ISSN 2078-9181 DOI 10.15622/sp.18.2

РОССИЙСКАЯ АКАДЕМИЯ НАУК
Отделение нанотехнологий и информационных технологий

САНКТ-ПЕТЕРБУРГСКИЙ
ИНСТИТУТ ИНФОРМАТИКИ И АВТОМАТИЗАЦИИ РАН

ТРУДЫ СПИИРАН

proceedings.spiiras.nw.ru

ТОМ 18 **№** 2

Санкт-Петербург
2019

18+

SPIIRAS PROCEEDINGS

Volume 18 № 2, 2019

Scientific, educational, and interdisciplinary journal primarily specialized
in computer science, automation, and applied mathematics

Trudy SPIIRAN ♦ Founded in 2002 ♦ Труды СПИИРАН

Founder and Publisher

St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences

Editorial Board's address

14-th line VO, 39, SPIIRAS, St. Petersburg, 199178, Russia,
e-mail: publ@iias.spb.su, web: http://www.proceedings.spiiras.nw.ru/

The journal is indexed in Scopus

ТРУДЫ СПИИРАН

Том 18 № 2, 2019

Научный, научно-образовательный, междисциплинарный журнал с базовой специализацией в области информатики, автоматизации и прикладной математики
Журнал основан в 2002 году

Учредитель и издатель

Федеральное государственное бюджетное учреждение науки
Санкт-Петербургский институт информатики и автоматизации Российской академии наук (СПИИРАН)

Адрес редакции

199178, Санкт-Петербург, 14-я линия, д. 39,
e-mail: publ@iias.spb.su, сайт: http://www.proceedings.spiiras.nw.ru/
Журнал индексируется в международной базе данных Scopus
Журнал входит в «Перечень ведущих рецензируемых научных журналов и изданий, в которых должны быть опубликованы основные научные результаты диссертации на соискание ученой степени доктора и кандидата наук»

CONTENTS

СОДЕРЖАНИЕ

УДК 681.5

DOI 10.15622/sp.18.2.267-301

И.В. Бычков, М.Ю. Кензин, Н.Н. Максимкин

ДВУХУРОВНЕВЫЙ ЭВОЛЮЦИОННЫЙ ПОДХОД К МАРШРУТИЗАЦИИ ГРУППЫ ПОДВОДНЫХ РОБОТОВ В УСЛОВИЯХ ПЕРИОДИЧЕСКОЙ РОТАЦИИ СОСТАВА

Бычков И.В., Кензин М.Ю., Максимкин Н.Н. **Двухуровневый эволюционный подход к маршрутизации группы подводных роботов в условиях периодической ротации состава.**

Аннотация. Применение скоординированных групп автономных подводных роботов представляется наиболее перспективной и многообещающей технологией, обеспечивающей решение самого широкого спектра океанографических задач. Групповое выполнение комплексных широкомасштабных миссий, как правило, связано с длительным пребыванием роботов в заданной акватории, что в условиях ограниченной энергоемкости аккумуляторных батарей возможно только при наличии специализированных док-станций для ее пополнения. Для обеспечения высокого уровня работоспособности действующей группировки возникают две параллельные задачи: эффективно распределить задания миссии между членами группы и определить порядок подзарядки роботов на длительном промежутке времени. При этом необходимо учитывать, что реальные робототехнические системы функционируют в динамической подводной среде, а значит, могут подвергаться влиянию непредвиденных событий и различного рода неполадок.

В данной статье предлагается двухуровневый подход к динамическому планированию групповой стратегии, основанный на декомпозиции миссии на последовательность рабочих периодов с обязательным сбором действующей группировки по окончанию каждого из них. Задача планировщика на верхнем уровне заключается в составлении такого расписания циклов зарядки для всех аппаратов в группе, которое обеспечивало бы своевременное пополнение батарей при недопущении одновременной зарядки большого количества роботов. На основе выбранного расписания осуществляется декомпозиция миссии таким образом, чтобы каждый сбор группы сопровождался либо выходом робота из группы для осуществления подзарядки, либо возвращением в группу уже заряженного аппарата. Такая схема позволяет отслеживать статус группы и осуществлять оперативное перепланирование при изменении ее состава. Маршрутизация группы на каждом рабочем периоде осуществляется низкоуровневым планировщиком, работающим на графе целей и учитывающим технические возможности всех аппаратов в группе, а также все действующие ограничения и требования к выполнению конкретных задач. В статье предлагается эволюционный подход к децентрализованной реализации обоих планировщиков с применением специализированных эвристик, процедур улучшения решений и оригинальных схем кодирования и оценки решений; приводятся результаты вычислительных экспериментов.

Ключевые слова: автономные подводные роботы, групповое управление, задача составления расписания, задача маршрутизации транспорта, эволюционные алгоритмы.

1. Введение.

На протяжении последних десятилетий автономные необитаемые подводные аппараты (АНПА) демонстрируют свою высокую эффективность при выполнении различного рода подводных работ по обследованию морского дна,

включая задачи картографирования, мониторинга, взятия проб, обнаружения и деактивации мин и другие. Эти мобильные роботы способны принимать участие в многоцелевых подводных миссиях большой длительности при значительно меньших ресурсных затратах в сравнении с применением обитаемых подводных транспортных средств [1]. При этом на данный момент подавляющее большинство реальных миссий АНПА осуществляется исключительно единичными аппаратами и на базе заранее спланированных траекторий и последовательностей действий [2]. Очевидно, что в условиях ограниченного времени и конечного энергозапаса батарей единичный аппарат не в состоянии обеспечить выполнение значительного количества работ в рамках комплексной подводной миссии. В связи с этим наиболее перспективным средством выполнения широкомасштабных океанографических операций представляется применение скоординированных робототехнических группировок, способных обеспечивать эффективный мониторинг значительных районов океана, а также производить измерения в заданной акватории с высоким разрешением как по времени, так и в пространстве [3].

Организация совместной работы группы подводных роботов в динамической подводной среде является сложным и нетривиальным процессом: заданный набор задач должен быть выполнен в результате коллективных действий нескольких аппаратов при действующих ограничениях и в условиях изменяющейся внешней обстановки. Разработка и развитие таких автономных многокомпонентных систем напрямую зависит от надежности и эффективности реализации механизмов планирования и маршрутизации [4]. Таким образом, на первый план выходит именно эффективная координация внутри сети АНПА, в том числе блок системы группового управления, который отвечает за распределение заданий между всеми роботами и планирование индивидуальных маршрутов.

В общем случае задача распределения целей и планирования маршрутов для группы АНПА представляет собой вариацию задачи маршрутизации транспорта (*vehicle routing problem*) со специализированными пространственно-временными ограничениями, накладываемыми особенностями подводных операций, неопределенной природой динамической среды, а также неточностью измерительных приборов. Многие виды подводных задач, такие как патрулирование, охрана, проведение замеров и другие, требуют для своего выполнения не разового проведения соответствующих работ в заданных областях акватории, а регулярной серии таких посещений роботами группы [5]. В этом отношении многозадачные подводные

операции могут быть поделены на два класса: обследовательские миссии и миссии по мониторингу.

Так, обследовательские миссии объединяют все виды подводных задач, для выполнения которых требуется разовое прибытие АНПА. В свою очередь, миссии по мониторингу включают все работы, требующие регулярных инспекций в соответствующих областях аппаратами группы с заданной периодичностью. В обоих случаях выполнение заданий может подразумевать наличие и других дополнительных условий, требований и ограничений (технических, временных и т.д.). Поскольку класс миссий по мониторингу совмещает в себе одновременно пространственные и комплексные временные ограничения, то именно он представляет больший научно-исследовательский и практический интерес.

Вместе со всеми установленными ограничениями необходимо учитывать, что реальные многокомпонентные робототехнические комплексы являются частично или полностью автономными системами и вынуждены самостоятельно реагировать на различные изменения во внешней среде, непрогнозируемые события и отказы оборудования. Помимо этого, необходимость периодической подзарядки аккумуляторных батарей АНПА в процессе выполнения продолжительных миссий также значительно затрудняет групповую маршрутизацию: во-первых, уход отдельных аппаратов на зарядку приводит к изменению действующего состава группировки, а значит, и к необходимости корректировки текущей групповой стратегии; во-вторых, в связи с этим возникает потребность в регулярных сборах всех роботов (рандеву) в установленном месте с целью актуализации состояния группы, обмена данными, отправки отдельных АНПА на зарядку либо приема роботов в группу после нее [6]. Хотя частые рандеву и способствуют повышению осведомленности группы и скорости ее реакции на любые существенные изменения, но вместе с тем каждый такой сбор может на значительное время отвлекать членов группы от своевременного выполнения текущих задач.

В данной статье предлагается использование двухуровневого подхода к решению задачи динамической маршрутизации группы АНПА при выполнении многозадачной миссии большой продолжительности. На верхнем уровне классический генетический алгоритм осуществляет составление расписания групповых рандеву в соответствии с ожидаемой ротацией роботов для пополнения заряда аккумуляторных батарей. Учитывая запланированные изменения состава при сборах группы, нижнеуровневый гибридный эволюционный алгоритм генерирует маршруты для каждого рабочего периода, которые обеспечивают максимальную эффективность

действующего состава при выполнении поставленных задач и гарантируют прибытие всех АНПА к месту следующего сбора. Данная работа является продолжением предыдущих исследований авторов в области группового управления [7, 8] и посвящена координации группы в условиях топливных ограничений и связанной с ними периодической ротацией состава группы.

2. Постановка задачи. В общем случае многозадачные миссии АНПА по мониторингу акватории состоят в посещении и обследовании (проведении ряда работ) группой подводных роботов заданного множества целей с рекомендуемой частотой. Задача маршрутизации группы при выполнении длительного мониторинга заключается в построении такого допустимого группового маршрута, который обеспечивал бы, насколько это возможно, своевременное обследование всех целей в условиях периодической смены состава действующей группировки для осуществления подзарядки аккумуляторных батарей. В дальнейшем мы будем называть задачу, исследуемую в рамках данной статьи, *задачей регулярного мониторинга.* Приведем ее формальную постановку.

Обозначим через *T* длительность всей миссии. Предполагается, что это большое (в масштабах задачи планирования) число. Пусть изначально миссия включает $N = \{1,...,n\}$ целей (заданий), расположенных в рамках обозначенной акватории. Кроме своего местоположения, каждая цель $i \in N$ характеризуется также требуемой периодичностью обследований p_i и длительностью разового обследования s_i. Значение периодичности p_i означает, что длительность временного интервала между каждыми двумя последовательными посещениями *i*-ой цели аппаратами группы должна составлять ровно p_i. Таким образом, в случае прибытия к цели ранее, чем через установленный интервал p_i, аппарат вынужден будет пребывать в режиме ожидания вплоть до истечения требуемого интервала. В случае прибытия АНПА с опозданием, новый интервал p_i будет отсчитываться не от ожидаемого, а от фактического времени последнего обследования цели.

Ограничения на периодичность посещений такого вида являются аналогом «жестких» временных окон из соответствующего класса задач маршрутизации и отвечают реальным задачам исследования динамики различных процессов [9], когда результаты обследования каждой цели (пробы, замеры, фото- и видеоизображения) должны быть получены аппаратами предпочтительно через равные промежутки времени. Постановки с подобными ограничениями практически не представлены в литературе, в отличие от ограничений «мягкого» вида,

когда допускается (а зачастую и поощряется) обследование целей чаще максимально допустимой периодичности (миссии по экологическому мониторингу, охране, патрулированию и т.п.). В то же время способность работать с «жесткими» временными окнами в задачах регулярного мониторинга позволит в дальнейшем рассматривать более сложные постановки, в которых будут одновременно представлены цели обоих типов.

Пусть группа подводных роботов, выполняющая миссию, изначально состоит из m аппаратов. Все аппараты в группе являются функционально идентичными, то есть каждый из них обладает всем необходимым оборудованием для обследования любой цели i за время s_i. При этом аппараты могут различаться между собой по своей крейсерской скорости v^k движения в водной среде и по емкости своих аккумуляторных батарей $b^k, k = 1,...,m$. Здесь мы допускаем, что все АНПА в течение миссии перемещаются между целями со своей равномерной крейсерской скоростью, а емкость аккумуляторов определяется не в единицах энергии, а в средней длительности функционирования АНПА.

Ограниченная емкость аккумуляторных батарей вынуждает роботов группы периодически прерывать работу для осуществления подзарядки путем стыковки со специальным зарядным доком или, в случае использования солнечных батарей, простого всплытия на поверхность. Конкретное расположение зарядных баз не является существенным, поскольку нам требуется лишь оценка сверху времени, которое требуется каждому АНПА для перемещения до ближайшего дока (всплытия) из рабочей области акватории. Также в данной работе мы допускаем, что количество зарядных баз и доков согласовано с количеством аппаратов в группе, а значит, в процессе выполнения миссии любой АНПА группы сможет быть обслужен на зарядной базе в любой момент времени. Обозначим среднюю скорость зарядки любой батареи через постоянную величину $c > 1$. Тогда для полной зарядки аккумуляторной батареи емкостью b потребуется b / c времени.

Поскольку все заданные цели расположены в ограниченной водной акватории и привязаны к конкретным ее участкам, они могут быть представлены в виде сети, где каждому узлу сети соответствует одна из целей миссии, что позволяет напрямую перейти к решению задачи маршрутизации транспорта на местности, представленной в виде графа [4]. Определим множество всех узлов сети как $V = N \cup V_0$, где V_0 соответствует установленной заранее точке группового сбора. Обозначим множество всех ребер графа как $\varepsilon = \{(i, j) : i, j \in V, i \neq j\}$.

Каждое ребро здесь соответствует кратчайшему пути между соответствующей парой целей. Предполагается, что траектории всех путей для исходного множества целей вычисляются заранее, а пути для новых заданий, возникающих в процессе выполнения миссии, рассчитываются онлайн на борту АНПА группы с использованием специализированного планировщика. Каждому ребру ставится в соответствие значение его веса — длина соответствующей траектории, которая при необходимости может быть легко пересчитана во временные затраты на перемещение АНПА по этому пути. Таким образом, ненаправленный (в отсутствие течений) полносвязный граф $G = (V, \varepsilon)$ представляет собой «дорожную карту» текущей многозадачной миссии.

На рисунке 1 схематически представлен процесс выполнения миссии по мониторингу группой из четырех АНПА в предложенной выше постановке. В изображенный момент времени действующая группировка состоит из двух роботов, выполняющих обследование текущих целей; еще один аппарат временно вышел из группы и движется к док-станции для восполнения заряда своих аккумуляторных батарей; четвертый аппарат недавно закончил подзарядку и находится на пути к точке группового сбора. После прибытия четвертого АНПА в точку сбора, действующая группировка уже в обновленном составе продолжит обход целей в соответствии с их состоянием (временем с последнего посещения).

Рис. 1. Схематическое представление выполнения миссии по длительному мониторингу акватории группой АНПА с непостоянным составом

Описанная задача содержит в себе особенности сразу нескольких известных NP-трудных транспортных задач: систематического покрытия и патрулирования (*persistent coverage and tasks patrolling problem*), мультикоммивояжера и нескольких вариаций задачи маршрутизации транспорта (*vehicle routing problem*). Постановка задачи маршрутизации транспорта (ЗМТ) и методы ее решения впервые были предложены в [10]. В рамках ЗМТ целью транспортных средств является вывоз и доставка грузов со склада до заданного набора клиентов, а сама задача маршрутизации заключается в поиске маршрута, оптимального по заданной характеристике (время, расстояние, стоимость и др.). Наиболее значимыми вариациями задачи маршрутизации являются периодическая ЗМТ (*periodic vehicle routing problem*), добавляющая в постановку многодневный горизонт планирования и требования клиентов по количеству посещений, и ЗМТ с временными окнами (*vehicle routing problem with time windows*), когда каждый клиент должен быть посещен строго в определенный период времени. В отсутствие необходимости непосредственной доставки грузов такие задачи сводятся к задаче мультикоммивояжера (*multiple traveling salesman problem*), когда необходимо найти кратчайшую групповую траекторию, обеспечивающую обход всех клиентов при действующих пространственно-временных ограничениях с последующим возвращением в стартовую точку.

Задача патрулирования (*coverage and patrol*) заключается в непрерывном посещении одним или несколькими агентами набора целей в рамках заданной области с целью минимизации периода между последовательными посещениями каждой точки [11]. В идеальном случае такой обход всего множества целей должен обеспечивать полное покрытие области патрулирования. В [12] исследуется расширенная постановка задачи патрулирования с предотвращением вторжения: задается граф потенциальных целей, где в соответствие каждой вершине ставится время, требующееся злоумышленнику для вторжения. Таким образом, маршрут для патрулирующих агентов строится с целью обеспечить посещение каждой цели с частотой, гарантирующей обнаружение злоумышленника.

Задача систематического мониторинга (*persistent visitation problem*), представленная в [13], лежит на пересечении патрульных постановок и задачи маршрутизации. Так, в [13] отсутствует понятие горизонта планирования в его классическом понимании. Вместо этого каждая цель характеризуется минимально допустимой частотой посещения. Также на транспортные средства накладывается ограничение по емкости топливного бака, при этом на карте имеется ряд заправочных станций с установленной ценой топлива. Ставится задача обеспечить непрерывный обход всех целей с требуемой

частотой и своевременную дозаправку транспортных средств при минимальных затратах на топливо.

