности длины p назовем периодически повторяющимся фрагментом. Если не оговорено иное, периодически повторяющимся фрагментом будем называть фрагмент длины p, начинающийся с первого символа периодического слова.

Введем следующие обозначения: $q^{\sigma}(m, p)$ – периодическое слово, содержащее $m \geq 8$ повторяющихся фрагментов длины p; $\tilde{q}^{\sigma}(m, p)$ – слово над тем же алфавитом, что и $q^{\sigma}(m, p)$, но с внесенным в него шумом; $\overline{q}^{\sigma}(m, p)$ – периодическое слово, полученное на основе анализа $\tilde{q}^{\sigma}(m, p)$ с помощью алгоритма из работы [25] с предлагаемыми в данной статье улучшениями.

Заметим, что слово $\overline{q}^{\sigma}(m, p)$ имеет такую же длину, как $\tilde{q}^{\sigma}(m, p)$ и служит аппроксимацией периодического слова $q^{\sigma}(m, p)$.

2. Постановка задачи

Пусть наблюдается некоторый непрерывный во времени, периодический процесс $g(t)$ от момента t_0 до момента $t_0 + n\Delta t$ в ходе которого через равные промежутки времени Δt измеряется величина $g(t_i)$, $i = 1, n$. Разбивая диапазон измеренных значений величины $g(t)$ на σ равных полусегментов и кодируя значения $g(t_i)$ символами алфавита $\sum^{\sigma}$, мы получаем символьную последовательность $q^{\sigma} = s_1, s_2, ..., s_n$ над этим алфавитом. Заметим, что некоторые символы алфавита $\sum^{\sigma}$ могут при этом не присутствовать в полученной символьной последовательности. Будем считать, что период p наблюдаемой функции g кратен длине интервала Δt между последовательными измерениями, т.е. $p = r\,\Delta t$, где r целое число, благодаря чему при кодировании периодической функции $g(t)$ на сегменте в m периодов получается периодическая символьная последовательность, также содержащая m периодов, при этом $n = mp = mr\,\Delta t$.

Теперь внесем случайные искажения в измеренные значения функции $g(t)$, будем рассматривать шумы вставки, замены и удаления. Вставка нового значения соответствует некоторому сбою в измерениях, когда между запланированными измерениями произошло внеочередное измерение, удаление означает потерю значения при вводе или передаче измерений, заменой считается неправильное измерение или сознательное искажение данных.

После внесения всех искажений мы получаем последовательность с шумом, которую кодируем в том же алфавите $\sum^{\sigma}$ и рассматриваем задачу восстановления периодической символьной последовательности по заданной последовательности с шумом.

В данной статье мы рассматриваем символьные последовательности, полученные кодированием (в отсчетах через Δt) последовательности значений непрерывной периодической функции (в частности, $\sin(t)$) на сегменте с длиной не менее восьми полных периодов ($m \geq 8$) в условиях их искажения шумами различных типов.

При кодировании непрерывной функции существенную роль играет выбор мощности алфавита. Мы кодируем каждую рассматриваемую функцию символами алфавитов мощности от 10 до 60 с шагом 10. Несмотря на то, что при кодировании с помощью разбиения диапазона значений кодируемой функции на равные полусегменты некоторые символы алфавита могут не встретиться в неискаженной периодической символьной последовательности, такие символы могут наблюдаться в последовательности, полученной кодированием искаженной последовательности измеренных значений функции.

Постановка задачи: изучить влияние мощности алфавита, вида функции и уровня шума на качество восстановления периодической символьной последовательности, полученной кодированием значений периодической функции в условиях ее искажения шумами вставки, замены и удаления значений.

3. Кодирование непрерывной периодической функции

Рассмотрим функцию $\sin(t)$ на сегменте $[0 - 16\pi]$, содержащем 8 полных периодов функции $\sin(t)$. Для построения числовой последовательности с периодом p разбиваем интервал $[0 - 16\pi]$ на $8p$ равных полусегментов длины $\Delta t = \frac{2\pi}{p}$ и вычисляем значение функции $\sin(t)$ в середине каждого полусегмента, получая последовательность из $8p$ вещественных чисел $y_1, y_2, ..., y_{8p}$.

В целях символьного кодирования полученных значений в алфавите мощности σ разбиваем область значений синуса – сегмент $[-1, 1]$ на σ последовательных полусегментов $I_1, I_2, ..., I_{\sigma}$ равной длины; при этом число полусегментов равно мощности алфавита. Ставим в соответствие полусегментам $I_1, I_2, ..., I_{\sigma}$ символы алфавита $\sum^{\sigma}$. Каждое число в последовательности $y_1, y_2, ..., y_{8p}$ кодируется символом, соответствующем тому полусегменту I_j, в которое это число попадает, в результате мы получаем символьную последовательность $q^{\sigma}(m, p)$. Заметим, что некоторые символы алфавита $\sum^{\sigma}$ могут не встретиться в этой последовательности.

4. Метод определения периода по периодической последовательности с шумом

Для понимания предлагаемых улучшений метода построения периодически повторяющегося фрагмента приведем краткое описание метода определения периода [25, 26].

Метод предполагает подсчет числа всех подслов длины $k = 10$ в зашумленной последовательности $\tilde{q}^{\sigma} = s_1, s_2, ..., s_n$. Подслова длины 10 в $\tilde{q}^{\sigma}$ берутся со сдвигом на один символ, т.е. рассматриваются подслова $s_1, s_2, ..., s_{10}$, $s_2, s_3, ..., s_{11}$ и т.д. Те подслова, которые встретились не менее 3-х раз, составляют множество R. Каждому подслову из R сопоставляется список номеров позиций символов последовательности $\tilde{q}^{\sigma}$, начиная с которых данное подслово входит в $\tilde{q}^{\sigma}$. Так, в последовательности «*abcabcdabcdeabcabcdabcde*» посдлово «*abcabcdabc*» входит, начиная с номеров позиций 1 и 13.