Задача систематического мониторинга частично пересекается с задачей регулярной маршрутизации, исследуемой в данной статье. В то же время между постановками существует ряд значительных различий, которые не позволяют отнести исследуемую задачу к вариации задачи систематического мониторинга и использовать для ее решения аналогичные методы и идеи. Так, требование ко времени посещения целей в [13] является аналогом «мягких» временных окон из соответствующей ЗМТ, то есть допускает преждевременное обследование цели чаще значения ее периодичности. В рассматриваемой нами задаче транспортное средство вынуждено ожидать строгого истечения такого отрезка времени, что соответствует принципу «жестких» временных окон. При этом именно «мягкие» временные требования позволяют авторам разбить обход всего множества целей на несколько последовательных циклических траекторий, непрерывное движение по которым обеспечивало бы постоянное приведение задачи в состояние, идентичное начальному. Кроме того, авторы [13] пренебрегают временными затратами на посещение целей и дозаправку, считая их ничтожно малыми относительно временных затрат на движение, и предлагают решение только для постановки с единичным транспортным средством. Даже при таких существенных допущениях доказывается, что задача систематического мониторинга, будучи новой оригинальной вариацией ЗМТ, является NP-полной.

В литературе представлено большое количество работ, развивающих модель систематического мониторинга, однако подавляющее их большинство посвящено поиску именно циклических траекторий ограниченной длины. Ограничения же в виде «жестких» временных окон делают такое разбиение невозможным, что приводит к необходимости оперирования ациклическими маршрутами с плавающим горизонтом планирования [14]. Кроме того, стоит отметить, что в подавляющем большинстве прикладных исследований по перечисленным задачам рассматривается их применимость к управлению беспилотными аппаратами, действующими в свободной воздушной среде [15-17], поэтому вопрос обеспечения межаппаратной коммуникации в них не поднимается. Авторам данной статьи не известны модели группового мониторинга, в которых ограничения в виде «жестких» временных окон сочетались бы с необходимостью периодической подзарядки, а аппараты группы и цели различались бы между собой по своим характеристикам. Такие задачи маршрутизации, объединяющие в себе целый спектр различных ограничений и требований, в современной литературе принято

относить к широкому классу комплексных задач маршрутизации *(rich vehicle routing problem*) [18].

Прежде чем перейти к решению рассматриваемой задачи как комплексной ЗМТ, необходимо отметить, что процесс выполнения подводной миссии является полностью автономным, то есть все вычисления производятся исключительно на бортовых системах АНПА. Очевидно, что для достижения максимальной суммарной вычислительной мощности при реализации децентрализованной архитектуры системы группового управления, все распараллеливаемые вычисления должны быть распределены между аппаратами. В то же время эффективная децентрализация возможна только в том случае, когда все узлы системы обладают актуальными данными. Так как коммуникационный обмен внутри группы осуществляется путем передачи данных между аппаратами по гидроакустическому каналу связи, полная синхронизация актуальных данных внутри группы может быть достигнута, только когда каждый робот будет в состоянии установить канал связи (напрямую или опосредованно) с любым другим роботом в группе. Ввиду низкой скорости и ограниченной дальности действия гидроакустического канала связи, предполагается, что полная синхронизация данных может быть осуществлена только во время проведения так называемых рандеву, когда вся действующая группировка АНПА одновременно прибывает в некоторую заранее оговоренную область. В дальнейшем мы будем называть групповые маршруты *коммуникационно устойчивыми*, если они гарантируют возможность регулярной синхронизации данных [7].

Требование *коммуникационной устойчивости* диктуется, в первую очередь, динамической природой реальных подводных миссий: во-первых, по результатам проводимых АНПА обследований может потребоваться изменение самого множества целей либо параметров отдельных целей; во-вторых, неопределенность внешней среды, ограниченные коммуникационные возможности, вероятность возникновения задержек и неисправностей оборудования могут приводить к непредвиденным изменениям в статусе действующей группировки. Все эти изменения могут происходить в реальном времени и приводить к потере эффективности выбранной ранее групповой стратегии. Таким образом, любые значимые изменения должны инициировать корректировку текущего группового маршрута с целью максимизации эффективности работы группы в новых условиях. В число возможных событий входят:

– добавление в миссию новых целей или отказ от текущих целей;

– изменение параметров цели или свойств аппаратов;

– потеря группой аппарата или выход АНПА из строя;

– расширение группы новыми роботами или обнаружение ранее потерянного АНПА;
– уход АНПА из группы для подзарядки батарей;
– присоединение АНПА после подзарядки к группе.

Отдельно стоит отметить, что изначально два последних события не должны быть указаны в приведенном списке, так как они могут быть спрогнозированы на основе рабочих циклов АНПА с высокой степенью точности. Однако, поскольку предполагаемое время наступления этих событий также может смещаться вследствие возникновения других изменений, мы предлагаем рассматривать их в качестве ключевых точек для процедуры планирования миссии наравне с другими случаями.

Эффективность групповой работы в целом определяется регулярностью обследований всех целей миссии при своевременной подзарядке АНПА. Ситуации, когда роботы прибывают с опозданием, задерживая обследование цели, являются нежелательными и должны по возможности быть исключены. Таким образом, ставится задача разработки алгоритмического обеспечения для системы управления, реализующей эффективное планирование групповых маршрутов. Разрабатываемый подход должен обеспечивать генерацию группового маршрута с минимальным опозданием при выполнении продолжительного регулярного обследования целей миссии вплоть до момента ее окончания в условиях высокой динамики внешней среды и постоянной ротации состава действующей группы.

3. Двухуровневая система управления. Эффективное планирование реальных подводных миссий для групп АНПА является комплексной и нетривиальной задачей, особенно в жестких временных рамках, когда необходима оперативная реакция группы на изменения внешней среды и непредвиденные возмущения. Групповая маршрутизация большой размерности сама по себе является задачей высокой вычислительной сложности даже в идеальной (полностью известной и неизменной) среде, поэтому в условиях динамической внешней среды поиск оптимального решения на периоде выполнения миссии T теряет свою целесообразность. В связи с этим на первый план выходит быстрое и точное локальное планирование с учетом ближайших ожидаемых изменений.

Мы предлагаем использование двухуровневой системы управления для обеспечения групповых миссий по регулярному мониторингу большой продолжительности. На двух уровнях предлагаемого подхода реализуется динамическое планирование групповой стратегии и локальное распределение целей (маршрутизация)

соответственно. Так, цель динамического планировщика миссии на верхнем уровне — производить декомпозицию миссии, обеспечивающую регулярность групповых сборов и общее снижение вычислительной нагрузки для расчета маршрутов аппаратов. В свою очередь, задача планировщика нижнего уровня — распределить цели между аппаратами группы и выбрать эффективный порядок их обхода с учетом действующих требований и ограничений.

В идеальном случае, когда точки разбиения при декомпозиции миссии соответствуют моментам возникновения любых значимых изменений, задача маршрутизации на каждом рабочем периоде группы (временном отрезке между двумя последовательными точками разбиения) будет являться статичной. Однако, поскольку среди приведенных выше событий достоверно предсказаны могут быть только те, которые связаны с процессом подзарядки аккумуляторных батарей АНПА, предлагается следующая схема декомпозиции миссии, основанная на ожидаемом цикле ротации роботов (см. рисунок 2). Чтобы в дальнейшем избежать путаницы терминов, уточним, что под *рабочим периодом группы* мы понимаем временной отрезок между двумя последовательными групповыми сборами, а под *рабочим циклом АНПА* — период работы аппарата между двумя его последовательными подзарядками.

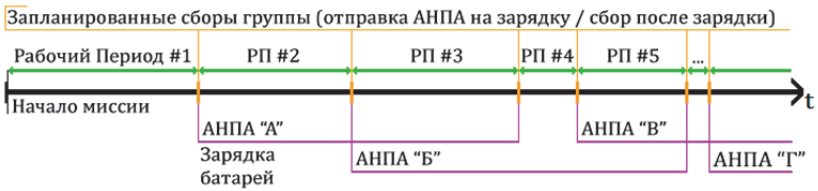


Рис. 2. Декомпозиция миссии на основе циклов зарядки АНПА

Возвращаясь к предложенной схеме, каждая точка разбиения миссии соответствует моменту сбора (рандеву) текущей группировки АНПА в заранее оговоренной области. Размер области сбора определяется в зависимости от предустановленной коммуникационной аппаратуры АНПА и глубины выполнения работ [19]. Рандеву включает в себя отправку нуждающихся в подзарядке аппаратов к док-станциям, сбор роботов, вернувшихся с зарядной базы, а также полную синхронизацию данных внутри группы и при необходимости обновление условий миссии. По завершении рандеву действующая группировка в обновленном составе возвращается к выполнению работ и осуществлению обследования целей миссии. При этом в процессе выполнения работ каждый робот группы в фоновом режиме

осуществляет поиск наилучшего группового маршрута уже на следующий рабочий период группы с учетом перегруппировки состава, предстоящей на ближайшем рандеву. Такое предварительное планирование позволяет одновременно сократить длительность каждого рандеву и распределить вычисления между аппаратами, обеспечив децентрализацию управления [20]. В этом случае полная синхронизация данных при каждом сборе должна дополнительно включать в себя обмен наилучшими найденными внутри группы решениями.

На рисунке 3 представлена схема функционирования группы АНПА на основе предложенного двухуровневого подхода. В соответствии с приведенной схемой, в процессе выполнения заданий на текущем рабочем периоде аппараты действующей группировки осуществляют планирование группового маршрута на следующий рабочий период группы согласно актуальному расписанию ротации. По прибытию в точку сбора в конце рабочего периода производится коммуникационный обмен между роботами группы, а также осуществляется отправка нуждающихся АНПА на зарядку и прием в группу уже зарядившихся аппаратов. В отсутствие незапланированных событий аппараты обмениваются наилучшими найденными решениями, выбирают среди них наиболее эффективный в качестве нового группового маршрута и начинают выполнение работ на новом рабочем периоде. В случае, когда были обнаружены существенные изменения в условиях миссии, инициируется сначала корректировка текущего расписания рабочих периодов, а затем генерация нового группового маршрута на ближайший рабочий период в соответствии с полученным от верхнего уровня расписанием ротации.

Таким образом, задача верхнеуровневого планировщика — регулировать действующий состав группы во времени, управляя рабочими циклами каждого аппарата. Поскольку от АНПА не требуется всегда работать в группе до полного разряда батарей и они могут покидать группу, имея некоторый энергозапас, а полная зарядка батарей также не обязательна (хоть и предпочтительна), то мы получаем возможность составлять расписание рабочих циклов АНПА, подстраивая циклы зарядки каждого робота с целью генерации единого группового расписания требуемого качества. Так, от расписания требуется, во-первых, удовлетворять критерию допустимости, то есть обеспечивать своевременную зарядку всех нуждающихся АНПА в отсутствие прецедентов выхода роботов из строя по причине нехватки энергии. Во-вторых, по возможности должна быть исключена одновременная зарядка большого количества аппаратов, так как это ведет к значительной потере производительности группировки, оставшейся для выполнения целей.

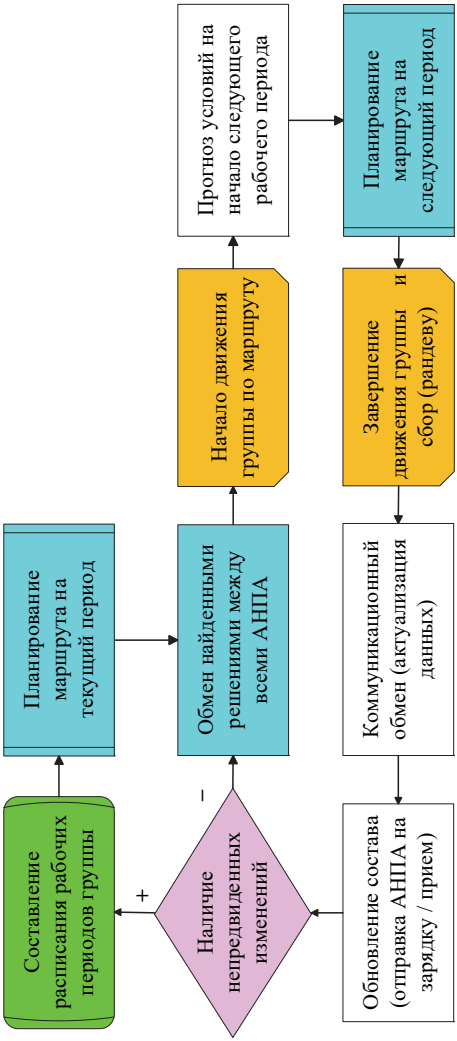


Рис. 3. Схема работы двухуровневой системы группового управления

В случае, если полностью избежать одновременной зарядки нескольких АНПА не представляется возможным, мы хотим добиться максимально равномерного распределения количества одновременно заряжающихся транспортных средств по времени. Для случая разнородной по скоростям группы роботов, этот критерий может быть обобщен как «равномерное распределение по времени суммарных крейсерских скоростей одновременно заряжающихся роботов». Так, для группы из трех роботов с крейсерскими скоростями 3 км/ч, 4 км/ч и 7 км/ч соответственно, при невозможности составить расписание с последовательными периодами подзарядки роботов пересечение периодов зарядки должно осуществляться предпочтительно для двух более медленных АНПА. В-третьих, в связи с тем, что каждый сбор группы ведет к временному прекращению проведения обследований, необходимо избегать избыточной частоты групповых рандеву. Уменьшение количества рандеву может быть осуществлено за счет объединения соседних близкорасположенных событий там, где это возможно, путем их сдвига к единой временной точке. Для примера, на рисунке 2 короткие рабочие периоды #4 и #6 (следующий за #5) могут быть исключены, если аппараты «C» и «D» будут отправлены на зарядку периодом раньше.

Алгоритмы, реализующие декомпозицию миссии на верхнем уровне должны быть простыми, быстрыми и надежными, чтобы обеспечивать оперативную корректировку с минимальным простоем группы после каждого непредвиденного события. В отличие от верхнего уровня, планировщик маршрутов на низком уровне не предназначен для «противодействия» динамическим изменениям среды, а преследует цель эффективного локального планирования групповых маршрутов и траекторий при действующих пространственно-временных ограничениях. Главная задача фазы маршрутизации — не только достичь своевременного обследования целей разнородной группой транспортных средств, но и обеспечить одновременное прибытие действующих АНПА к месту рандеву в конце рабочего периода. Кроме того, нельзя забывать, что задача маршрутизации на одном рабочем периоде не должна рассматриваться обособленно от глобальной задачи всей миссии, поскольку внутренние таймеры (время до следующего запланированного обследования) целей миссии не «останавливаются» на время проведения групповых сборов, что может привести к возникновению опозданий уже после того, как обследование целей будет продолжено на следующем рабочем периоде.

В силу того, что оба уровня предложенной системы планирования являются в равной степени значимыми для эффективного осуществления регулярного и своевременного

мониторинга в течение миссии, обеспечение должного уровня кооперации и синхронизации между планировщиками позволит конечной системе управления справляться с комплексными требованиями и ограничениями, присущими масштабным детализированным моделям реальных прикладных задач [4].

4. Планировщик миссий. Обобщая вышеизложенное, цель планировщика на верхнем уровне — сконструировать эффективное допустимое расписание рабочих периодов группы на основе циклов подзарядки АНПА с целью обеспечить максимальную производительность постоянно действующей группировки наряду с возможностью оперативного реагирования на любые непредвиденные изменения условий задачи.

Предполагается, что миссия должна выполняться в течение времени T значительной продолжительности, поэтому, в условиях высокой динамики среды и необходимости регулярных корректировок, разумным представляется построение расписания не на всю миссию, а на некоторый менее длительный интервал времени. Обозначим за T_P длительность такого интервала, который будем называть *периодом планирования*. Для получения долгосрочного группового расписания высокого качества, такой период должен включать по крайней мере несколько рабочих циклов каждого АНПА. В задачах составления расписаний большой размерности, как правило, применяется дискретизация пространства поиска, позволяющая ценой незначительных погрешностей одновременно облегчить кодирование решения и ускорить процесс его поиска. Таким образом, мы будем рассматривать период планирования как последовательность равных отрезков времени $T_P = \left\langle T^1, ..., \mathrm{T}^e \right\rangle$, $e = T_P / T_0$, где T_0 — длительность каждого такого отрезка.

В этом случае рабочее расписание единичного АНПА может быть представлено в виде e-мерного двоичного вектора, в котором i-ый элемент принимает значение нуля, когда аппарат работает вместе с группой на соответствующем отрезке времени T^i, а значение единицы во всех остальных случаях (аппарат находится на пути к зарядной базе, заряжается либо возвращается после зарядки). Расписание единичного робота считается допустимым, если временная длительность каждого из его рабочих циклов (последовательностей нулевых элементов) не превышает запас энергии аккумуляторов на момент завершения последней подзарядки. Таким образом, расписание группы будет представлять собой двоичную матрицу $H = \{h_{ij}\}$ размерности $e \times m$ (см. рисунок 4).

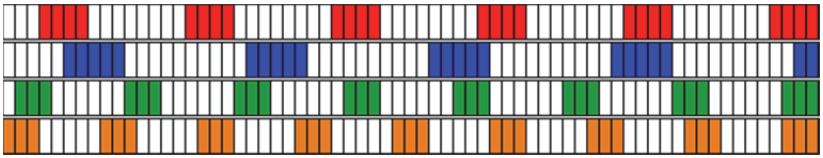


Рис. 4. Расписание ротации для группы из четырех аппаратов с различной емкостью аккумуляторов. Цветные клетки соответствуют периодам подзарядки

По сути, задача минимизации количества одновременно заряжающихся аппаратов на протяжении миссии является вариацией задачи составления циклов зарядки группы с ограничением на ресурсы пополнения аккумуляторов. При ограниченных возможностях подзарядки наилучшим образом показывает себя стратегия разрешения конфликтов (*conflict resolution*), когда действия агентов определяются одновременно двумя противоречащими поведенческими паттернами: уступать нуждающимся и быть жадными [21]. Жадность отвечает за самосохранение агента, стимулируя преждевременную подзарядку при наличии такой возможности (свободной док-станции), а способность уступать обеспечивает выживание группы, поощряя преждевременное прекращение пользования топливным ресурсом в пользу более нуждающихся агентов. При составлении критерия оценки групповых действий на верхнем уровне мы будем основываться на аналогичных принципах.

В нашем случае за самосохранение будет отвечать желание АНПА покинуть группу и пополнять запас энергии батарей независимо от их текущего уровня, уже не возвращаясь обратно для выполнения работ. Так как возврат любого АНПА к работе всегда инициирует проведение группового сбора с целью обновления состава, то этот критерий может быть сформулирован как требование минимизации количества групповых рандеву (пар $\left\langle T^i, T^{i+1} \right\rangle$, на которых хотя бы один из роботов меняет свой статус):

$$f_G(H) = \sum_{i=2}^{e} \left(1 - \prod_{j=1}^{m} (1 - \left| h_{ij} - h_{i-1\,j} \right|) \right) \to \min. \qquad (1)$$

Действуя по оптимальному расписанию согласно (1), все аппараты группы должны будут сохранять свой начальный статус (работа или зарядка) на протяжении всего периода планирования, чтобы не инициировать сборов действующей группировки. Однако, так как критерием допустимости расписания является поддержание всех АНПА в рабочем состоянии, то

оптимальным допустимым расписанием будет матрица единиц, то есть незамедлительная отправка всех АНПА на зарядку.

В свою очередь, альтруистический шаблон поведения [21], направленный на улучшение командной работы, будет заключаться в поощрении присутствия роботов в группировке, выполняющей цели миссии. Чтобы стимулировать такое поведение, мы будем штрафовать АНПА за время отсутствия в действующей группировке. В этом случае, функция оценки эффективности группового расписания H будет иметь следующий вид (v^j — крейсерские скорости движения аппаратов):

$$f_A(H) = \sum_{i=1}^{e} \left(\left(\sum_{j=1}^{m} h_{ij} \right) \left(\sum_{j=1}^{m} h_{ij} \cdot v^j \right) \right) \to \min. \qquad (2)$$

Штрафная функция (2) оценивает рабочие характеристики (в данном случае *скорость*) всех АНПА, отсутствующих в группе на каждом отрезке T^i текущего периода планирования вследствие ухода на подзарядку. Используя два множителя под знаком основной суммы, мы пытаемся одновременно исключить зарядку большого числа АНПА, а также подгрупп из самых быстрых аппаратов. В результате, при реализации командного поведения, планировщик будет минимизировать потерю производительности (2) всей группы на каждом отрезке времени T^i. На рисунке 5 слева представлен пример допустимого расписания рабочих циклов для группы из четырех АНПА, рассчитанный по критерию (2) с учетом требования по недопущению полной разрядки какого-либо аппарата. Очевидно, что многие рандеву здесь являются избыточными. Их можно исключить, если в процессе принятия решений дополнительно использовать описанный выше критерий жадного поведения.

Таким образом, ставится задача составления таких расписаний, которые в равной степени удовлетворяли бы сразу двум противоречащим критериям. При этом необходимо найти точный баланс между влиянием каждого из них на конечный результат. Как правило, это осуществляется нормированием двух величин относительно друг друга. Экспериментальным путем нами был определен следующий вид конечного критерия эффективности:

$$f(H) = f_G^2(H) \cdot f_A(H) \to \min. \qquad (3)$$

Используя те же входные данные для генерации расписания, что и в примере выше, оптимизация по критерию (3) позволила получить более качественное решение, обеспечивающее почти

двукратное уменьшение частоты групповых сборов при сохранении средней производительности работающей в ходе миссии группировки (см. рисунок 5 справа).

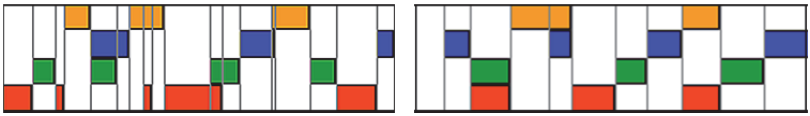


Рис. 5. Примеры рассчитанных расписаний по критериям (2) и (3) (вертикальными линиями отмечены моменты групповых сборов)

В общем случае задача составления расписания является NP-полной, а значит, применение эвристических методов для ее решения представляется наиболее разумным. Несмотря на большое разнообразие представленных в литературе алгоритмов составления расписаний и их модификаций, нацеленных на улучшение тех или иных аспектов задачи, наиболее часто применимыми на протяжении уже многих лет остаются генетические алгоритмы (ГА) ввиду их крайне высокой эффективности, в особенности на задачах большой размерности [22, 23]. Естественный параллелизм ГА позволяет легко реализовывать распределенное решение задачи на нескольких вычислительных узлах, в то время как «any-time» — природа алгоритма гарантирует получение некоторого допустимого решения в любой момент остановки вычислений.

Принимая во внимание вышесказанное, в качестве планировщика миссий на верхнем уровне мы предлагаем использование модификации классического генетического алгоритма (ГА), адаптированной под специфику задачи. При реализации ГА мы используем сжатое векторное представление матрицы-расписания H в качестве хромосомы (см. рисунок 6). Согласно такому представлению, на e-мерном векторе отмечаются только моменты смены статуса аппаратами группы, что позволяет значительно понизить размерность задачи и, следовательно, вычислительно-временные затраты на поиск ее решения. При расчете приспособленности хромосом сначала специализированная процедура восстанавливает решение в матричном виде, а уже затем определяет по нему значение целевой функции (3). Используемая процедура восстановления осуществляет не только декодирование решения, но и его локальную оптимизацию, объединяя или переупорядочивая смежные события. Допустимость всех хромосом проверяется в ходе работы алгоритма дополнительной алгоритмической процедурой, определяющей энергетическую потребность каждого аппарата на всех его рабочих циклах.

Мы используем классическую схему ГА, в которой на каждой итерации последовательно выполняются процедуры оценки всех хромосом популяции, отбора хромосом для создания потомства, скрещивания и мутации. При создании новых хромосом используются генетические операторы, которые, согласно [24], демонстрируют наилучшие результаты именно на задачах составления расписаний: двухточечное скрещивание, PPOX-скрещивание, случайная мутация и swap-мутация. Также применяется турнирная процедура отбора, реализованы принципы элитизма и параллельных популяций (островов).

Рис. 6. Одномерное представление группового расписания с рисунка 4

5. Планировщик маршрутов. Декомпозиция миссии, осуществляемая на верхнем уровне системы управления, позволяет значительно понизить размерность пространства поиска в задаче маршрутизации, а, следовательно, и необходимый объем вычислительных ресурсов, что является дополнительным преимуществом предлагаемого подхода. Ограниченная длительность каждого рабочего периода группы между двумя последовательными рандеву позволяет нам сформулировать задачу групповой маршрутизации. Входными параметрами задачи маршрутизации на каждом рабочем периоде служат характеристики всех целей миссии и аппаратов группы, обозначенные в постановке задачи (раздел 2), а также актуальная дорожная карта (граф *G)*. Начальное состояние каждого объекта миссии наследуется с окончания предыдущего рабочего периода группы, длительность нового рабочего периода составляет T_P.

Определим маршрут единичного АНПА как вектор-строку вида $r = \left\langle V_0, V_1^r, V_2^r, ..., V_h^r, V_0 \right\rangle$, представляющую собой индексированный список целей, предписанных текущему аппарату и расположенных в порядке их запланированного обхода. На каждом рабочем периоде все АНПА начинают движение из точки сбора группы (V_0) и возвращаются туда же по завершению выполнения своего маршрута. Необходимо отметить, что каждая цель может входить в маршрут одного робота более чем один раз, а также в маршруты нескольких роботов группы. Итоговый маршрут группы $R = \{r_1, ..., r_k\}$, $k \le m$ представляет собой совокупность маршрутов всех АНПА, действующих на текущем рабочем периоде. Таким образом, задача маршрутизации заключается в поиске такого группового маршрута, который обеспечивал бы:

– минимальный объем опозданий при посещении всех целей миссии, требующих обследования;

– прибытие всех действующих аппаратов в точку сбора в конце рабочего периода;

– благоприятные условия миссии (отсутствие целей с истекшим/истекающим внутренним таймером, что может привести к опозданиям) на момент завершения рабочего периода, которые будут «переданы» на следующий период работы группы.

Для оценки эффективности различных групповых маршрутов с учетом перечисленных требований и ограничений, предлагается использовать схему на основе списка сценариев желательного и нежелательного поведения группы. Список всех возможных сценариев должен быть составлен человеком-оператором предварительно и загружен на бортовые системы АНПА перед началом миссии. Каждый такой сценарий содержит свое значение приоритета относительно других сценариев, описание инициирующих его событий, тип сценария и условия начисления. Согласно предлагаемой схеме, оценка качества решения производится по результатам виртуального «прогона» группы по оцениваемому маршруту с начислением штрафных баллов и баллов поощрения, взвешенных согласно приоритетам соответствующих сценариев. Для рассматриваемой задачи регулярной маршрутизации мы предлагаем список из четырех сценариев, представленный в таблице 1.

Таблица 1. Список сценариев для групповой миссии по мониторингу

Приоритет	Инициирующие события (триггеры)	Тип сценария	Условия начисления
#1	АНПА прибывает к цели для обследования	Штраф	Наличие опозданий
#2	Сбор группы в конце рабочего периода	Штраф	Наличие целей с задержанным обследованием
#3	Сбор группы в конце рабочего периода	Штраф	Долгое прибытие всех АНПА
#4	АНПА прибывает к цели для обследования	Поощрение	Минимальное ожидание обследования

Для нормирования баллов, начисляемых по целям с различной периодичностью, мы применяем подход, предложенный нами в более ранних работах [5, 7] и использующий дополнительную функцию $a_i(t)$, определяющую актуальность (степень необходимости) обследования

цели в заданный момент времени. Такая функция актуальности ставится в соответствие каждой цели $i \in N$ и ведет себя согласно следующим трем правилам: обследование цели любым АНПА обнуляет значение ее актуальности; актуальность цели растет экспоненциально, достигая заданного порогового значения $\overline{a}$ (единого для всех целей) от нулевого значения за временной период p_i; в случае непосещения цели вовремя ее актуальность продолжает экспоненциально расти (см. рисунок 7).

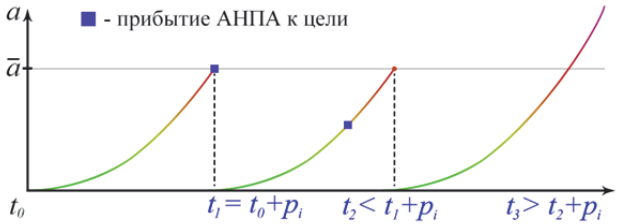


Рис. 7. Графическое представление изменения актуальности цели во времени

Определим функцию штрафа по первому сценарию из представленного выше списка (запоздалое обследование цели):

$$\varphi(i,t) = \begin{cases} a_i(t) - \overline{a}, a_i(t) > \overline{a} \\ 0, \quad a_i(t) \le \overline{a} \end{cases}. \tag{4}$$

Аналогично определим штрафную функцию по второму сценарию (наличие целей с задержанным обследованием на момент завершения группового движения):

$$\phi(i, t_0 + T_P) = a_i(t_0 + T_P), \tag{5}$$

где момент t_0 соответствует началу текущего рабочего периода. Фактически, использование функции (5) позволяет обеспечить генерируемые маршруты сразу несколькими полезными свойствами: во-первых, таким образом группе запрещается пропускать и игнорировать требующие обследования цели; во-вторых, значение функции (5) соответствует уровню благоприятности начальных условий для группы на следующем рабочем периоде; наконец, в-третьих, использование (5) косвенным образом нормирует длительности маршрутов всех роботов группы. Таким образом, нам не требуется вводить отдельную штрафную функцию для третьего сценария, поскольку он уже учитывается в (5).

Функция поощрения по четвертому сценарию определяется следующим образом (параметр ε задает ту максимальную длительность простоя, которую мы можем классифицировать как незначительную):

$$\psi(i,t)=\begin{cases} a_i(t)+\varepsilon-\overline{a}\,, 0\le \overline{a}-a_i(t)\le \varepsilon \\ 0 \quad , \quad \overline{a}-a_i(t)\le 0 \\ 0 \quad , \quad \overline{a}-a_i(t)\ge \varepsilon \end{cases}. \tag{6}$$

Стоит отметить, что функция актуальности (рисунок 7) построена таким образом, чтобы ее значение для целей с меньшей периодичностью росло быстрее. В этом случае при невозможности своевременного посещения всех целей миссии обследования с опозданием будут приходиться преимущественно на задания с большой периодичностью, представляющие, предположительно, меньший интерес.

Возвращаясь к процедуре планирования, необходимо напомнить, что групповой маршрут на текущем рабочем периоде группы всегда рассчитывается во время предыдущего рабочего периода с учетом ожидаемых изменений в составе действующей группы на ближайшем рандеву. Возникновение каких-либо непредвиденных событий в течение рабочего периода, как правило, ведет к потере актуальности всех заранее рассчитанных решений, однако, они все еще могут выступать в качестве своего рода базы знаний для более интеллектуального перепланирования. Тем не менее при наличии свободных вычислительных ресурсов мы можем частично устранить влияние отдельных непрогнозируемых событий за счет заблаговременной генерации дополнительных резервных планов. Например, параллельно может обсчитываться резервный маршрут для группы меньшего размера на случай потери или выхода из строя одного из роботов. В этом случае целесообразно строить маршрут с прогнозом на отсутствие наиболее быстрого АНПА в группе, так как он сможет заменить любой другой отсутствующий аппарат без потери эффективности. При необходимости для выявления наиболее вероятных изменений, по которым может потребоваться расчет резервных маршрутов, могут быть использованы более продвинутые техники диагностики и ситуационной осведомленности [25].

Планирование на нижнем уровне является гораздо более сложной и трудоемкой задачей, поскольку уже изначально NP-трудный поиск оптимального маршрута для фиксированного флота (*fixed fleet*) транспортных средств здесь осложняется комплексным набором действующих ограничений и, следовательно,

плохой окрестностной структурой задачи, что затрудняет поиск и прогнозирование качественных допустимых решений, поскольку они могут не находиться в окрестности других эффективных или допустимых решений в пространстве поиска. Таким образом, не существует алгоритмов, которые бы решали такую задачу за полиномиальное время, что приводит нас к классу приближенных эвристических алгоритмов, которые позволяют получать близкие к оптимальным решения за приемлемое вычислительное время.

За последние 10 лет эвристические и метаэвристические подходы к решению ЗМТ получили стремительное развитие во многом благодаря появлению гибридных алгоритмов, сочетающих в себе несколько изначально независимых вычислительных процедур. Лучшие метаэвристики применяют сложные стратегии поиска по окрестностям, принципы работы точных методов оптимизации и декомпозиционные схемы. Эвристики также становятся более гибкими и могут применяться к широкому диапазону вариаций ЗМТ без каких-либо структурных изменений. Современные методы реализуют высокоуровневые стратегии управления, базируясь на различных структурах памяти (хромосомы, феромоны, нейроны) и опираясь на блок локального поиска, направляющего работу алгоритма в перспективные области пространства поиска [26].

Эволюционные методы неоднократно доказывали свою эффективность при решении различных вариаций ЗМТ, включая маршрутизацию с временными окнами, комплексную маршрутизацию и ряд других постановок, пересекающихся с рассматриваемой в статье задачей. Основным преимуществом эволюционных алгоритмов (ЭА) является их способность строить решения для постановок со сложным набором ограничений, так как они требуют относительно небольшое количество информации о природе самой задачи. Согласно представленным в литературе аналитическим исследованиям [27, 28], гибридные эволюционные алгоритмы позволяют находить близкие к оптимальным решения с лучшей масштабируемостью и за лучшее время, чем в среднем любые другие эвристические и метаэвристические подходы. Кроме того, эволюционные алгоритмы, хорошо зарекомендовавшие себя при решении различных задач многомерной оптимизации, в последнее время приобрели актуальность в решении задач группового управления мобильными роботами [29]. Эффективность гибридных ЭА обеспечивается сочетанием генетических операторов, учитывающих специфику задачи, с правильно подобранными процедурами локального поиска и аккуратной балансировкой между исследованием пространства поиска (*exploration*) и его разработкой (*exploitation*), что позволяет

избежать преждевременной сходимости алгоритма. Еще одной общей чертой, объединяющей успешные реализации ЭА, является отказ от работы с исключительно допустимыми решениями в пользу процедуры наложения штрафов.