Далее для каждого подслова длины $k = 10$ подсчитываются разности между номерами последовательных вхождений, после чего из таких разностей строится множество Ω. Каждому элементу множества Ω сопоставлено количество раз, когда наблюдалась такая разность номеров последовательных вхождений первых символов подслов длины 10. Например, в слове «*abcabcdabcdeabcabcdabcde*» разность $r = 13 - 1 = 12$ наблюдается 3 раза (для слов «*abcabcdabc*», «*bcabcdabcd*», «*cabcdabcde*»). За счет внесенного шума разности не всегда равны или даже кратны периоду, но большинство разностей принимает значения, близкие к периоду или к числу, кратному периоду. В связи с этим для каждого значения разности подсчитывается, сколько раз встречалась в Ω разность, близкая к ней. В наших экспериментах близость определялась как попадание в интервал ±20%. Анализ полученных разностей и позволяет получить оценку периода для неизвестной строго периодической последовательности [25].

5. Усовершенствованный метод построения периодически повторяющегося фрагмента

Решением, которое доставляет алгоритм из работы [25] в части построения периодически повторяющегося фрагмента является подслово $\tilde{f}^{\sigma} = s_1, s_2, \ldots, s_{\tilde{p}}$ длины $\tilde{p}$ анализируемого слова $\tilde{q}^{\sigma}(m, p)$, минимизирующее (на множестве полученных вариантов фрагментов) редакционное расстояние от символьной последовательности длины $|\tilde{q}^{\sigma}(m, p)|$, построенной из данного фрагмента, до последовательности $\tilde{q}^{\sigma}(m, p)$.

Построение аппроксимирующего периодически повторяющегося фрагмента $\tilde{f}^{\sigma}$ в работе [25] проводится путем разбиения искаженной последовательности на последовательные подслова длины $\tilde{p}$ (последнее подслово длины меньше $\tilde{p}$ не учитывалось) и выбора такого из них, у которого редакционное расстояние до одного из оставшихся подслов минимально. В случае, если таких подслов несколько, выбирается первое подслово с минимальным редакционным расстоянием до другого подслова.

Предлагаемый нами и описываемый далее метод пытается улучшить фрагмент $\tilde{f}^{\sigma}$ с целью получения меньшего значения редакционного расстояния между искаженной и аппроксимирующей периодической последовательностями. Предлагаемое улучшение достигается за счет использования информации об определенных ранее частотах подслов длины $k = 10$, наблюдаемых в $\tilde{q}^{\sigma}(m, p)$.

Вначале на основе фрагмента $\tilde{f}^{\sigma}$ строится слово $\tilde{f}_4^{\sigma}$, содержащее фрагмен $\tilde{f}^{\sigma}$, записанный четыре раза подряд. Затем каждое подслово фрагмента $\tilde{f}_4^{\sigma}$ длины $k = 10$ ($s_1, s_2, ..., s_{10}$; $s_2, s_3, ..., s_{11}$ и т.д.) проверяется на встречаемость в слове $\tilde{q}^{\sigma}(m, p)$ не менее 3 раз, т.е. на принадлежность множеству R. Первое из проверяемых подслов $\omega_0 = s_{t+1}, s_{t+2}, ..., s_{t+10}$, принадлежащее множеству R, становится стартовым подсловом, с него будет начинаться улучшенный периодический фрагмент. Если такого подслова не нашлось, периодический фрагмент остается без улучшения.

После того, как найдено подслово ω_0, последовательно просматриваются подслова ω из $\tilde{f}_4^{\sigma}$ со сдвигом на один символ, т.е. $\omega = s_{t+h}, s_{t+h+1}, ..., s_{t+h+9}$, $h = 2, 3, ...$, выполняется проверка принадлежности этих подслов множеству R. Если подслово принадлежит R, то соответствующее подслово слова $\tilde{f}_4^{\sigma}$ остается без изменений, иначе начинаем подсчет последовательных подслов (со сдвигом на один символ) не входящих в R.

Поскольку длина ω равна 10, то пока не наберется 10 последовательных подслов, не входящих в R, слово $\tilde{f}_4^{\sigma}$ остается без изменений. Если после менее 10 последовательных ω, не входящих в R очередное ω оказалось входящим в R, то обнуляется счетчик последовательных подслов, не входящих в R. Если же в R не нашлось 10 последовательных подслов ω, но одиннадцатое ω входит в R то начинается поиск

подслова ω^R в R, такого, что первые m символов ω^R совпадают с последними m символами ω^- – последнего из рассматриваемых подслов, входящего в R, $3 \le m \le 9$. Значения m берутся от 9 до 3, т.е. сначала пытаемся найти ω^R, у которого первые 9 символов совпадают с последними 9 символами ω^-, если такое ω^R найдено, то в $\tilde{f}_4^\sigma$ заменяем символ, следующий за последним символом подслова ω^-, на последний символ ω^R.

Если при $m = 9$ подслово ω^R не найдено, переходим к $m = 8$ и т.д. до $m = 3$. Пусть при некотором m от 3 до 9 нашлось ω^R, тогда в $\tilde{f}_4^\sigma$ заменяем на символ, следующий за последним символом подслова ω^-, на $(m + 1)$-й символ ω^R. Если ни для какого m от 3 до 9 не нашлось ω^R, то фрагмент $\tilde{f}_4^\sigma$ остается без улучшения.

В какой-то момент или будут просмотрены $2\tilde{p}$ последовательных подслов ω и не встретится ни разу фрагмент ω_0 или очередное подслово ω совпадет с ω_0. В первом случае в качестве улучшенного фрагмента $\tilde{f}^\sigma$ берем первые $\tilde{p}$ символов $\tilde{f}_4^\sigma$ начиная с первого символа ω_0 (т.е. из $\tilde{f}_4^\sigma$ вырезаем ω_0 и следующие за ним в $\tilde{f}_4^\sigma$ символы, всего $\tilde{p}$ символов). Во втором случае $\tilde{f}^\sigma$ состоит из символов $\tilde{f}_4^\sigma$, начиная с первого вхождения ω_0 до второго ω_0, при этом проверяется возможность того, что в конце периодического фрагмента за счет шума вставки или удаления потерялся или добавился один символ.

Проверка делается следующим образом. Если после очередного ω, входящего в R в $\tilde{f}_4^\sigma$ было 9 последовательных подслов, не входящих в R, после которых нашлось слово из R и это слово совпало с ω_0, то возможно, имел место шум удаления. Для проверки в множестве R ищем ω^R, у которого первые m символов совпадают с последними m символами ω. Если такое слово ω^R найдено и у него символы, начиная с $(m + 2)$-го совпадают с первыми символами , ω_0 то в конец $\tilde{f}^\sigma$ добавляем $(m + 1)$-й символ ω^R.