На основании вышеизложенного, для решения задачи групповой маршрутизации предлагается использовать оригинальный гибридный эволюционный подход, включающий в себя специализированные генетические операторы, эвристики локального поиска и набор дополнительных процедур для повышения эффективности работы алгоритма при комплексных пространственно-временных ограничениях и в условиях большой размерности задачи. Блок-схема разработанного алгоритма представлена на рисунке 8, цветом отмечены модифицированные блоки ЭА.

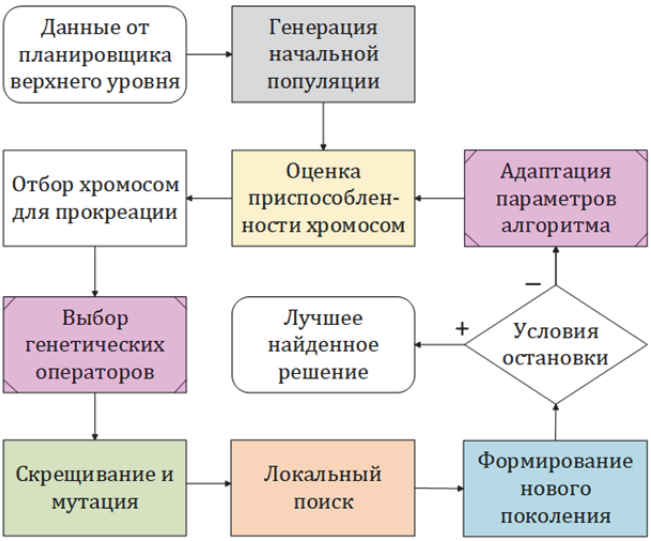


Рис. 8. Блок схема гибридного эволюционного алгоритма

Групповой маршрут R выступает здесь в качестве хромосомы, а формулы (4)-(6) обеспечивают оценку приспособленности с учетом приоритетов соответствующих им сценариев. Генерация начальной популяции хромосом является первым и определяющим шагом к быстрому получению допустимых решений приемлемого качества. Цель этого шага — сконструировать качественный начальный набор решений-хромосом, равномерно покрывающий пространство поиска. Это требование достигается одновременным использованием трех различных конструктивных эвристик: простой случайной

вставки (*insertion heuristic*) и двух вариаций жадной эвристики «*time-oriented nearest-neighbor*».

Все полученные решения проходят процедуру оценки, после чего, по результатам ранжирования, производится турнирная селекция решений для воспроизводства потомства. Мы предлагаем использование сразу нескольких специализированных генетических операторов: два скрещивания и многорежимную мутацию. В качестве операторов скрещивания используются модификации двухточечного кроссовера и оператора адаптивной памяти (*adaptive memory*) [30]. В свою очередь, многорежимная мутация состоит из четырех операторов: добавление новой цели в маршрут, удаление цели из маршрута, изменение цели и смена двух целей местами. Полученные решения-потомки подвергаются локальному поиску, где с помощью процедуры спуска с чередующимися окрестностями (*variable neighborhood descent*) [31] производится их направленное улучшение. Механизмы параллельных популяций с миграцией и элитизма применяются на этапе формирования новой популяции, обеспечивая выход из локальных оптимумов, а процедура удаления клонов обеспечивает разнообразие новых популяций.

Для повышения эффективности механизма создания новых популяций в целом, параметры, задающие вероятность применения генетических операторов, корректируются на каждой итерации алгоритма. Эти изменения основываются на оценке текущей способности каждого из операторов привести к улучшению решений, что позволяет осуществлять непрерывную адаптацию алгоритма на всех этапах вычислений. Использование такого механизма самоадаптации позволяет значительно увеличить скорость работы алгоритма в тех случаях, когда одни операторы начинают работать заметно лучше других.

Подробное описание структуры и схемы работы всех описанных эвристик, генетических операторов и других вычислительных процедур может быть найдено в наших предыдущих работах [7, 8], посвященных исключительно задачам групповой маршрутизации на низком уровне, поэтому в рамках данной статьи мы не будем вдаваться в детали.

На рисунке 9 приведен пример группового маршрута, построенного разработанным гибридным эволюционным алгоритмом. На изображенном примере группа из трех АНПА осуществляет регулярный мониторинг семи целей, представленных на рисунке горизонтальными осями, различной периодичности (от 10 до 17 минут). Линиями отмечены перемещения аппаратов между целями, а прямоугольники на осях соответствуют периодам обследования целей аппаратами группы: полностью прозрачные отвечают

своевременному посещению цели, а частично закрашенные — обследованию с опозданием.

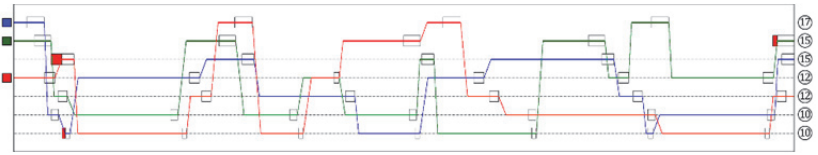


Рис. 9. Построенный групповой маршрут для трех АНПА на семи целях различной периодичности

6. Реализация. Описанные планировщики для верхнего и нижнего уровня были программно реализованы на языке C++ в виде системы группового управления предложенной структуры и включены в состав разрабатываемого нами моделирующего комплекса «AUV Multiobjective Mission Planner» (см. рисунок 10) для проведения вычислительных экспериментов. Для построения дорожной карты на основе заданного рельефа был использован алгоритм поиска и оценки длин путей на графе HPA* (*hierarchical pathfinding A-star*), позволяющий при малых вычислительных затратах получать близкие к оптимальным оценки как для фиксированного множества целей, так и при появлении новых объектов.

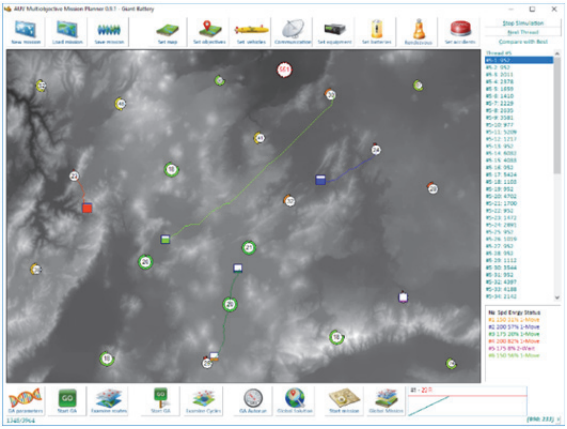


Рис. 10. Снимок главного окна моделирующего комплекса

Поскольку исследуемые задачи планирования являются оригинальными и не имеют прямых аналогов в литературе, судить об эффективности работы разработанных методов мы можем либо в сравнении с найденными для тестовых задач оптимальными

решениями, либо на основании собственных статистических данных, полученных при многократном запуске алгоритмов. Для задачи планирования на верхнем уровне была реализована процедура, основанная на методе ветвей и границ, которая позволяет путем сокращения перебора отсекать подмножества всех возможных решений задачи, которые являются хуже, чем построенное генетическим алгоритмом групповое расписание. Таким образом, для каждого полученного в результате вычислительных экспериментов решения может быть найдена оценка снизу на процент заведомо менее эффективных решений. Результаты тестирования приведены в таблице 2 (для задачи со звездочкой найдено оптимальное решение).

Таблица 2. Эффективность работы верхнеуровневого планировщика

Длина расписания (e)	Размер группы АНПА (k)	Оценка снизу на качество решения	
		3 секунды вычислений	30 секунд вычислений
20	3*	100%	100%
	5	99,1%	99,9%
	10	98,8%	99,9%
100	3	98%	99,5%
	5	97,5%	98,35%
	10	97,3%	98,25%
1000	3	96,7%	98,1%
	5	96,2%	97,9%
	10	95,9%	97,5%

Результаты, приведенные в таблице 2, получены при расчетах на одном ядре процессора Intel Core 2 Duo E6750 2,66 Ггц. Отсечка в 30 секунд выбрана экспериментальным путем как время, за которое может быть получено приемлемое допустимое решение, дальнейшее улучшение которого требует уже заметно бо́льших временных затрат.

На рисунке 11 приведен пример рассчитанного расписания ротации для группы из четырех АНПА со следующими параметрами: период планирования $T_p = 50000$ секунд ($T_0 = 500$), длина расписания $e = 100$; скорости аппаратов $v^1 = 1.5$ м/с, $v^2 = 2$ м/с, $v^3 = 1.75$ м/с и $v^4 = 1$ м/с; емкости батарей АНПА рассчитаны на 8000 секунд работы; расстояние до единственной зарядной базы 900м; скорость зарядки $c = 2$. Выделенные на рисунке цветом отрезки соответствуют периодам выхода аппаратов из группы с целью пополнения заряда батарей, а вертикальными линиями отмечены точки сбора

действующей группировки АНПА с целью обновления состава. На рисунке 12 представлен график изменения заряда батарей всех роботов в группе при работе по расписанию, изображенному на рисунке 11. Черным цветом отмечены те отрезки времени, которые тратятся аппаратами на перемещение до зарядной док-станции и на возвращение обратно в группу.

Рис. 11. Расписание рабочих циклов для группы из четырех АНПА

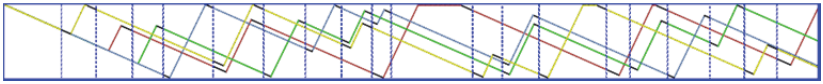


Рис. 12. График уровня заряда батарей АНПА в ходе выполнения миссии

Критерием эффективности предложенного подхода в целом является оценка его влияния на качество конечных групповых маршрутов и скорость их планирования. Для этого мы сравним результаты работы группы на длительном отрезке времени согласно маршрутам, полученным двумя различными способами (см. рисунок 13): первый целиком построен низкоуровневым планировщиком без учета групповых сборов (верхний график); второй состоит из трех последовательных маршрутов меньшей длины с промежуточными сборами группы (нижний график). Графики на рисунке 13 отслеживают изменение максимального значения актуальности среди целей миссии во времени. Горизонтальная линия на каждом графике — пороговое значение $\overline{a}$, а вертикальными линиями отмечены точки групповых рандеву. Так как планировщик на нижнем уровне не приспособлен для построения маршрутов, включающих подзарядку, то в данном примере мы не учитываем ограничение на емкость батарей, а ротация группы отсутствует. Рассматриваемый пример содержит 30 целей, которые обследуются группой из пяти АНПА. Период планирования составляет 13000 секунд, что включает 150 запланированных обследований. Рассматривается наихудший случай, когда область сбора задана в виде точки. Выбранный пример относится к категории задач высокой сложности, для которых нами не было найдено решений, обеспечивающих полное выполнение всех целей без опозданий.

Применение только низкоуровневого планировщика позволило за 342 секунды расчетов получить маршрут, обеспечивающий 357 секунд суммарной задержки обследований, в то время как предложенная схема

на основе групповых рандеву обеспечила инспекцию целей с опозданием в 392 секунды при общей длительности расчетов всего в 48 секунд. Как можно видеть на рисунке 13, главной причиной увеличения объема опозданий во втором маршруте является необходимость групповых сборов в области малого размера и в условиях плотного расписания инспекций целей. При этом в масштабах длительности миссии разница в эффективности двух маршрутов (0,26% в данном примере; 0,37% в среднем на всех тестовых примерах) может считаться несущественной, а предложенный подход при значительно меньших вычислительных затратах обеспечивает еще и коммуникационную устойчивость движения.

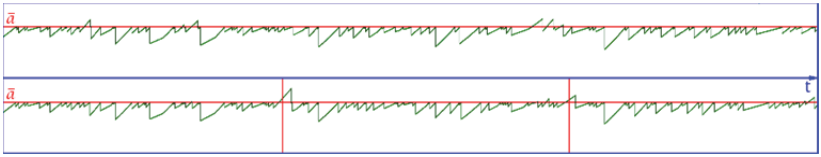


Рис. 13. Сравнение эффективности обследования целей при долгосрочном (наверху) и краткосрочном (внизу) планировании

7. Заключение. В статье представлен двухуровневый подход к динамической задаче регулярной маршрутизации АНПА, в рамках которого реализованы алгоритм планирования расписания групповых рандеву, обеспечивающий эффективную декомпозицию миссии, и процедура распределения целей и генерации маршрутов движения действующей группы на каждом рабочем периоде.

Была проведена серия вычислительных экспериментов, в рамках которых моделировалось выполнение заданной многозадачной миссии группой АНПА с непостоянным составом в условиях комплексных требований и пространственно-временных ограничений, вытекающих из регулярного характера задач миссии. Результаты моделирования продемонстрировали высокую эффективность предлагаемого подхода: долгосрочный планировщик на верхнем уровне обеспечивает быструю и надежную динамическую генерацию качественных расписаний группы, в то время как низкоуровневая система маршрутизации осуществляет интеллектуальную процедуру рационального распределения целей и построения детальных маршрутов при строгих ограничениях и большой размерности задачи. Предложенная схема взаимодействия между двумя планировщиками позволяет группе сохранять свою эффективность в постоянно меняющейся среде и при непрерывной ротации группировки.

Дальнейшее развитие исследуемых в статье моделей связано с добавлением в постановку функциональной разнородности

транспортных средств по типам предустановленного бортового исследовательского оборудования. В этом случае каждый аппарат группы будет в состоянии выполнять лишь те задания, техническим требованиям которых он удовлетворяет. При управлении подобной гетерогенной группировкой АНПА верхний уровень системы управления должен будет помимо рассмотренных в данной работе критериев дополнительно обеспечивать доступность любого типа оборудования на каждом рабочем периоде группы.

Литература

1. *Blidberg D.R.* The Development of Autonomous Underwater Vehicles (AUV); A Brief Summary // Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). 2001. vol. 4. pp. 1.
2. *Deng Y., Beaujean P.P.J., An E., Carlson E.* Task allocation and path planning for collaborative AUVs operating through an underwater acoustic network // OCEANS 2010 MTS/IEEE SEATTLE. 2010. pp. 1–9.
3. *Агеев М.Д. и др.* Автономные подводные роботы. Системы и технологии // М.: Наука. 2005. 398 с.
4. *Zadeh S.M., Powers D.M.W., Yazdani A.M.* Development of an Autonomous Reactive Mission Scheduling and Path Planning (ARMSP) Architecture Using Evolutionary Algorithms for AUV Operation in a Sever Ocean Environment // Computing Research Repository. 2016. 22 p.
5. *Kenzin M.Yu., Bychkov I.V., Maksimkin N.N.* A hybrid approach to solve the dynamic patrol routing problem for group of underwater robots // 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). 2016. pp. 1114–1119.
6. *Zeng Z. et al.* Path planning for rendezvous of multiple AUVs operating in a variable ocean // 2014 IEEE 4th Annual International Conference on Cyber Technology in Automation, Control, and Intelligent Systems. 2014. pp. 451–456.
7. *Kenzin M.Yu., Bychkov I.V., Maksimkin N.N.* Hybrid evolutionary approach to multi-objective mission planning for group of underwater robots // International Conference on Computational and Information Technologies in Science, Engineering and Education. 2015. pp. 73–84.
8. *Kenzin M., Bychkov I., Maksimkin N.* Task allocation and path planning for network of autonomous underwater vehicles // International Journal of Computer Networks & Communications (IJCNC). 2018. vol. 10. no. 2. pp. 33–42.
9. *Ozaslan T. et al.* Inspection of penstocks and featureless tunnel-like environments using micro UAVs // Field and Service Robotics. 2015. pp. 123–136.
10. *Christofides N.* The Vehicle Routing Problem // Revue Française d'Automatique, Informatique, Recherche Opérationelle: Recherche Opérationelle. 1976. vol. 10(V1). pp. 55–70.
11. *Chevaleyre Y.* Theoretical Analysis of the multi-agent patrolling problem // Proceedings of the IEEE/WIC/ACM International Conference on Intelligent Agent Technology. 2004. pp. 302–308.
12. *Basilico N., Gatti N., Villa F.* Asynchronous multi-robot patrolling against intrusions in arbitrary topologies // Proceedings of the 24th AAAI Conference on Artificial Intelligence. 2010. pp. 1224–1229.
13. *Fargeas J.L., Hyun B., Kabamba P., Girard A.* Persistent visitation under revisit constraints // 2013 International Conference on Unmanned Aircraft Systems (ICUAS). 2013. pp. 952–957.