Кроме того, если после очередного ω, входящего в R в $\tilde{f}_4^\sigma$ было 9 последовательных подслов, не входящих в R, после которых нашлось слово из R и это слово совпало с ω_0, возможно, был шум вставки, от которого пытается избавиться так. Если первый символ ω_0 совпадает с последним символом ω, то в $\tilde{f}^\sigma$ не включаем последний символ перед вторым вхождением ω_0 в $\tilde{f}_4^\sigma$.

Для того, чтобы одновременно с поиском улучшенного фрагмента $\tilde{f}^\sigma$ уточнить значение периода, просматриваются $2\tilde{p}$ последовательных подслов ω слова $\tilde{f}_4^\sigma$, пока на очередном шаге среди ω не встретится слово ω_0, тогда улучшенным фрагментом считается подслово улучшенного $\tilde{f}_4^\sigma$, начиная с первого вхождения ω_0 до второго, если только при этом период получается не менее 3. Если за счет изменений в $\tilde{f}_4^\sigma$ в процессе улучшения не нашлось еще одного ω_0, то после окончания процесса улучшения происходит еще один просмотр $2\tilde{p}$ последовательных подслов ω слова $\tilde{f}_4^\sigma$ начиная с первого вхождения ω_0, при этом если нашлось второе вхождение ω_0, то улучшенным фрагментом считается подслово улучшенного $\tilde{f}_4^\sigma$, начиная с первого вхождения ω_0 до второго. Если второе вхождение ω_0 не нашлось, то в качестве улучшенного фрагмента берутся $\tilde{p}$ последовательных символов улучшенного $\tilde{f}_4^\sigma$, начиная с первого вхождения ω_0.

В случае, если удалось улучшить периодический фрагмент, рассматриваются его циклические сдвиги, и в качестве окончательного фрагмента берется такой сдвиг, у которого минимально редакционное расстояние до начала зашумленной последовательности, т.е. ее первых $\tilde{p}^*$ символов, где $\tilde{p}^*$ – длина улучшенного периодического фрагмента.

Если в результате улучшения фрагмента получено подслово длины больше трех, и редакционное расстояние от периодической последовательности, построенной повторением улучшенного фрагмента, до зашумленной меньше, чем в случае $\tilde{f}^\sigma$ до улучшения, в качестве аппроксимации периодически повторяющегося фрагмента используем улучшенный $\tilde{f}^\sigma$, иначе – исходный.

6. Оценка качества восстановления периода

Оценку точности определения периода и качества восстановления периодического фрагмента будем проводить отдельно. Пусть период последовательности до внесения искажений был равен p, а наш алгоритм определил, что период равен $\overline{p} = \overline{p}\left(\overline{q}^\sigma\right)$, тогда точность δ нахождения периода определяется как

$$\delta = \frac{|p - \overline{p}|}{p}. \qquad (1)$$

Для периодической последовательности, полученной кодированием периодической функции на восьми периодах, мы получаем серию из 100 случайно зашумленных последовательностей, для каждой из них определяем период и находим значение δ, затем вычисляем среднее значение δ и медиану выборки по 100 зашумленным последовательностям.

Качество восстановления периодического фрагмента будем оценивать путем вычисления отношения редакционного расстояния $d\left(\overline{q}^{\sigma}, q^{\sigma}(m,p)\right)$ между восстановленной и исходной периодической символьными последовательностями, к длине исходной периодической последовательности [27]. Обозначим это отношение $\varepsilon\left(\overline{q}^{\sigma}, q^{\sigma}\right)$:

$$\varepsilon\left(\overline{q}^{\sigma}, q^{\sigma}\right) = \frac{d\left(\overline{q}^{\sigma}, q^{\sigma}(m,p)\right)}{mp}. \qquad (2)$$

При таком подходе хорошую оценку будет давать случай, когда полученный алгоритмом период в два-три раза больше исходного, но при этом периодический фрагмент близок к оригиналу, повторенному нужное число раз.

Кроме того, проводилось сравнение исходной и восстановленной последовательностей с зашумленной, в этом случае от каждой периодической последовательности для сравнения были взяты первые n символов, где n – длина зашумленной последовательности.

7. Схема вычислительного эксперимента

В вычислительном эксперименте проводилось исследование предложенного в работах [25, 26] и улучшенного в данной статье метода на следующих функциях:

- $\sin(t)$ на сегменте $[0 - 16\pi]$;
- $\left(t - 8\left\lfloor\frac{t}{8}\right\rfloor + 1\right)\sin(\pi t)$ на сегменте $[0 - 64]$;
- $\left(t - 16\left\lfloor\frac{t}{16}\right\rfloor + 1\right)\sin(\pi t)$ на сегменте $[0 - 128]$;
- $\left|7 - t + 16\left\lfloor\frac{t}{16}\right\rfloor\right|\sin(\pi t)$ на сегменте $[0 - 128]$,

где $\lfloor ... \rfloor$ означает целую часть числа.

Таком образом, для всех функций рассматривался сегмент аргумента, содержащий 8 полных периодов. Кодирование функций производилось в алфавитах $\sum^{\sigma}$ мощности σ от 10 до 60 с шагом 10. Область значений функции разбивалась на σ интервалов, при этом каждый интервал I_j кодировался символом s_j алфавита $\sum^{\sigma}$.

Значения периода p выбирались равными 20, 30 и 50, при этом один период функции разбивался на p равных полусегментов, затем для каждого полусегмента Δt_i находилось значение функции в середине полусегмента, и по этому значению определялся полусегмент I_j, которому это значение принадлежит, после чего на i-ую позицию кодирующего слова записывался символ s_j, кодирующий полусегмент I_j. Так были получены периодические символьные последовательности в алфавитах мощности от 10 до 60.

На основе построенных периодических последовательностей были получены случайные зашумленные последовательности, по 100 последовательностей на каждую чисто периодическую. Шум вносился в соответствии с ранее предложенной нами вероятностной моделью шумов для периодических символьных последовательностей [28].

В первой серии экспериментов вносили шум с общим уровнем 5%, при этом уровни шума вставки, замены и удаления принимали значения от 1 до 5% с шагом 1%, так что сумма уровней шума равна 5%. Во второй серии был внесен шум, равномерно распределенный по типам, т.е. уровень шума вставки, замены и удаления был одинаковым в отдельном эксперименте, принимая значения от 1 до 4% с шагом 1%.

8. Результаты и обсуждение

Результаты экспериментальных исследований по всем четырем функциям приведены на *рисунках 1–4* и в *таблицах 1* и *2*. На *рисунках 1–4* мы показываем числовые последовательности значений функций на двух периодах, содержащие 50 отсчетов на период, при мощности алфавита кодирования 20. На рисунках по оси абсцисс отложены номера отсчетов, а по оси ординат – номера полусегментов кодирования, которым соответствуют символы алфавита мощности 20.

Для всех функций, в качестве примера, показан один вариант внесения шума из исследованных разнообразных вариантов. А именно – вариант с общим уровнем шума в 5%, состоящим из шума удаления в 2% и шума вставки в 3%. На всех рисунках черные точки соответствуют исходной периодической последовательности, полученной из указанной функции, серые точки – это зашумленная последовательность значений функции в отсчетах, а белые с обводкой – это восстановленная исследуемым алгоритмом периодическая последовательность. Пунктир показывает последовательность точек по отсчетам.

Результаты исследования по оценке точности δ определения периода показали, что для всех функций полученные результаты не сильно отличаются друг от друга, т.е. исследуемый метод слабо чувстви-

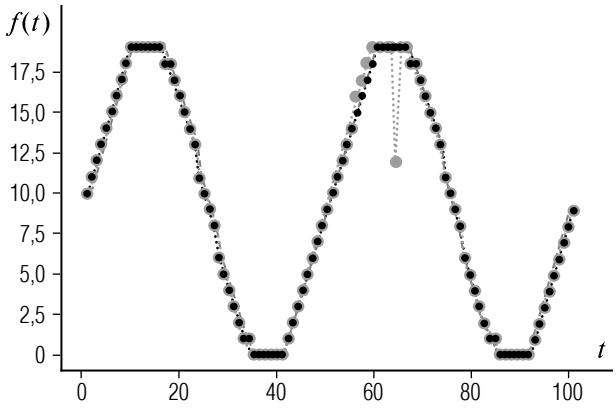

Рис. 1. Два периода функции $\sin(t)$, общий шум 5%, шум удаления 2%, шум вставки 3%, 50 отсчетов на период, мощность алфавита кодирования – 20 символов

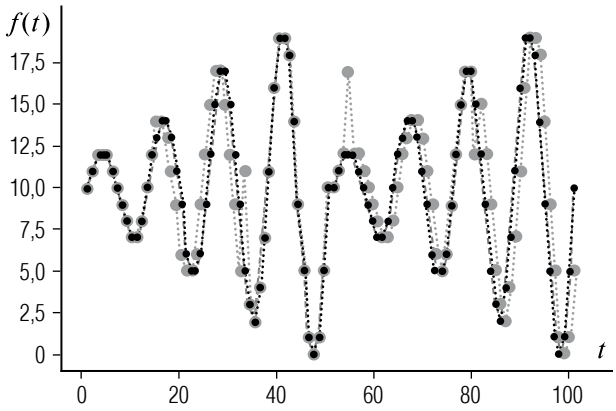

Рис. 2. Два периода функции $\left(t-8\left\lfloor\frac{t}{8}\right\rfloor+1\right)\sin(\pi t)$, общий шум 5%, шум удаления 2%, шум вставки 3%, 50 отсчетов на период, мощность алфавита кодирования – 20 символов

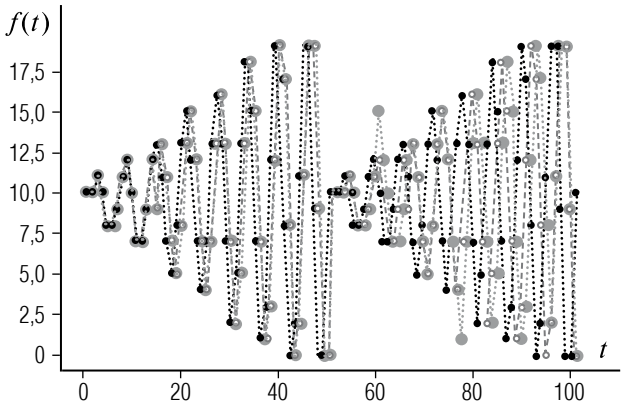

Рис. 3. Два периода функции $\left(t-16\left\lfloor\frac{t}{16}\right\rfloor+1\right)\sin(\pi t)$, общий шум 5%, шум удаления 2%, шум вставки 3%, 50 отсчетов на период, мощность алфавита кодирования – 20 символов

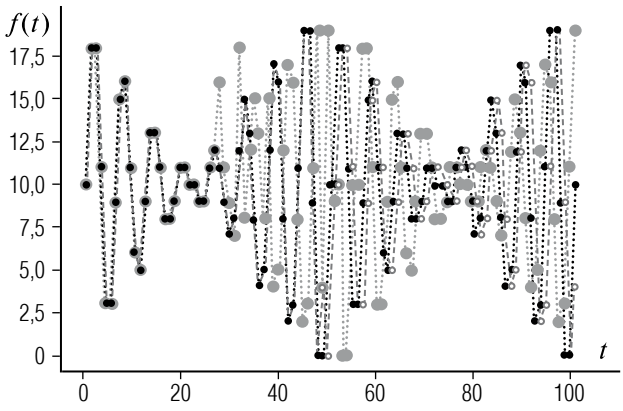

Рис. 4. Два периода функции $\left|7-t+16\left\lfloor\frac{t}{16}\right\rfloor\right|\sin(\pi t)$, общий шум 5%, шум удаления 2%, шум вставки 3%, 50 отсчетов на период, мощность алфавита кодирования – 20 символов

телен к виду периодической функции (по крайней мере для данных четырех функций). С связи с этим мы приводим результаты только для одной функции.

Для лучшей наглядности мы приводим в *таблице 1* не среднее значение δ, усредненное по 100 экспериментам путем внесения случайных шумов с заданным уровнем, а долю последовательностей (из 100 зашумленных) с восстановленным периодом, значение которого не более чем на 2% отличается от исходного периода.

Результаты исследований по качеству восстановления периодически повторяющегося фрагмента улучшенным алгоритмом приведены в *таблице 2.* Значения медианы $\varepsilon\left(\overline{q}^{\sigma}, q^{\sigma}\right)$ приведены для всех четырех исследованных функций при шуме равномерной структуры. При этом в каждую из 100 исходных последовательностей случайно вносился одинаковый уровень шума каждого типа (вставка, удаление, замена) Этот уровень изменялся от 1 до 4% с шагом 1%. Эксперименты были проведены для всех мощностей алфавита – 10, 20, 30, 40 и 50 и всех значений периода. В *таблице 2* приведены результаты для значения периода $p = 50$.