14. *Stump E., Michael N.* Multi-robot persistent surveillance planning as a Vehicle Routing Problem // 2011 IEEE International Conference on Automation Science and Engineering. 2011. vol. 1. pp. 569–575.
15. *Manyam S.G. et al.* Multi-UAV routing for persistent intelligence surveillance & reconnaissance missions // 2017 International Conference on Unmanned Aircraft Systems (ICUAS). 2017. pp. 573–580.
16. *Leahy K. et al.* Persistent surveillance for unmanned aerial vehicles subject to charging and temporal logic constraints // Autonomous Robots. 2016. vol. 40. no. 8. pp. 1363–1378.
17. *Drucker N., Penn M., Strichman O.* Cyclic Routing of Unmanned Aerial Vehicles // 13th International Conference on Integration of AI and OR Techniques in Constraint Programming (CPAIOR 2016). 2016. vol. 9676. pp. 125–141.
18. *Hartl R.F., Hasle G., Janssens G.K.* Special Issue on Rich Vehicle Routing Problems // Central European Journal of Operations Research. 2006. vol. 14. no. 2. pp. 103–104.
19. *Вершинин А.С.* Экспериментальная оценка скорости передачи данных макета гидроакустического модема // Труды СПИИРАН. 2016. Вып. 3(46). С. 40–48.
20. *Каляев А.И., Каляев И.А.* Метод децентрализованного управления группой роботов при выполнении потока заданий // Робототехника и техническая кибернетика. 2015. Вып. 1(6). С. 26–35.
21. *Sempe F., Munoz A., Drogoul A.* Autonomous Robots Sharing a Charging Station with no Communication: a Case Study // Distributed Autonomous Robotic Systems. 2002. vol. 5. pp. 91–100.
22. *Calis B., Bulkan S.* A research survey: review of AI solution strategies of job shop scheduling problem // Journal of Intelligent Manufacturing. 2015. vol. 26. no. 5. pp. 961–973.
23. *Vincent L., Durai C.* A survey on various optimization techniques with respect to fexible job shop scheduling // International Journal of Scientifc and Research Publications. 2014. vol. 4. no. 3. pp. 1–7.
24. *Amjad M.K. et al.* Recent Research Trends in Genetic Algorithm Based Flexible Job Shop Scheduling Problems // Mathematical Problems in Engineering. 2018. vol. 2018. pp. 1–32.
25. *Liu C., Coombes M., Li B., Chen W-H.* Enhanced situation awareness for unmanned aerial vehicle operating in terminal areas with circuit flight rules // Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering. 2016. vol. 230. no. 9. pp. 1683–1693.
26. *Laporte G., Ropke S., Vidal T.* Chapter 4: Heuristics for the Vehicle Routing Problem // Vehicle Routing: Problems, Methods, and Applications, Second Edition. 2014. pp. 87–116.
27. *Braysy O., Gendreau M.* Vehicle routing problem with time windows, part I: route construction and local search algorithms // Transportation science. 2005. vol. 39. no. 1. pp. 104–118.
28. *Koc C., Bektas T., Jabali O., Laporte G.* Thirty years of heterogeneous vehicle routing // European Journal of Operational Research. 2016. vol. 249. no. 1. pp. 1–21.
29. *Пшихопов В.Х., Медведев М.Ю.* Групповое управление движением мобильных роботов в неопределенной среде с использованием неустойчивых режимов // Труды СПИИРАН. 2018. Вып. 5(60). С. 39–63.
30. *Yong M.* Solving vehicle routing problem with time windows with hybrid evolutionary algorithm // 2010 Second WRI Global Congress on Intelligent Systems (GCIS). 2010. vol. 1. pp. 335–339.
31. *Affi M., Derbel H., Jarboui B.* Variable neighborhood search algorithm for the green vehicle routing problem // International Journal of Industrial Engineering Computations. 2018. vol. 9. no. 2. pp. 195–204.

Бычков Игорь Вячеславович — д-р техн. наук, профессор, академик РАН, директор, Федеральное государственное бюджетное учреждение науки Институт динамики систем и теории управления им. В.М. Матросова (ИДСТУ СО РАН). Область научных интересов: искусственный интеллект, геоинформационные системы, системы интеллектуального анализа данных, математическое моделирование. Число научных публикаций — 295. idstu@icc.ru, www.idstu.irk.ru; ул. Лермонтова, 134, а/я 292, Иркутск, 664033, РФ; р.т. +7(3952)42-71-00, факс +7(3952)51-16-16.

Кензин Максим Юрьевич — младший научный сотрудник лаборатории информационно-управляющих систем, Федеральное государственное бюджетное учреждение науки Институт динамики систем и теории управления им. В.М. Матросова (ИДСТУ СО РАН). Область научных интересов: методы оптимального управления роботами, групповое управление, задачи маршрутизации транспорта, поиск пути, искусственный интеллект. Число научных публикаций — 41. gorthauers@gmail.com, www.idstu.irk.ru; ул. Лермонтова, 134, а/я 292, Иркутск, 664033, РФ; р.т. +7(3952)45-30-85, факс +7(3952)51-16-16.

Максимкин Николай Николаевич — канд. техн. наук, ведущий научный сотрудник лаборатории информационно-управляющих систем, Федеральное государственное бюджетное учреждение науки Институт динамики систем и теории управления им. В.М. Матросова (ИДСТУ СО РАН). Область научных интересов: методы исследования сложных систем, групповое управление, искусственный интеллект. Число научных публикаций — 81. mnn@icc.ru, www.idstu.irk.ru; ул. Лермонтова, 134, а/я 292, Иркутск, 664033, РФ; р.т. +7(3952)45-30-05, факс +7(3952)51-16-16.

Поддержка исследований. Работа выполнена при финансовой поддержке РНФ (проект № 16-11-00053).

DOI 10.15622/sp.18.2.267-301

I.V. BYCHKOV, M.Yu. KENZIN, N.N. MAKSIMKIN

TWO-LEVEL EVOLUTIONARY APPROACH TO PERSISTENT SURVEILLANCE FOR MULTIPLE UNDERWATER VEHICLES WITH ENERGY CONSTRAINTS

Bychkov I.V., Kenzin M.Yu., Maksimkin N.N. **Two-Level Evolutionary Approach to Persistent Surveillance for Multiple Underwater Vehicles with Energy Constraints.**

Abstract. Currently, the coordinated use of autonomous underwater vehicles groups seems to be the most promising and ambitious technology to provide a solution to the whole range of oceanographic problems. Complex and large-scale underwater operations usually involve long stay activities of robotic groups under the limited vehicle's battery capacity. In this context, available charging station within the operational area is required for long-term mission implementation. In order to ensure a high level of group performance capability, two following problems have to be handled simultaneously and accurately – to allocate all tasks between vehicles in the group and to determine the recharging order over the extended period of time. While doing this, it should be taken into account, that the real world underwater vehicle systems are partially self-contained and could be subjected to any malfunctions and unforeseen events.

The article is devoted to the suggested two-level dynamic mission planner based on the rendezvous point selection scheme. The idea is to divide a mission on a series of time-limited operating periods with the whole group rendezvous at the end of each period. The high-level planner's objective here is to construct the recharging schedule for all vehicles in the group ensuring well-timed energy replenishment while preventing the simultaneous charging of a plenitude of robots. Based on this schedule, mission is decomposed to assign group rendezvous to each regrouping event (robot leaving the group for recharging or joining the group after recharging). This scheme of periodic rendezvous allows group to keep up its status regularly and to re-plan current strategy, if needed, almost on-the-fly. Low-level planner, in return, performs detailed group routing on the graph-like terrain for each operating period under vehicle's technical restrictions and task's spatiotemporal requirements. In this paper, we propose the evolutionary approach to decentralized implementation of both path planners using specialized heuristics, solution improvement techniques, and original chromosome-coding scheme. Both algorithm options for group mission planner are analyzed in the paper; the results of computational experiments are given.

Keywords: Autonomous Underwater Vehicles, Group Control, Scheduling Problem, Vehicle Routing Problem, Evolutionary Algorithms.

Bychkov Igor Vyacheslavovich — Ph.D., Dr. Sci., Associate Professor, Academician of RAS, Head of Matrosov Institute for System Dynamics and Control Theory of Siberian Branch of Russian Academy of Sciences (IDSTU SB RAS). Research interests: artificial intelligence, geographic information systems, intellectual data analysis, mathematical modeling. The number of publications — 295. IDSTU@icc.ru , www.idstu.irk.ru; 134, Lermontov str., Irkutsk, 664033, Russia; office phone: +7(3952)42-71-00, fax +7(3952)51-16-16.

Kenzin Maksim Yurievich — Junior Researcher of Information and Control Systems Laboratory, Matrosov Institute for System Dynamics and Control Theory of Siberian Branch of Russian Academy of Sciences (IDSTU SB RAS). Research interests: optimal robot motion planning, group control, vehicle routing problem, pathfinding, artificial intelligence. The number of publications — 41. GORTHAUERS@gmail.com, www.idstu.irk.ru; 134, Lermontov str., Irkutsk, 664033, Russia; office phone +7(3952)45-30-85, fax +7(3952)51-16-16.

Maksimkin Nikolai Nikolayevich — Ph.D., Leading Researcher of Information and Control Systems Laboratory, Matrosov Institute for System Dynamics and Control Theory of Siberian Branch of Russian Academy of Sciences (IDSTU SB RAS). Research interests: dynamics of complex systems, group control, artificial intelligence. The number of publications — 81. MNN@icc.ru, www.idstu.irk.ru; 134, Lermontov str., Irkutsk, 664033, Russia; office phone +7(3952)45-30-05, fax +7(3952)51-16-16.

Acknowledgements. This research is supported by RSF (grant 16-11-00053).

References

1. Blidberg D.R. The Development of Autonomous Underwater Vehicles (AUV); A Brief Summary. Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). 2001. vol. 4. pp. 1.
2. Deng Y., Beaujean P.P.J., An E., Carlson E. Task allocation and path planning for collaborative AUVs operating through an underwater acoustic network. OCEANS 2010 MTS/IEEE SEATTLE. 2010. pp. 1–9.
3. Ageev M.D. et al. *Avtonomnie podvodnie roboty. Sistemy i tekhnologii* [Autonomous Underwater Robots: Systems and Technologies]. M.: Nauka. 2005. 398 p. (In Russ.).
4. Zadeh S.M., Powers D.M.W., Yazdani A.M. Development of an Autonomous Reactive Mission Scheduling and Path Planning (ARMSP) Architecture Using Evolutionary Algorithms for AUV Operation in a Sever Ocean Environment. Computing Research Repository. 2016. 22 p.
5. Kenzin M.Yu., Bychkov I.V., Maksimkin N.N. A hybrid approach to solve the dynamic patrol routing problem for group of underwater robots. 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). 2016. pp. 1114–1119.
6. Zeng Z. et al. Path planning for rendezvous of multiple AUVs operating in a variable ocean. 2014 IEEE 4th Annual International Conference on Cyber Technology in Automation, Control, and Intelligent Systems. 2014. pp. 451–456.
7. Kenzin M.Yu., Bychkov I.V., Maksimkin N.N. Hybrid evolutionary approach to multi-objective mission planning for group of underwater robots. International Conference on Computational and Information Technologies in Science, Engineering and Education. 2015. pp. 73–84.
8. Kenzin M., Bychkov I., Maksimkin N. Task allocation and path planning for network of autonomous underwater vehicles. *International Journal of Computer Networks & Communications (IJCNC)*. 2018. vol. 10. no. 2. pp. 33–42.
9. Ozaslan T. et al. Inspection of penstocks and featureless tunnel-like environments using micro UAVs. Field and Service Robotics. 2015. pp. 123–136.
10. Christofides N. The Vehicle Routing Problem. *Revue Française d'Automatique, Informatique, Recherche Opérationelle: Recherche Opérationelle.* 1976. vol. 10(V1). pp. 55–70.
11. Chevaleyre Y. Theoretical Analysis of the multi-agent patrolling problem. Proceedings of the IEEE/WIC/ACM International Conference on Intelligent Agent Technology. 2004. pp. 302–308.
12. Basilico N., Gatti N., Villa F. Asynchronous multi-robot patrolling against intrusions in arbitrary topologies. Proceedings of the 24th AAAI Conference on Artificial Intelligence. 2010. pp. 1224–1229.
13. Fargeas J.L., Hyun B., Kabamba P., Girard A. Persistent visitation under revisit constraints. 2013 International Conference on Unmanned Aircraft Systems (ICUAS). 2013. pp. 952–957.
14. Stump E., Michael N. Multi-robot persistent surveillance planning as a Vehicle Routing Problem. 2011 IEEE International Conference on Automation Science and Engineering. 2011. vol. 1. pp. 569–575.

15. Manyam S.G. et al. Multi-UAV routing for persistent intelligence surveillance & reconnaissance missions. 2017 International Conference on Unmanned Aircraft Systems (ICUAS). 2017. pp. 573–580.
16. Leahy K. et al. Persistent surveillance for unmanned aerial vehicles subject to charging and temporal logic constraints. *Autonomous Robots.* 2016. vol. 40. no. 8. pp. 1363–1378.
17. Drucker N., Penn M., Strichman O. Cyclic Routing of Unmanned Aerial Vehicles. 13th International Conference on Integration of AI and OR Techniques in Constraint Programming (CPAIOR 2016). 2016. vol. 9676. pp. 125–141.
18. Hartl R.F., Hasle G., Janssens G.K. Special Issue on Rich Vehicle Routing Problems. *Central European Journal of Operations Research*. 2006. vol. 14. no. 2. pp. 103–104.
19. Vershinin A.S. [Experimental Estimation of the Data Transfer Rate of a Hydroacoustic Modem Model]. *Trudy SPIIRAN – SPIIRAS Proceedings*. 2016. vol. 3(46). pp. 40–48. (In Russ.).
20. Kalyaev A.I., Kalyaev I.A. [Method of decentralized control of robot group during execution of task flow]. *Robototehnika i tehnicheskaya kibernetika – Robotics and Technical Cybernetics*. 2015. vol. 1. no. 6. pp. 26–35. (In Russ.).
21. Sempe F., Munoz A., Drogoul A. Autonomous Robots Sharing a Charging Station with no Communication: a Case Study. Distributed Autonomous Robotic Systems. 2002. vol. 5. pp. 91–100.
22. Calis B., Bulkan S. A research survey: review of AI solution strategies of job shop scheduling problem. *Journal of Intelligent Manufacturing.* 2015. vol. 26. no. 5. pp. 961–973.
23. Vincent L., Durai C. A survey on various optimization techniques with respect to fexible job shop scheduling. *International Journal of Scientifc and Research Publications*. 2014. vol. 4. no. 3. pp. 1–7.
24. Amjad M.K. et al. Recent Research Trends in Genetic Algorithm Based Flexible Job Shop Scheduling Problems. *Mathematical Problems in Engineering*. 2018. vol. 2018. pp. 1–32.
25. Liu C., Coombes M., Li B., Chen W-H. Enhanced situation awareness for unmanned aerial vehicle operating in terminal areas with circuit flight rules. *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering*. 2016. vol. 230. no. 9. pp. 1683–1693.
26. Laporte G., Ropke S., Vidal T. Chapter 4: Heuristics for the Vehicle Routing Problem. Vehicle Routing: Problems, Methods, and Applications, Second Edition. 2014. pp. 87–116.
27. Braysy O., Gendreau M. Vehicle routing problem with time windows, part I: route construction and local search algorithms. *Transportation science*. 2005. vol. 39. no. 1. pp. 104–118.
28. Koc C., Bektas T., Jabali O., Laporte G. Thirty years of heterogeneous vehicle routing. *European Journal of Operational Research*. 2016. vol. 249. no. 1. pp. 1–21.
29. Pshikhopov V.K., Medvedev M.Y. [Group control of autonomous robots motion in uncertain environment via unstable modes]. *Trudy SPIIRAN – SPIIRAS Proceedings*. 2018. vol. 5. no. 60. pp. 39–63. (In Russ.).
30. Yong M. Solving vehicle routing problem with time windows with hybrid evolutionary algorithm. 2010 Second WRI Global Congress on Intelligent Systems (GCIS). 2010. vol. 1. pp. 335–339.
31. Affi M., Derbel H., Jarboui B. Variable neighborhood search algorithm for the green vehicle routing problem. *International Journal of Industrial Engineering Computations*. 2018. vol. 9. no. 2. pp. 195–204.

УДК 004.021 **DOI** 10.15622/sp.18.2.302-325

С.Г. Попов, В.С. Заборовский, Л.М. Курочкин, М.П. Шарагин, Л. Чжан

МЕТОД ДИНАМИЧЕСКОГО ВЫБОРА СПУТНИКОВОЙ НАВИГАЦИОННОЙ СИСТЕМЫ В АВТОНОМНОМ РЕЖИМЕ ПОЗИЦИОНИРОВАНИЯ

Попов С.Г., Заборовскй В.С., Курочкин Л.М., Шарагин М.П., Чжан Л. **Метод динамического выбора спутниковой навигационной системы в автономном режиме позиционирования.**

Аннотация. Перечень прикладных задач, требующих точного оперативного позиционирования, постоянно растет. К таким задачам относятся: задачи управления группами автономных мобильных роботов; геодезические задачи высокоточного позиционирования; задачи навигации и мониторинга в интеллектуальных транспортных системах. Источником данных для оперативного позиционирования в таких задачах являются спутниковые навигационные системы. Активно используются глобальные и локальные спутниковые навигационные системы: GPS, GLONASS, BeiDou, Galileo. Их характеризует разная полнота развертывания спутниковой группировки, что определяет точность оперативного позиционирования в конкретной географической точке, которая зависит от числа доступных для наблюдения спутников, а также характеристик приемника, особенностей ландшафта, погодных условий и возможности использования дифференциальных поправок.