По полученным экспериментальным данным усовершенствованный метод восстановления периода и периодически повторяющегося фрагмента

Таблица 1.

Доля символьных последовательностей с восстановленным периодом, не более чем на 2% отличающимся от исходного, функция –

$$\left(t-8\left\lfloor\frac{t}{8}\right\rfloor+1\right)\sin(\pi t)$$

Уровень шума, %			Мощность алфавита (количество символов)					
удаление	замена	вставка	10	20	30	40	50	60
0	0	5	25	26	33	24	26	31
0	1	4	25	16	22	26	24	26
0	2	3	98	97	94	95	96	93
0	3	2	99	100	98	99	99	100
0	4	1	100	98	100	100	100	100
0	5	0	100	100	100	100	100	100
1	0	4	94	93	94	90	95	96
1	1	3	100	99	96	100	97	97
1	2	2	99	100	99	100	98	97
1	3	1	99	97	97	97	98	99
1	4	0	94	96	97	96	97	95
2	0	3	99	96	99	99	96	97
2	1	2	98	97	99	100	99	98
2	2	1	99	100	99	97	100	97
2	3	0	97	98	97	98	98	98
3	0	2	100	100	98	97	99	100
3	1	1	98	97	98	96	98	98
3	2	0	85	94	91	91	87	90
4	0	1	88	85	88	83	90	88
4	1	0	22	34	28	26	35	31
5	0	0	8	9	7	7	17	12

показывает в целом удовлетворительные результаты. Данные *таблиц 1* и *2* показывают, что метод обладает слабой чувствительностью по мощности кодирующего алфавита и по виду функции как при определении периода, так и при восстановлении периодически повторяющегося фрагмента.

При определении периода (см. *табл. 1*) метод оказывается восприимчивым к разности между уровнями шумов вставки и удаления, поскольку именно эта разность влияет на расстояния между повторяющимися подсловами длины 10 в зашумленной последовательности. Заметим, что наличие только шума замены в 5% приводит к наилучшему наблюдаемому в экспериментах результату. При определении периодически повторяющего фрагмента при шуме равномерной структуры ни вид функции, ни мощность алфавита не оказывают заметного влияния на результаты. Единственный влияющий в этом случае фактор – общий уровень шума.

Заключение

Применение моделей символьных циклов с шумами дает возможность решения задач по вероятностному прогнозированию символьных зашумленных последовательностей и позволяет

Таблица 2.

Влияние уровня шума и мощности алфавита на медиану $\varepsilon\left(\overline{q}^{\sigma}, q^{\sigma}\right)\cdot 100\%$

Суммарный шум %	Мощность алфавита (количество символов)	Значения медианы $\varepsilon\left(\overline{q}^{\sigma}, q^{\sigma}\right)\cdot 100\%$ для различных функций			
		$\sin(t)$	$\left(t-8\left\lfloor\frac{t}{8}\right\rfloor+1\right)\sin(\pi t)$	$\left(t-16\left\lfloor\frac{t}{16}\right\rfloor+1\right)\sin(\pi t)$	$\left\lvert 7-t+16\left\lfloor\frac{t}{16}\right\rfloor\right\rvert\sin(\pi t)$
3%	10	0	0	0	0
	20	0	0	0	0
	30	0	0	0	0
	40	0	0	0	0
	50	0	0	0	0
6%	10	2	2	2	2
	20	2	2	2	2
	30	2	2	2	2
	40	2	2	2	2
	50	2	2	2	2
9%	10	4	4	4	4
	20	6	4	4	4
	30	4	4	4	4
	40	4	4	5	4
	50	4	4	5	4
12%	10	7	6	8	6
	20	8	6	8	6
	30	6	6	6	6
	40	8	8	6	8
	50	8	8	7	6

разрабатывать эффективные методы прогнозирования, реконструкции и аппроксимации данных в форме символьных кодов по фрагментарной, неполной и искаженной информации.

В статье предложен усовершенствованный метод решения задачи восстановления периодической символьной последовательности на основе исходной последовательности, полученной путем внесения шумов вставки, удаления и замены в неизвестную периодическую последовательность. Метод основан на исследовании частотной встречаемости и расстояний между совпадающими подсловами фиксированной длины. В качестве синтетических данных рассматриваются символьные последовательности в алфавитах разной мощности, кодирующие зашумленные периодические функции. Помимо описания усовершенствованного метода нахождения периодически повторяющегося фрагмента статья содержит результаты экспериментального исследования зависимости характеристик качества метода восстановления периода и периодически повторяющегося фрагмента от мощности кодирующего алфавита и уровней шума различных типов. Исследование проводилось для зашумленных символьных кодов периодических функций, являющихся моделями зашумленных (квазипериодических) временных рядов. Такого рода исходные данные часто возникают в задачах анализа и прогнозирования временных рядов в бизнес-информатике и менеджменте.

Вычислительный эксперимент показал, что качество метода зависит не только от общего уровня шума, но и от соотношения уровня шумов. Предложенный метод обладает слабой чувствительностью по мощности кодирующего алфавита и по виду периодической функции, как при определении длины периода, так и при восстановлении периодически повторяющегося фрагмента. Исследование для шума равномерной структуры показало, что единственным фактором, влияющим на качество, является уровень шума, при этом ни вид функции, ни мощность алфавита не оказывают заметного влияния на результаты.

Полученные в статье результаты позволяют дать рекомендации о возможном применении метода при решении задач анализа символьных кодов зашумленных периодических непрерывных функций в алфавитах небольшой мощности, с уровнем шума, не превышающим10–12%. Такого рода задачи возникают при анализе как динамических процессов и временных рядов в бизнес-информатике и менеджменте, так и при анализе бизнес-процессов в условиях неполной и фрагментарной информации. ■

Благодарности

Работа выполнена при поддержке грантов РФФИ 19-07-00150 и 19-07-00151.

Литература

1. Andersen B. Business process improvement toolbox. Milwaukee, Wisconsin: ASQ Quality Press. 2007.
2. Mitsyuk A.A., Lomazova I.A., van der Aalst W. Using event logs for local correction of process models // Automatic Control and Computer Sciences. 2017. Vol. 51. No 7. P. 709–723. DOI: 10.3103/S0146411617070306.
3. Keogh E.J., Pazzani M.J. An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback // Fourth International Conference on Knowledge Discovery and Data Mining (KDD '98). New York, USA, 27–31 August 1998. P. 239–241. [Электронный ресурс]: https://www.aaai.org/Papers/KDD/1998/KDD98-041.pdf (дата обращения 16.11.2021).
4. Bemdt D.J., Clifford J. Using dynamic time warping to find patterns in time series. AAAI Technical Report // Workshop on Knowledge Discovery in Databases (KDD '94). Seattle, Washington, USA, 31 July – 1 August 1994. P. 359–370. [Электронный ресурс]: https://www.aaai.org/Papers/Workshops/1994/WS-94-03/WS94-03-031.pdf (дата обращения 16.11.2021).
5. Wu Y.-L., Agrawal D., el Abbadi A. A comparison of DFT and DWT based similarity search in time-series databases // Ninth International Conference on Information and Knowledge Management (CIKM' 00). McLean, Virginia, USA, 6–11 November 2000. P. 488–495. DOI: 10.1145/354756.354857.
6. Querying and mining of time series data: Experimental comparison of representations and distance measures / H. Ding [et al.] // VLDB Endowment. Auckland, New Zealand, 23–28 August 2008. Vol. 1. No 2. P. 1542–1552. DOI: 10.14778/1454159.1454226.
7. Kurbalija V., Radovanović M., Geler Z., Ivanović M. The influence of global constraints on DTW and LCS similarity measures for time-series databases // Advances in Intelligent and Soft Computing. 2011. Vol. 101. P. 67–74. DOI: 10.1007/978-3-642-23163-6_10.
8. Dreyer W., Dittrich A.K., Schmidt D. Research perspectives for time series management systems. ACM SIGMOD Record. 1994. Vol. 23. No 1. P. 10–15. [Электронный ресурс]: https://dl.acm.org/doi/abs/10.1145/181550.181553 (дата обращения 16 November 2021).
9. Розенберг Г.С., Шитиков В.К., Брусиловский П.М. Экологическое прогнозирование (функциональные предикторы временных рядов). Тольятти: ИЭВБ РАН, 1994.
10. Скляр А.Я. Анализ и устранение шумовой компоненты во временных рядах с переменным шагом // Кибернетика и программирование. 2019. № 1. С. 51–59. DOI: 10.25136/2306-4196.2019.1.27031.
11. Mauleón I. Aggregated world energy demand projections: Statistical assessment // Energies. 2021. Vol. 14. No 15. P. 1–13. [Электронный ресурс]: https://www.mdpi.com/1996-1073/14/15/4657 (дата обращения 16.11.2021). DOI: 10.3390/en14154657.
12. Suganthi L., Samuel A.A. Energy models for demand forecasting – A review // Renewable and Sustainable Energy Reviews. 2012. Vol. 16. No 2. P. 1223–1240. DOI: 10.1016/j.rser.2011.08.014.
13. Boßmann T., Staffell I. The shape of future electricity demand: Exploring load curves in 2050s Germany and Britain // Energy. 2015. Vol. 90. No 2. P. 1317–1333. DOI: 10.1016/j.energy.2015.06.082.
14. Akca A., Canakoğlu E. Adaptive stochastic risk estimation of firm operating profit // Journal of Industrial and Business Economics. 2021. Vol. 48. No 3. P. 463–504. DOI: 10.1007/s40812-021-00184-z.
15. Abbasimehr H., Shabani M. A new methodology for customer behavior analysis using time series clustering. A case study on a bank's customers // Kybernetes. 2021. Vol. 50. No 2. P. 221–242. DOI: 10.1108/K-09-2018-0506.
16. Fildes R., Wei Y., Ismail S. Evaluating the forecasting performance of econometric models of air passenger traffic flows using multiple error measures // International Journal of Forecasting. 2011. Vol. 27. No 3. P. 902–922. DOI: 10.1016/j.ijforecast.2009.06.002.
17. Bi J.-W., Li H., Fan Zh.-P. Tourism demand forecasting with time series imaging: A deep learning model // Annals of Tourism Research. 2021. Vol. 90. Article no 103255. DOI: 10.1016/j.annals.2021.103255.
18. Neural networks for enhanced stress prognostics for encapsulated electronic packages – A comparison / P. Meszmer [et al.] // Microelectronics Reliability. 2021. Vol. 123, art 114181. DOI: 10.1016/j.microrel.2021.114181.

19. Дещеревский А.В., Журавлев В.И., Никольский А.Н., Сидорин А.Я. Проблемы анализа временных рядов с пропусками и методы их решения в программе WINABD // Геофизические процессы и биосфера. 2016. Т. 15. № 3. С. 5–34.

20. Плотников Д.Е., Миклашевич Т.С., Барталев С.А. Восстановление временных рядов данных дистанционных измерений методом полиномиальной аппроксимации в скользящем окне переменного размера // Современные проблемы дистанционного зондирования Земли из космоса. 2014. Т. 11. № 2. С. 103–110.

21. Lin J., Keogh E., Wei L., Lonardi S. Experiencing SAX: A novel symbolic representation of time series // Data Mining and Knowledge Discovery. 2007. Vol. 15. No 2. P. 107–144. DOI: 10.1007/s10618-007-0064-z.

22. Zhukova G., Smetanin Y., Uljanov M. Informative symbolic representations as a way to qualitatively analyses time series // 2019 International Conference on Engineering Technologies and Computer Science: Innovation & Application. Moscow, Russia, 26–27 March 2019. P. 43–47.

23. Lin J., Keogh E., Lonardi S., Chiu B. A symbolic representation of time series, with implications for streaming algorithms // 8th ACM SIGMOD Workshop on Research Issues in Data mining and Knowledge Discovery. San Diego, California, USA, 13 June 2003. P. 2–11. DOI: 10.1145/882082.882086.

24. Нестеренко А.Ю. Алгоритмы поиска длин циклов в последовательностях и их приложения // Фундаментальная и прикладная математика. 2010. Т. 16. № 6. С. 109–122.

25. Жукова Г.Н., Жуков А.В., Сметанин Ю.Г., Ульянов М.В. Метод определения периода зашумленной периодической символьной последовательности, основанный на позициях подслов в последовательности // Современные информационные технологии и ИТ-образование. 2020. Т. 16. № 1. С. 23–32. DOI: 10.25559/SITITO.16.202001.23-32.