Повсеместное использование дифференциальных поправок на данный момент невозможно ввиду того, что количество стабильно работающих опорных станций ограничено — Земля покрыта ими неравномерно; надежные сети передачи данных, необходимые для передачи дифференциальных поправок также развернуты не повсеместно; широкое применение нашли бюджетные версии одноканальных приемников навигационного сигнала, не позволяющие использовать дифференциальные поправки. В этом случае возникает задача оперативного выбора системы или комбинации систем спутникового позиционирования, предоставляющей наиболее точные навигационные данные. Приведено сравнение статического и динамического методов выбора системы или комбинации систем спутникового позиционирования, обеспечивающих наиболее точное определение собственных координат объекта при использовании одноканального приемника навигационных сигналов в автономном режиме (без использования сторонних поправок). Выбор осуществляется на основе статистического анализа данных, получаемых от систем спутникового позиционирования. При проведении анализа выполнялось сравнение результатов, сформированных при постобработке данных, полученных от спутниковых навигационных систем и уточненных с применением дифференциальных поправок навигационных данных.

Ключевые слова: навигация автономных мобильных объектов; статистический анализ навигационных данных; методы выбора системы спутникового позиционирования.

1. Введение. Основным средством определения координат наземного объекта являются данные, получаемые от спутниковых систем позиционирования. На разных стадиях развертывания, эксплуатации и использования находится глобальные и

региональные национальные системы спутникового позиционирования, такие как: GPS, GLONASS, GALILEO, BeiDou. Постоянное наращивание группировок спутников и их модернизация обеспечивают повышение точности позиционирования. Однако из-за влияния различных факторов позиционирование с использованием различных систем обеспечивается с разной точностью, особенно при использовании бюджетных приемников мобильными автономными объектами. Ключевыми особенностями бюджетных приемников являются сравнительно невысокая стоимость и компактное исполнение, что достаточно значимо при массовом производстве, например, мобильных роботов. К важным особенностям бюджетных приемников следует отнести: использование одного диапазона для получения данных от спутников и возможность одновременного использования нескольких систем спутниковой навигации. Использование одного диапазона получения данных от спутников обеспечивает меньшую точность позиционирования, по сравнению с одновременным использованием нескольких диапазонов приема навигационных данных.

На точность координат, получаемых от спутниковых навигационных систем, влияют: взаимное расположение спутников, от которых может быть получен сигнал; погодные условия; особенности рельефа в конкретной точке определения координат, ионосферная и тропосферная рефракции [1-3].

При оценке точности навигационных данных оценивается ряд факторов [4]:

– геометрический фактор снижения точности (GDOP);
– горизонтальный фактор снижения точности (HDOP);
– фактор снижения точности определения положения (PDOP);
– относительный фактор снижения точности (RDOP);
– временной фактор снижения точности (TDOP);
– вертикальный фактор снижения точности (VDOP).

Анализ точности определения координат и выбор наиболее точной системы спутникового позиционирования или комбинации таких систем вызывает интерес как у исследователей, так и у специалистов, решающих прикладные задачи. Точность определения координат в автономном режиме (т.е. без сторонних поправок) может быть повышена за счет применения фильтрации данных, использования поправок, например RTK [5], средствами библиотеки программ с открытым исходным кодом RTKLib [6] или проприетарных решений [7], использования комбинации спутниковых

навигационных систем, разработки и использования методов повышения точности навигационного обеспечения.

Подготовку и распространение дифференциальных поправок, используемых для передачи на абонентские приемники для корректировки принимаемых ими навигационных сигналов, обеспечивают несколько систем: американская WAAS (Wide Area Augmentation System) для GPS, Европейская — EGNOS (European Geostationary Navigation Overlay Service) для Galileo, Японская — MSAS (Multi-functional Satellite Augmentation System), СДКМ (система дифференциальной коррекции и мониторинга) для GLONASS. Каждая из систем имеет собственный набор спутников, оборудованных передатчиками навигационных сигналов, а приемники — специализированным программным обеспечением для проведения обработки и уточнения навигационных данных.

Для снижения влияния случайной погрешности на точность определения координат спутниковых систем при решении навигационных задач используются методы фильтрации, например, фильтр Калмана и расширенный фильтр Калмана [8-11]. Предложенный в работе [8] метод, основанный на применении фильтра Калмана, позволяет сократить погрешность определения пространственных координат заданной точки более чем в 10 раз. Для борьбы с узкополосными помехами успешно разрабатываются адаптивные методы фильтрации сигнала, основанные на оценке интерференционной частоты [12].

Для решения задач управления воздушными судами, в частности захода на посадку, могут использоваться заранее сформированные поля точности спутниковой системы навигации, подготовленные с учетом повторяемости наблюдаемой орбитальной группировки. Работа [13] посвящена описанию процесса построения полей точности воздушного пространства на основание значений горизонтального (HDOP, Horizontal Dilution of Precision) и вертикального (VDOP Vertical Dilution of Precision) геометрического фактора в выбранных точках воздушного пространства. Результаты, представленные в работе, подтверждают целесообразность использования описанных методов для оценки условий навигационного сеанса, а также построения полей точности GPS в заданной зоне воздушного пространства.

Исследование методов высокоточного навигационного обеспечения, учитывающего погрешности, которые вызваны ошибками эфемеридно-временной информации, влияния релятивистских, гравитационных и приливных эффектов, ошибки

многолучевости, приведены в работе [14]. Анализ приведенного метода показывает, что его точность сравнима с точностью позиционирования, выполненного с учетом RTK поправок.

Анализ результатов совместного использования навигационных систем GPS, GLONASS, BeiDou, Galileo также привлекает внимание исследователей [15]. Представленный в работе [16] анализ показывает, что при совместном использовании BeiDou и Galileo точность позиционирования составляет менее 0,1 м, при совместном использовании GLONAS и GPS — менее 0,05 м. Совместное использование систем BeiDou, Galileo, GLONASS и GPS сокращает время конвергенции почти на 70%, тогда как точность позиционирования возрастает примерно на 25%. В современной литературе представлены и другие методы повышения точности позиционирования, основанные на одновременном использовании нескольких спутниковых систем позиционирования в различных режимах работы. Большинство описанных методов требуют значительных временных затрат либо использования многодиапазонного приемника навигационных сигналов [17].

Анализ зависимостей доступности различных спутниковых систем позиционирования и точности предоставляемых ими данных, от периода наблюдения, состояния конкретной группировки или их комбинаций, позволяет сформировать рекомендации для выбора комбинации систем спутниковой навигации, предоставляющей наиболее точные данные в конкретный момент в заданной географической точке. Результаты такого рода исследований представлены в работе [18]. В некоторых исследования проводится анализ траекторий спутников и предугадывание потери связи со спутником [19].

Сбором, анализом и распространением параметров функционирования спутниковых группировок и данных, полученных от глобальных систем позиционирования, занимается глобальная служба навигационных спутниковых систем [20]. К таким данным относятся оценки задержки передачи сигнала, параметры вращения Земли и спутниковые эфемериды. Указанные параметры и данные предоставляют более двухсот агентств, исследовательских институтов, университетов. Перечисленные параметры и данные применяются при постобработке навигационных данных. Постобработка позволяет повысить точность навигационных данных, используемых как при решении фундаментальных, так и прикладных задач.

Целью данной работы является сравнение методов оперативного выбора системы или комбинации систем спутникового позиционирования, предоставляющей наиболее точные

навигационные данные. Указанный выбор производится в автономном режиме с использованием методов статистического анализа.

2. Технические средства получения данных спутниковых навигационных систем. На рынке представлено множество приемников сигналов спутниковых навигационных систем от модулей, предназначенных для реализации встроенных применений в устройствах, требующих определения собственного местоположения, до геодезических станций, обеспечивающих точность определения местоположения до сантиметров. Большинство мобильных устройств оснащаются приемниками сигналов спутниковых навигационных систем, которые способны работать с несколькими системами спутниковой навигации. Все приемники отличаются по стоимости от ультра-бюджетных до дорогостоящих.

На рисунке 1 представлена классификация современных приемников сигналов систем спутникового позиционирования.

1) Используемые рабочие частоты приемника. На данный момент наиболее часто используемыми являются частоты L1, L2, L5 либо их комбинации. Большинство бюджетных приемников поддерживают только частоты L1, L2.

2) Количество каналов приемника: существуют одноканальные и многоканальные приемники. Количество каналов позволяет параллельно получать данные от спутников разных навигационных систем, что увеличивает скорость получения данных. Наиболее популярные реализации многоканальных приемников поддерживают от 12 до 200 каналов.

3) Частота обновления навигационных данных, формируемых приемником. В большинстве реализаций бюджетных приемников частота обновления навигационной информации предустановлена и зафиксирована. В специализированных реализациях приемников обновления навигационной информации предусмотрена возможность выбора частоты обновления навигационной информации. Как правило, частоты обновления менее 1Гц не используются.

4) Формат передачи навигационных данных — возможность приемника передавать навигационные данные потребителю в различных форматах. Для бюджетных приемников наиболее часто используется протокол NMEA (National Marine Electronics Association).

5) Время старта перед выдачей фиксированных навигационных данных — наименьший временной интервал, необходимый приемнику для подготовки решения по данным, полученным от спутниковой навигационной системы.

6) Форм-фактор. Встраиваемые, портативные, kit-наборы либо промышленные.

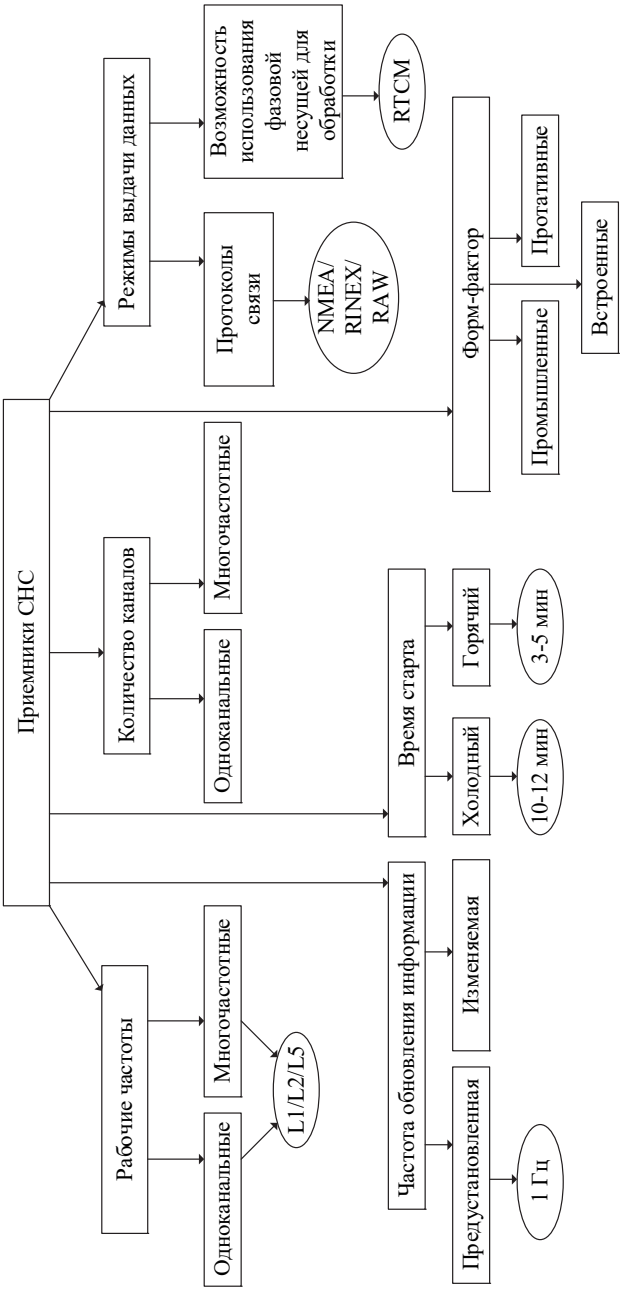


Рис. 1. Классификация приемников сигналов спутниковых навигационных систем по свойствам и потребительским качествам

3. Методы выбора системы спутникового позиционирования, предоставляющей наиболее точные навигационные данные. Необходимость оперативного выбора системы спутникового позиционирования или комбинации таких систем, предоставляющих наиболее точные навигационные данные, возникает в задачах автономного группового управления наземными мобильными робототехническими системами; в задачах навигационного/информационного обеспечения интеллектуальных транспортных систем; в прочих задачах, требующих точного оперативного определения собственных координат при использовании бюджетных многоканальных приемников навигационных сигналов.

Мобильный объект/робот, не обладающий возможностью оперативного получения дифференциальных поправок навигационной информации, способен самостоятельно выбрать наиболее точную систему спутникового позиционирования (систему, предоставляющую наиболее точные навигационные данные), используя статический или динамический метод.

Статический метод описан в работе [21]. Указанный метод основан на статистическом анализе наборов собственных координат неподвижного объекта/робота, полученных от различных систем спутниковой навигации, или их комбинаций за равные интервалы времени $\boldsymbol{T}$. Для каждого набора собственных координат вычисляется положение «эталонного» центра — точки с координатами, равными среднему значению соответствующих координат в наборе. Далее вычисляются среднее евклидово расстояние от всех точек набора собственных координат до вычисленного «эталонного» центра и дисперсия указанных расстояний до «эталонного» центра.

Рекомендуемой для использования системой спутникового позиционирования или комбинацией таких систем принимается система или комбинация, показавшая наименьшие среднее расстояние до центра и дисперсию.

Метод, предлагаемый в данной работе, отличается от метода, приведенного в работе [21], использованием динамических оценок расстояния до вычисленного «эталонного» центра и дисперсии. Указанные оценки формируются и уточняются по мере получения новых координат. Для формирования оценок используются наборы собственных координат неподвижного объекта/робота, которые были предоставлены различными системами спутниковой навигации, или их комбинациями. По мере получения новых значений собственных координат указанный набор смещается в последовательной выборке на s значений, для оценки используются значения собственных координат

от $k+s$ до $k+m+s$. Применяется следующая терминология: m — размер «окна» — количество собственных координат, используемых для формирования оценок, s — «шаг окна» — количество собственных координат, последовательно полученных во времени, на которое смещается «окно». Блок-схема динамического метода выбора системы спутникового позиционирования приведена на рисунке 2.

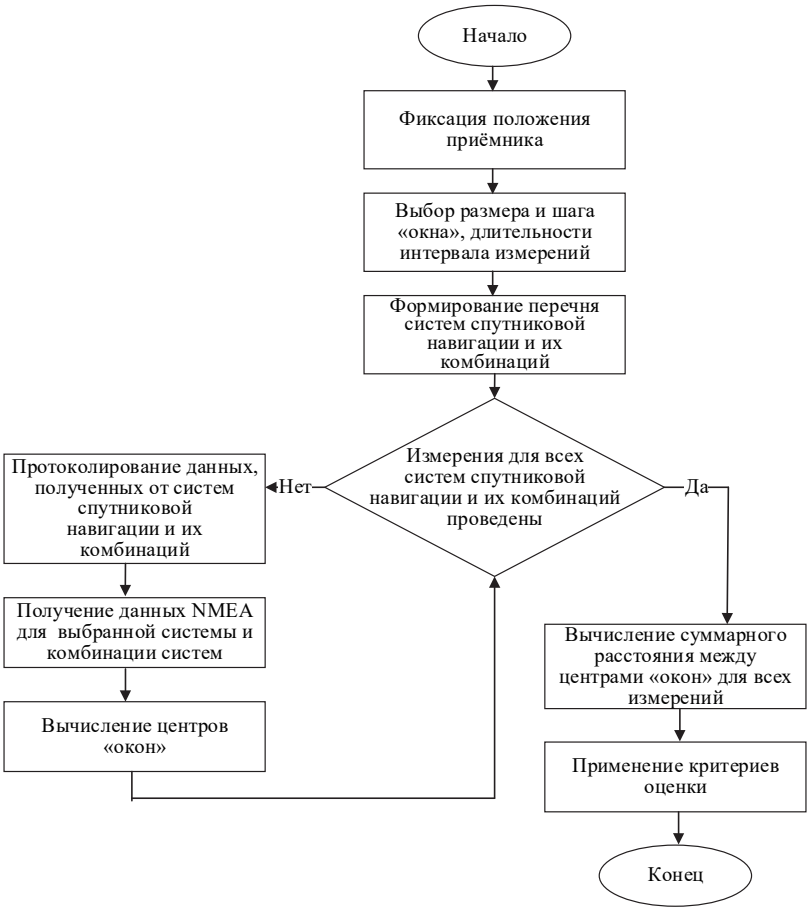


Рис. 2. Блок-схема динамического метода выбора спутниковой навигационной системы или комбинации систем

В рамках проводимого апостериорного анализа навигационных данных вычислялись оценки: суммарного расстояния между центрами «окон» в рамках одного эксперимента и суммарного расстояния между

центрами «окон» и уточненным центром полной выборки, вычисленным с использованием дифференциальных поправок для навигационных данных GPS. Центры «окон» вычислялись с помощью алгоритмов k-means, c-means с одним центром. Центр, полученный по средствам применения k-means с одним центром, фактически совпадает со средним значением, вычисленным по выборке. Центр, вычисленный по средствам применения c-means с одним центром, используется для сравнительной оценки среднего значения по выборке.