26. Ульянов М.В. Подход к идентификации длины цикла в символьных последовательностях с шумом, основанный на энтропии слов // III Международный научно-технический форум «Современные технологии в науке и образовании» (СТНО–2020). Рязань, 4–6 марта 2020 г. Т. 4. С. 120–124.

27. Левенштейн В.И. Двоичные коды с исправлением выпадений, вставок и замещений символов // Доклады АН СССР. 1965. Т. 163. С. 707–710.

28. Жукова Г.Н., Сметанин Ю.Г., Ульянов М.В. Вероятностная модель шумов для периодических символьных последовательностей // Современные информационные технологии и ИТ-образование. 2019. Т. 15. № 2. С. 431–440. DOI: 10.25559/SITITO.15.201902.431-440.

Об авторах

Жукова Галина Николаевна

кандидат физико-математических наук;

доцент департамента программной инженерии, факультет компьютерных наук, Национальный исследовательский университет «Высшая школа экономики», 101000, г. Москва, ул. Мясницкая, д. 20;

E-mail: galinanzhukova@gmail.com

ORCID: 0000-0003-1835-7422

Ульянов Михаил Васильевич

доктор технических наук, профессор;

ведущий научный сотрудник, Институт проблем управления им. В.А. Трапезникова Российской академии наук, 117997, г. Москва, ул. Профсоюзная, д. 65;

профессор кафедры алгоритмических языков, факультет вычислительной математики и кибернетики, Московский государственный университет им. М.В. Ломоносова, 119991, г. Москва, Ленинские горы, д. 1;

E-mail: muljanov@mail.ru

ORCID: 0000-0002-5784-9836

To the question of restoring symbol sequences encoding noisy periodic functions

Galina N. Zhukova [a]
E-mail: galinanzhukova@gmail.com

Mikhail V. Ulyanov [b,c]
E-mail: muljanov@mail.ru

[a] National Research University Higher School of Economics
Address: 20, Myasnitskaya Street, Moscow 101000, Russia

[b] Trapeznikov Institute of Control Sciences, Russian Academy of Sciences
Address: 65, Profsoyuznaya Street, Moscow 117997, Russia

[c] Lomonosov Moscow State University
Address: 1, Leninskie Gory, Moscow 119991, Russia

Abstract

In business informatics, one of the research subjects is the analysis of data on processes in applied subject areas; here problems of qualitative analysis arise. Such problems arise, for example, in the qualitative study of log files of business processes, in the analysis and prediction of time series and other processes of a different nature. Quite often, to represent information about the processes under study, the methods of qualitative analysis use symbolic coding, which makes it possible to remove unnecessary detailing of numerical descriptions. The relevance of this study is due to the fact that when working with the raw data, researchers often face the presence of noise and distortions of the data, which significantly complicates the solution of the problems of qualitative analysis. When working with symbolic representations of the processes under study, which quite often have a periodic nature, we observe noise of deletion, insertion and replacement of symbols, which complicate the solution of the problem of revealing and analyzing the periodicity. This article deals with the problem of recovering periodic symbolic sequences obtained by coding from samples of continuous periodic functions and distorted by noise of insertion, replacement and deletion of symbols. Trigonometric functions are considered as a specific example of synthetic time series data. To encode trigonometric functions, alphabets of various cardinalities are used. The article presents an experimental study of the dependence of the quality characteristics of the method of period and a periodically repeating fragment recovery, previously proposed by the authors and improved in this study. For alphabets of different cardinalities at fixed sampling intervals, the fraction of sequences with a satisfactorily reconstructed period and the relative error in determining the period are given. The quality of reconstruction of a periodically repeating fragment is estimated by the edit distance from the reconstructed periodic sequence to the original sequence distorted by noise.

Key words: symbolic sequence; cardinality of an alphabet; periodic sequence; sequence with noise; noise of insertion; noise of deletion; noise of change.

Citation: Zhukova G.N., Ulyanov M.V. (2021) To the question of restoring symbol sequences encoding noisy periodic functions. *Business Informatics*, vol. 15, no 4, pp. 22–35. DOI: 10.17323/2587-814X.2021.4.22.35

References

1. Andersen B. (2007) *Business process improvement toolbox.* Milwaukee, Wisconsin: ASQ Quality Press.
2. Mitsyuk A.A., Lomazova I.A., van der Aalst W. (2017) Using event logs for local correction of process models. *Automatic Control and Computer Sciences*, vol. 51, no 7, pp. 709–723. DOI: 10.3103/S0146411617070306.
3. Keogh E.J., Pazzani M.J. (1998) An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback. Proceedings of the *Fourth International Conference on Knowledge Discovery and Data Mining (KDD '98), New York, USA, 27–31 August 1998,* pp. 239–241. Available at: https://www.aaai.org/Papers/KDD/1998/KDD98-041.pdf (accessed 16 November 2021).
4. Bemdt D.J., Clifford J. (1994) Using dynamic time warping to find patterns in time series. *AAAI Technical Report. Workshop on Knowledge Discovery in Databases (KDD '94), Seattle, Washington, USA, 31 July – 1 August 1994,* pp. 359–370. Available at: https://www.aaai.org/Papers/Workshops/1994/WS-94-03/WS94-03-031.pdf (accessed 16 November 2021).
5. Wu Y.-L., Agrawal D., el Abbadi A. (2000) A comparison of DFT and DWT based similarity search in time-series databases. Proceedings of the *Ninth International Conference on Information and Knowledge Management (CIKM' 00), McLean, Virginia, USA, 6–11 November*, pp. 488–495. DOI: 10.1145/354756.354857.
6. Ding H., Trajcevski G., Scheuermann P., Wang X., Keogh E. (2008) Querying and mining of time series data: Experimental comparison of representations and distance measures. Proceedings of the *VLDB Endowment, Auckland, New Zealand, 23–28 August 2008*, vol. 1, no 2, pp. 1542–1552. DOI: 10.14778/1454159.1454226.
7. Kurbalija V., Radovanović M., Geler Z., Ivanović M. (2011) The influence of global constraints on DTW and LCS similarity measures for time-series databases. *Advances in Intelligent and Soft Computing*, vol. 101, pp. 67–74. DOI: 10.1007/978-3-642-23163-6_10.
8. Dreyer W., Dittrich A.K., Schmidt D. (1994) Research perspectives for time series management systems. *ACM SIGMOD Record*, vol. 23, no 1, pp. 10–15. Available at: https://dl.acm.org/doi/abs/10.1145/181550.181553 (accessed 16 November 2021).
9. Rozenberg G.S., Shitikov V.K., Brusilovskij P.M. (1994) *Environmental forecasting (time series functional predictors)*. Tolyatti: IEVB RAS (in Russian).
10. Sklyar A.Ya. (2019) Analysis and elimination of noise component in time series with variable step. *Cybernetics and Programming*, no 1, pp. 51–59 (in Russian). DOI: 10.25136/2306-4196.2019.1.27031.

11. Mauleón I. (2021) Aggregated world energy demand projections: Statistical assessment. *Energies*, vol. 14, no 15, pp. 1–13. Available at: https://www.mdpi.com/1996-1073/14/15/4657 (accessed 16 November 2021). DOI: 10.3390/en14154657.
12. Suganthi L., Samuel A.A. (2012) Energy models for demand forecasting – A review. *Renewable and Sustainable Energy Reviews*, vol. 16, no 2, pp. 1223–1240. DOI: 10.1016/j.rser.2011.08.014.
13. Boßmann T., Staffell I. (2015) The shape of future electricity demand: Exploring load curves in 2050s Germany and Britain. *Energy*, vol. 90, no 2, pp. 1317–1333. DOI: 10.1016/j.energy.2015.06.082.
14. Akca A., Canakoğlu E. (2021) Adaptive stochastic risk estimation of firm operating profit. *Journal of Industrial and Business Economics*, vol. 48, no 3, pp. 463–504. DOI: 10.1007/s40812-021-00184-z.
15. Abbasimehr H., Shabani M. (2021) A new methodology for customer behavior analysis using time series clustering. A case study on a bank's customers. *Kybernetes,* vol. 50, no 2, pp. 221–242. DOI: 10.1108/K-09-2018-0506.
16. Fildes R., Wei Y., Ismail S. (2011) Evaluating the forecasting performance of econometric models of air passenger traffic flows using multiple error measures. *International Journal of Forecasting*, vol. 27, no 3, pp. 902–922. DOI: 10.1016/j.ijforecast.2009.06.002.
17. Bi J.-W., Li H., Fan Zh.-P. (2021) Tourism demand forecasting with time series imaging: A deep learning model. *Annals of Tourism Research,* vol. 90, article no 103255. DOI: 10.1016/j.annals.2021.103255.
18. Meszmer P., Majd M., Prisacaru A., Gromala P.J., Wunderle B. (2021) Neural networks for enhanced stress prognostics for encapsulated electronic packages – A comparison. *Microelectronics Reliability,* vol. 123, article no 114181. DOI: 10.1016/j.microrel.2021.114181.
19. Deshcherevsky A.V., Zhuravlev V.I., Nikolskij A.N., Sidorin A.Ya. (2016) Time series analysis problems with gaps and methods for solving them in the program WINABD. *Geophysical processes and the biosphere*, vol. 15, no 3, pp. 5–34 (in Russian).
20. Plotnikov D.E., Miklashevich T.S., Bartalyov S.A. (2014) Reconstruction of time series of remote sensing data by the method of polynomial approximation in a sliding window of variable size. *Current Problems in Remote Sensing of the Earth from Space*, vol. 11, no 2, pp. 103–110 (in Russian).
21. Lin J., Keogh E., Wei L., Lonardi S. (2007) Experiencing SAX: A novel symbolic representation of time series. *Data Mining and Knowledge Discovery*, vol. 15, no 2, pp. 107–144. DOI: 10.1007/s10618-007-0064-z.
22. Zhukova G., Smetanin Y., Uljanov M. (2019) Informative symbolic representations as a way to qualitatively analyses time series. Proceedings of the *2019 International Conference on Engineering Technologies and Computer Science: Innovation & Application, Moscow, Russia, 26–27 March 2019*, pp. 43–47.
23. Lin J., Keogh E., Lonardi S., Chiu B. (2003) A symbolic representation of time series, with implications for streaming algorithms. Proceedings of the *8th ACM SIGMOD Workshop on Research Issues in Data mining and Knowledge Discovery, San Diego, California, USA, 13 June 2003*, pp. 2–11. DOI: 10.1145/882082.882086.
24. Nesterenko A.Yu. (2010) Algorithms for finding the lengths of cycles in sequences and their applications. *Fundamental and Applied Mathematics*, vol. 16, no 6, pp. 109–122 (in Russian).
25. Zhukova G.N., Zhukov A.V., Smetanin Yu.G., Ulyanov M.V. (2020) The method of estimating the period of a symbolic periodic sequence with noise, based on the sub-words positions in the sequence. *Modern information technologies and IT education*, vol. 16, no 1, pp. 23–32 (in Russian). DOI: 10.25559/SITITO.16.202001.23-32.
26. Ulyanov M.V. (2020) An approach to identifying the cycle length in noisy character sequences based on the entropy of words. Proceedings of the *III International Scientific and Technical Forum "Modern Technologies in Science and Education", Ryazan, 4–6 March 2020*, vol. 4, pp. 120–124 (in Russian).
27. Levenshtejn V.I. (1965) Binary codes with corrected dropouts, insertions and character replacements. *Proceedings of the Academy of Sciences*, vol. 163, pp. 707–710 (in Russian).
28. Zhukova G.N., Zhukov A.V., Smetanin Yu.G., Ulyanov M.V. (2019) Stochastic model of noises for periodic symbol sequences. *Modern information technologies and IT education,* vol. 15, no 2, pp. 431–440. DOI: 10.25559/SITITO.15.201902.431-440 (in Russian).

About the authors

Galina N. Zhukova

Cand. Sci. (Phys.-Math.);

Associate Professor, School of Software Engineering, Faculty of Computer Science, National Research University Higher School of Economics, 20, Myasnitskaya Street, Moscow 101000, Russia;

E-mail: galinanzhukova@gmail.com

ORCID: 0000-0003-1835-7422

Mikhail V. Ulyanov

Dr. Sci. (Tech.);

Leading Researcher, Laboratory of Scheduling Theory and Discrete Optimization, V.A. Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, 65, Profsoyuznaya Street, Moscow 117997, Russia;

Professor, Department of Algorithmic Languages, Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, 1, Leninskie Gory, Moscow 119991, Russia;

E-mail: muljanov@mail.ru

ORCID: 0000-0002-5784-9836