Уточненный центр полной выборки использовался в качестве точного значения координат приемника для оценки погрешности определения координат при обработке данных, полученных от систем спутниковой навигации и их комбинаций без использования дифференциальных поправок. Суммарное расстояние между центрами «окон», полученное в рамках одного эксперимента, является альтернативной оценкой точности определения собственных координат при обработке данных от систем спутниковой навигации и их комбинаций без использования дифференциальных поправок. Графическая иллюстрация оценки положения центров «окон», уточненного центра полной выборки представлена на рисунке 3.

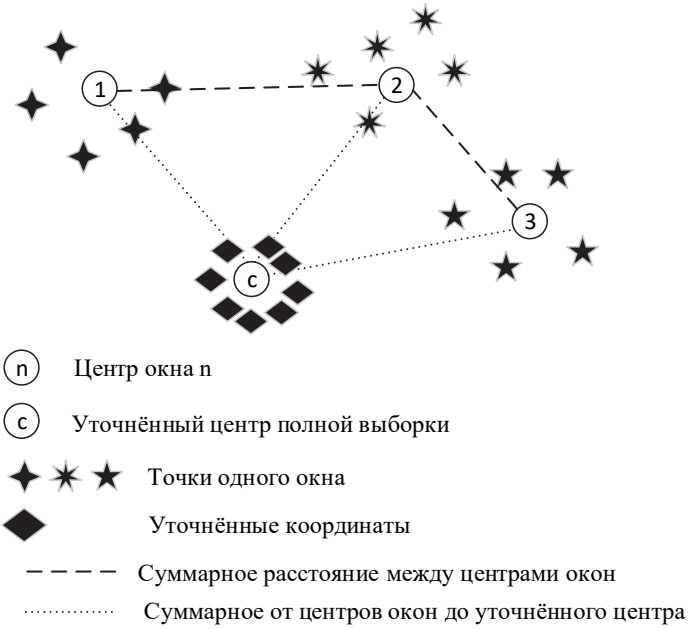


Рис. 3. Положения центров «окон», уточненного центра полной выборки

В рамках апостериорного анализа данных, полученных от систем спутниковой навигации и их комбинаций, проводилась оценка точности определения собственных координат для полных выборок навигационных данных и для выборок, прошедших фильтрацию выбросов. При фильтрации данных из выборок удалялись значения, среднеквадратичное отклонение которых превышало *n* сигма, где *n* принимало значения: 3, 2.75, 2.5, ... 1.25, 1. После каждой фильтрации вычислялись центры выборок. При оценке точности определения собственных координат проводилось сравнение расстояний от центров «фильтрованных» выборок, вычисленных методами k-means, c-means до центра, выборки GPS, уточнённой с использованием дифференциальных поправок.

4. Оборудование и условия проведения экспериментов. При проведении серии экспериментов использовался приемник EVK-7 Multi-GNSS evaluation kit. Под серией экспериментов подразумеваются измерения, проводимые в одной географической точке. Антенна приемника располагалась неподвижно на протяжении всех экспериментов серии. Осуществлялся выбор системы спутникового позиционирования или парного набора таких систем, затем осуществлялось протоколирование навигационных данных на протяжении интервала — 15 минут. Частота получения данных от приемника — 1 Гц. В экспериментах участвовали системы спутникового позиционирования GPS, Glonass, BeiDou, Galilleo, QZSS, SBAS и их парные комбинации. То есть для каждой системы или комбинации систем в рамках каждого эксперимента было получено порядка 900 координат, всего в каждой точке проводилось 20 экспериментов (6 — для каждой системы, 14 — для каждой парной комбинации систем). Эксперименты проводились в Санкт-Петербурге (Россия) и в Мадриде (Испания). В Санкт-Петербурге и Мадриде были выбраны по две точки, в каждой из которых проводилась серия измерений. Серии измерений проводились в разные дни и время суток. Анализ данных проводился апостериорно. Полученные данные обрабатывались согласно методам, приведенным выше. Точки проведения эксперимента были удалены друг от друга с целью оценки влияния неоднородности покрытия Земли сигналом спутниковых навигационных систем.

Для статического метода проводилась оценка отклонений среднего расстояния от центра обрабатываемой выборки до вычисленного «эталонного» центра выборки, сформированного по полной пятнадцатиминутной выборке. В обрабатываемой выборке использовались наборы точек от первого измерения до k-го, где k

принимало значения 30 с, 60 с, 90 с, …, 900 с. То есть проводилось сравнение «эталонного» центра, полученного по пятнадцатиминутной выборке, и значений центров, полученных по обрабатываемым выборкам указанных размеров. Также проводилось сравнение дисперсий значений, присутствующих в обрабатываемых выборках. Выборку размером 30 с можно считать репрезентативной, так как среднее значение по выборке отличается менее чем на 1 % от среднего значения по выборке размером 60 с.

Для динамического метода размер «окна» принимал значения 30 с, 60 с, 90 с, …, 300 с, шаг «окна» был равен 60 с.

5. Анализ экспериментальных данных. Результаты исследований статического метода, приведенные в [21], показывают, что значения среднего расстояния до «эталонного» центра выборки для наборов измерений, полученных при использовании различных систем спутниковой навигации и их парных комбинаций, может значительно отличаться в рамках проведения одной серии эксперимента (серии пятнадцатиминутных измерений). Например, для экспериментов, проведенных в Мадриде, значение среднего отклонения центра выборки от «эталонного» центра, вычисленного в результате обработки пятнадцатиминутной выборки, отличается более чем в 7.5 раз для данных, полученных от навигационных систем Gallileo и GPS (рисунок 4). По оси абсцисс (размер выборки) приведены значения объемов выборок для которых вычислялось отклонение.

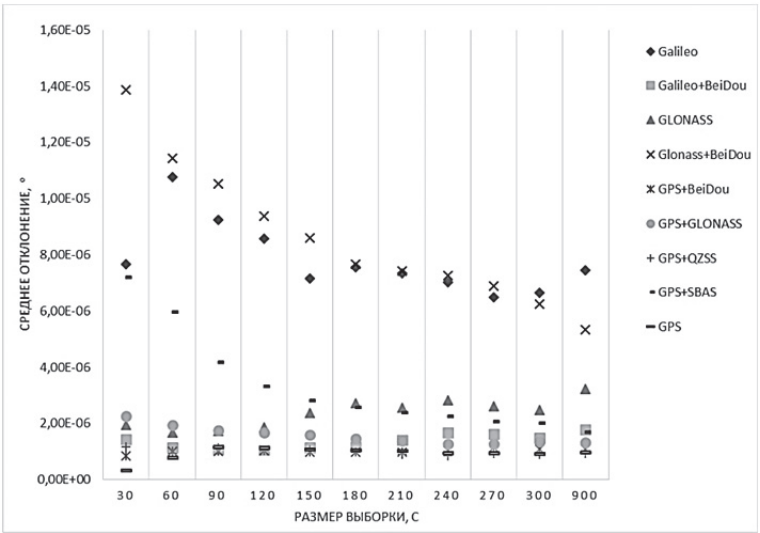


Рис. 4. Среднее отклонение от «эталонного» центра (Мадрид)

Дисперсия соответствующих измерений (Gallileo и GPS) отличается более чем в 97 раз (рисунок 5). Наименьшую дисперсию и, соответственно, расстояние до эталонного центра обеспечивает GPS.

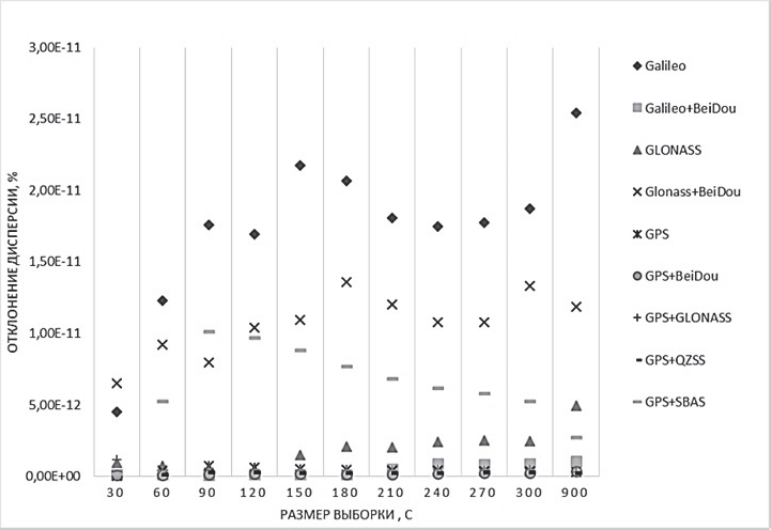


Рис. 5. Зависимость дисперсии среднего расстояния до «эталонного» центра (Мадрид) от размера выборки

Зависимость значений дисперсии навигационных данных для систем спутниковой навигации, показавших наименьшие значения среднего расстояния до «эталонного» центра, изображены на рисунке 6.

Представленная диаграмма показывает, что для выборок, навигационных данных, полученных за временные интервалы 150 и более секунд, значение дисперсии отличается от «эталонного» не более чем на 0.01 %.

Диаграмма, представленная на рисунке 7, показывает, что для выборки, соответствующей временному интервалу 150 с и более, значение среднего расстояния до «эталонного» центра, полученного для конкретной выборки, отличается не более чем на 0.001 % от значения, полученного для полной выборки. Вычисленные значения дисперсий для указанных выборок принимают следующие значения (рисунок 8).

Представленные зависимости показывают, что комбинации спутниковых систем GPS + BaiDou и GPS + QZSS обеспечивают наименьшие значения дисперсий навигационных данных. Отношения

дисперсии к математическому ожиданию (оценка расстояния до «эталонного» центра) составляют величину порядка $5*10^{-7}$, что подтверждает, что процессы сходятся.

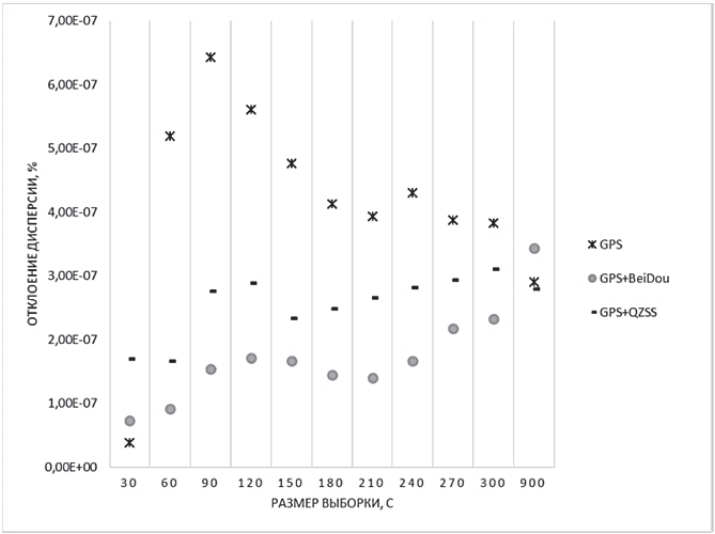


Рис. 6. Зависимость отклонения значений дисперсии (%) от результирующего значения по полной выборке

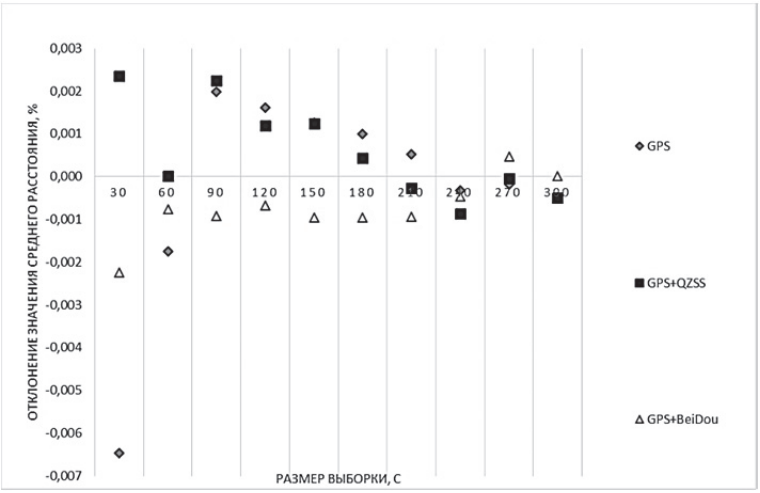


Рис. 7. Зависимость отклонения значения среднего расстояния до «эталонного» центра от размера выборки

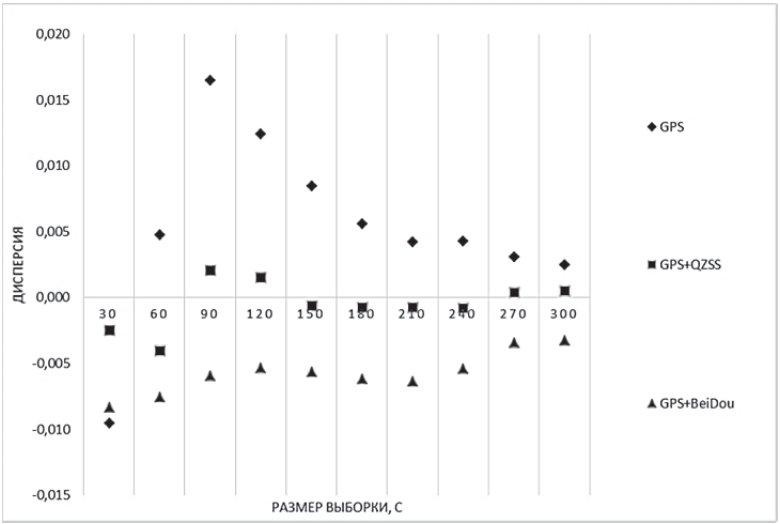


Рис. 8. Зависимость дисперсии выборки от размера выборки. GPS, GPS+BeiDou, GPS+QZSS

По результатам представленных экспериментов можно сделать вывод, что для оценки точности навигационных данных, получаемых от системы спутниковой навигации или парной комбинации таких систем, необходима обработка выборок навигационных данных, полученных за время порядка трех минут.

При исследовании динамического метода выбора навигационной системы для каждой серии экспериментов проводилось сравнение вычисленных центров «окон» с уточненным значением положения приемника. Уточненное значение положения приемника вычислялось с использованием дифференциальных поправок для данных, полученных в данной серии эксперимента от навигационной системы GPS.

Значения отклонений от уточненного значения положения приемника использовались для оценки точности метода и выбора его параметров. На рисунке 9 представлена зависимость изменения суммарного расстояния между центрами «окон» для первой серии эксперимента, проводимого в Санкт-Петербурге. В данном эксперименте размер «окна» принимал значения 60 с, 90 с, 120 с … 300 с, «окно» двигалось с шагом 60 с. Для данной серии экспериментов наименьшее значение суммарного расстояния между центрами «окон» обеспечивает использование комбинации систем спутниковой навигации GPS и SBAS.

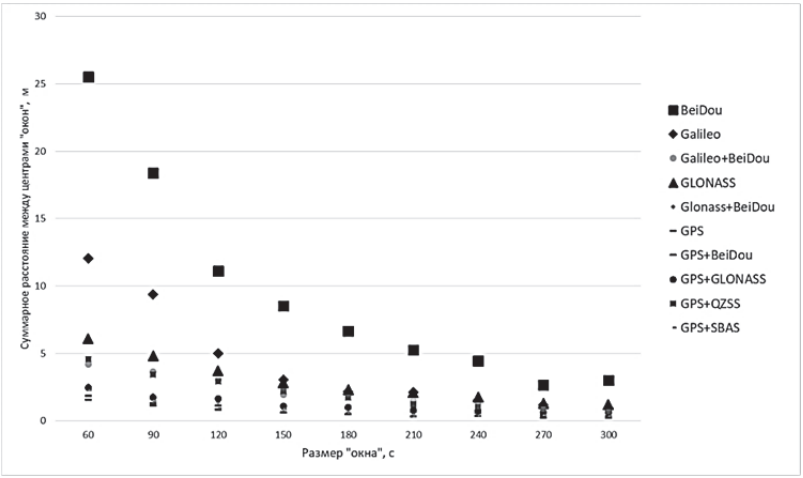


Рис. 9. Суммарное расстояние между центрами окон (Санкт-Петербург, первый эксперимент)

Эта же комбинация систем спутниковой навигации обеспечивает наименьшее расстояние от центра «окна» до уточненного центра, вычисленного с использованием дифференциальных поправок для навигационных данных, полученных от GPS (рисунок 10).

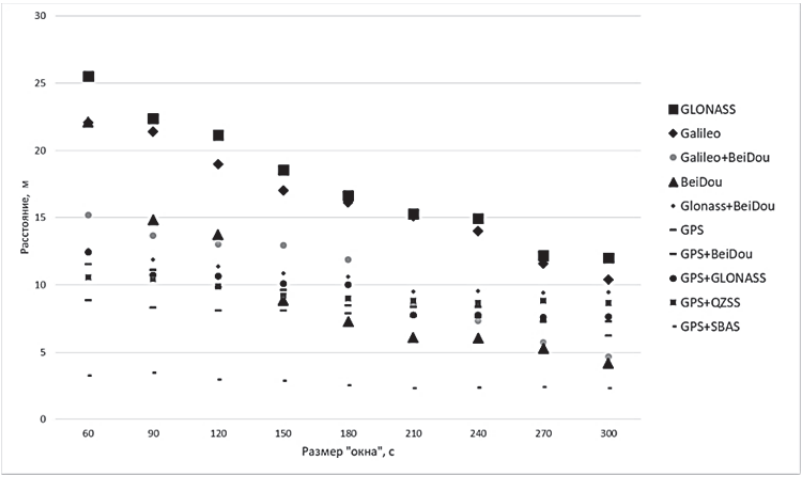


Рис. 10. Расстояние от центра «окна» до уточненного по GPS центра (Санкт-Петербург, первый эксперимент)

В таблице 1 приведены значения отношений суммарных расстояний между центрами окон для наименьшего суммарного расстояния, обеспечиваемого в данной серии экспериментов комбинацией спутниковых систем GPS и SBAS. В таблице приведены значения для нескольких комбинаций систем спутниковой навигации, вычисленные в зависимости от размера «окна», используемого в серии экспериментов, а также усредненные по данным конкретного эксперимента дисперсии. Данные, приведенные в таблице, позволяют сделать вывод, что при использовании динамического метода выбора спутниковой навигационной системы использование «окна» размером 90 с. достаточно для формирования оценки точности навигационных данных, получаемых от спутниковой навигационной системы или комбинации систем. Среднее значение дисперсии подтверждает обоснованность выбора.

Таблица 1. Оценка точности по размеру окна (Санкт-Петербург, первый эксперимент)

		Отношение суммарного расстояния между центрами окон		
Размер «окна» сек.	Количество шагов	GPS+SBAS и BeiDou	GPS+SBAS и Galileo	GPS+SBAS и GPS+QZSS
60	8	11,18722	5,29396	2,00966
90	7	17,25133	8,81791	3,23396
120	7	13,05250	5,84847	3,41301
150	6	9,79725	3,48373	2,52817
180	6	12,37874	3,99748	3,20094
210	5	15,12715	6,08051	3,52487
240	5	12,26723	4,22035	2,94589
270	4	6,38588	2,86823	1,96481
300	4	9,52387	2,88521	2,16126
Средняя (по экспериментам) дисперсия навигационных данных		2,07669E-12	2,30942E-10	3,97981E-12

Серия экспериментов, проведенных в Мадриде, подтверждает утверждение о том, что при использовании динамического метода для оценки точности навигационных данных достаточно использовать «окно» размером 90 с. В первой серии экспериментов, проводимых в Мадриде, наибольшая точность навигационных данных была достигнута при использовании комбинации навигационных систем GPS и GLONASS. В рамках проводимой серии экспериментов указанная комбинация навигационных систем позволила сократить отклонение

центра «окна» от уточненного центра в 6.2 раза по сравнению с результатами обработки данных, полученных от комбинаций систем GLONASS и BeiDou. Зависимости суммарного расстояния между центрами окон от размера «окна» представлена на рисунке 11. Зависимости расстояний от центров «окон» до уточненного центра изображены на рисунке 12. Значения дисперсий по полной выборке навигационных данных отличаются на 2 порядка. Для комбинации GPS и Glonass значение составляет $5.4*10^{-12}$, для комбинации Glonass и BeiDou, показавшей наименьшую точность, — $3.8 * 10^{-10}$.

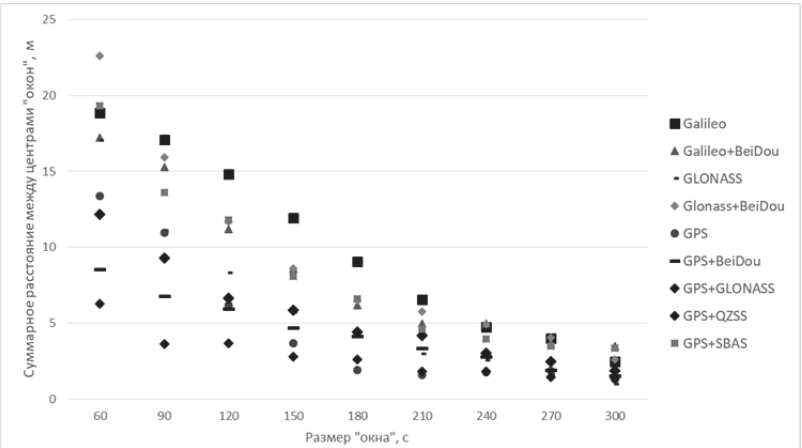


Рис. 11. Суммарное расстояние между центрами окон (Мадрид, первый эксперимент)

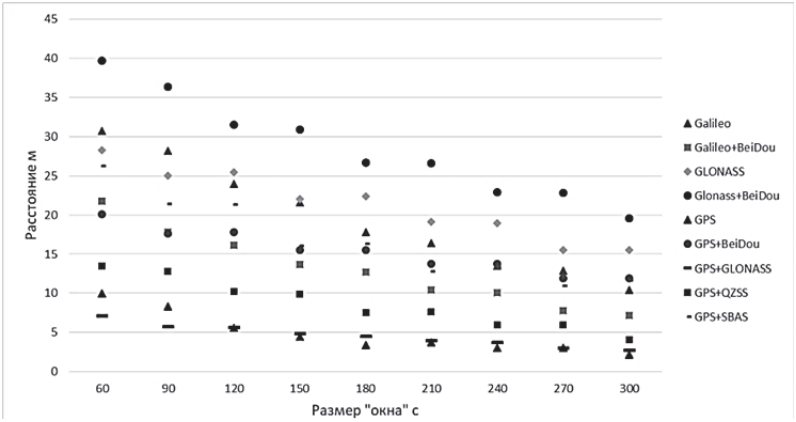


Рис. 12. Расстояние от центра «окна» до уточнённого по GPS центра (Мадрид, первый эксперимент)

В ходе анализа выборок данных, полученных после фильтрации, были вычислены расстояния от центров, вычисленных по методам k-means, c-means выборок, прошедших фильтрацию до центра уточненной с использованием дифференциальных поправок выборки GPS, прошедшей аналогичную фильтрацию. Результаты анализа первого эксперимента, проводимого в Мадриде, представлены в таблице 2.

Таблица 2. Расстояния между центрами выборок после фильтрации (Мадрид, второй эксперимент)

	GPS+SBAS		GPS+ GLONASS		GPS+QZSS		GPS+BeiDou	
sigm	k-means м	c-means м	k-means м	c-means м	k-means м	c-means м	k-means м	c-means м
3	0,649	0,660	0,347	0,375	0,845	0,815	1,258	1,275
2,75	0,653	0,669	0,358	0,378	0,839	0,815	1,275	1,274
2,5	0,672	0,674	0,365	0,371	0,831	0,777	1,287	1,303
2,25	0,732	0,728	0,366	0,368	0,821	0,786	1,295	1,299
2	0,740	0,752	0,359	0,355	0,800	0,779	1,314	1,311
1,75	0,771	0,792	0,373	0,364	0,805	0,761	1,312	1,307
1,5	0,797	0,812	0,388	0,378	0,814	0,770	1,318	1,322
1,25	0,817	0,826	0,393	0,382	0,816	0,772	1,320	1,316
1	0,830	0,841	0,393	0,385	0,820	0,797	1,324	1,315

Результаты, представленные в таблице 2, показывают, что фильтрация не дает ощутимого увеличения точности. При этом точность определения координат является достаточно высокой, а именно: для комбинации GPS+GLONASS расстояние между вычисленным центром и уточненным с использованием данных системы IGS (International GNSS Service) GPS центром составляет величину порядка 34 см. Для экспериментов, проведенных в Санкт-Петербурге, расстояние между центрами выборок, которые были получены для комбинаций систем спутниковой навигации, выбранными динамическим методом, составили величины порядка 45 см и 55 см для двух экспериментов. Для первого эксперимента, проводимого в Мадриде, — величину порядка 8 см. Значения расстояний до уточненного центра, вычисленные при обработке выборок навигационных данных, полученных от остальных систем

спутниковой навигации в рамках проводимых серий экспериментов, превосходили данные значения от 2 до 12 раз, что подтверждает корректность работы динамического метода выбора спутниковой навигационной системы в автономном режиме позиционирования.

6. Заключение. В работе предложен метод выбора системы спутникового позиционирования или комбинаций таких систем, обеспечивающий наибольшую точность определения собственных координат при использовании одноканальных приемников и без использования поправок навигационных данных.

По результатам проведенных исследований можно сделать вывод о том, что представленный в данной работе динамический метод позволяет сократить время выбора спутниковой навигационной системы вдвое по сравнению со статическим методом.

Точность навигационных данных, получаемых от системы спутниковой навигации или комбинации таких систем, выбранных с использованием предложенного метода, позволяет достигнуть точности определения собственных координат, сравнимой с точностью, получаемой при использовании постобработки с использованием данных системы IGS. Проведенные эксперименты показали возможность достижения точности определения собственных координат от единиц дециметров до единиц сантиметров.

Представленное решение может использоваться для навигационного обеспечения в задачах управления автономными наземными объектами. Использование представленного решения позволяет сократить время выбора наиболее точной системы спутникового позиционирования или парных комбинаций систем в условиях отсутствия дифференциальных поправок и полей точности систем спутниковой навигации, сформированных для данной местности, и сведений о начальном географическом положении объекта; ограниченных вычислительных мощностей автономного объекта; использования одноканального приёмника навигационных сигналов.

Литература

1. *Schüler T.* On ground-based GPS tropospheric delay estimation // Doctor's Thesis, Studiengang Geodsie und Geoinformation. 2001. vol. 73. 364 p.
2. *Хуторова О.Г. и др.* Пассивное зондирование структуры коэффициента преломления радиоволн в тропосфере сетью приёмников спутниковых навигационных систем в г. Казани // Известия высших учебных заведений. Радиофизика. 2011. Т. 54. № 1. С. 1–8.
3. *Попов С.Г., Попов М.В.* Исследование методов повышения точности навигационных // Материалы научного форума с международным участием «Неделя науки СПбПУ». 2016. С. 8–10.

4. *Adrados C. et al.* Global Positioning System (GPS) location accuracy improvement due to Selective Availability removal // Comptes Rendus Biologies. 2012. vol. 325. no. 2. pp. 165–170.
5. *Siejka Z.* Validation of the accuracy and convergence time of real time kinematic results using a single galileo navigation system // Sensors. 2018. vol. 18(8). pp. 2412.
6. *Takács B., Siki Z., Markovits-Somogyi R.* Extension of RTKLIB for the Calculation and Validation of Protection Levels // International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences. 2017. vol. 42. 6 p.
7. *Suldi A.M., Samad A.M.A., Hashim H.* Feasible study on establishment of a GPS reference station // 2012 IEEE 8th International Colloquium on Signal Processing and its Applications. 2012. pp. 362–366.
8. *Макаренко Г.К., Алешечкин А.М.* Исследование алгоритма фильтрации при определении координат объекта по сигналам спутниковых радионавигационных систем // Доклады Томского государственного университета систем управления и радиоэлектроники. 2012. № 2-2(26). С. 15–18.
9. *Гаврилов А.В.* Использование фильтра Калмана для решения задач уточнения координат БПЛА // Современные проблемы науки и образования. 2015. № 1-1. С. 1784–1784.
10. *Wans S.C. et al.* Study on an intelligent fault-tolerant technique for multiple satellite configured navigation under highly dynamic conditions // Science China Information Sciences. 2011. vol. 54. pp. 529–541.
11. *Hu J. et al.* An extended Kalman filter and back propagation neural network algorithm positioning method based on anti-lock brake sensor and global navigation satellite system information // Sensors. vol. 18. no. 19. pp. 2753.
12. *Ferrara N.G. et al.* A new implementation of narrowband interference detection, characterization, and mitigation technique for a software-defined multi-GNSS receiver // GPS Solutions. 2018. vol. 22(4). pp. 106.
13. *Скрыпник О.Н., Нечаев Е.Е., Арефьев Р.О.* Построение и анализ полей точности GPS на основе программно-аппаратных средств NI GPS Simulation Toolkit // Научный вестник Московского государственного технического университета гражданской авиации. 2014. № 209. С. 5–12.
14. *Дворкин В.В., Карутин С.Н.* Высокоточные навигационные определения по сигналам ГНСС // Сибирский журнал науки и технологий. 2013. № 6(52). С. 70–76.
15. *Guo J. et al.* Multi-GNSS precise point positioning for precision agriculture // Precision Agriculture. 2018. vol. 19(5). pp. 895–911.
16. *Li X. et al.* Accuracy and reliability of multi-GNSS real-time precise positioning: GPS, GLONASS, BeiDou, and Galileo // Journal of Geodesy. 2015. vol. 89. no. 6. pp. 607–635.
17. *Серапинас Б.Б.* Глобальные системы позиционирования // М.: ИФК «Каталог». 2002. 105 с.
18. *Li X. et al.* Precise Positioning with Current Multi-Constellation Global Navigation Satellite Systems: GPS, GLONASS, Galileo and BeiDou // Scientific Reports. 2015. vol. 5. pp. 8328.
19. *Xiaolei Y., Yongrong S., Jianye L., Jianfeng M.* Fast Algorithm of Selecting Satellites for Multiple Satellite Integrated Navigation Engineering // 2009 WRI World Congress on Computer Science and Information Engineering. 2009. vol. 5. pp. 121–125.
20. The International GNSS Service. URL: http://www.igs.org/ (дата обращения: 14.09.18).
21. *Курочкин Л.М., Курочкин М.А., Попов С.Г., Попов М.В.* Результаты экспериментальных исследований точности позиционирования при использовании различных систем спутниковой навигации // Научно-технические ведомости СПбГПУ. Информатика. Телекоммуникации. Управление. 2018. Т. 10. № 4. С. 79–88.

Попов Сергей Геннадьевич — канд. техн. наук, доцент, доцент, кафедра телематики (при ЦНИИРТК) института прикладной математики и механики, Санкт-Петербургский политехнический университет Петра Великого (СПбПУ). Область научных интересов: СУБД, робототехника, спутниковые навигационные системы. Число научных публикаций — 70. popovserge@spbstu.ru; 29, Политехническая, 195251, Санкт-Петербург, Российская Федерация; р.т.: +79219613493.

Заборовский Владимир Сергеевич — д-р техн. наук, профессор, профессор, кафедра телематики (при ЦНИИРТК) института прикладной математики и механики, Санкт-Петербургский политехнический университет Петра Великого (СПбПУ). Область научных интересов: киберфизика, квантовые вычисления. Число научных публикаций — 135. vlad@neva.ru; 29, Политехническая, 195251, Санкт-Петербург, Российская Федерация; р.т.: +7(921)9398954.

Курочкин Леонид Михайлович — канд. техн. наук, доцент, доцент, кафедра телематики (при ЦНИИРТК) института прикладной математики и механики, Санкт-Петербургский политехнический университет Петра Великого (СПбПУ). Область научных интересов: робототехника, интеллектуальные транспортные сети. Число научных публикаций — 30. kurochkinl@spbstu.ru; 29, Политехническая, 195251, Санкт-Петербург, Российская Федерация; р.т.: +7(921)3465767.

Шарагин Максим Павлович — аспирант, кафедра телематики (при ЦНИИРТК) института прикладной математики и механики, Санкт-Петербургский политехнический университет Петра Великого (СПбПУ). Область научных интересов: машинное обучение, робототехника. Число научных публикаций — 3. msharagin@gmail.com; 29, Политехническая, 195251, Санкт-Петербург, Российская Федерация; р.т.: +7(812)5526521.

Чжан Лэй — канд. техн. наук, профессор, Школа компьютерных наук и программного обеспечения, Восточно-китайский педагогический университет (ЭКНУ). Область научных интересов: спутниковые навигационные системы, робототехника. Число научных публикаций — 18. lzhang@ce.ecnu.edu.cn; 500, дор. Дунчуань, окр. Миньхан, 200241, Шанхай, Китай; р.т.: +86-18818223050.

Поддержка исследований. Госзадание, базовая часть: 2.9198.2017/8.9 Устойчивое управление гетерогенной группировкой киберфизических объектов в условиях пространственно-временной неопределённости.

DOI 10.15622/sp.18.2.302-325

S.G. POPOV, V.S. ZABOROVSKY, L.M. KUROCHKIN, M.P. SHARAGIN, L. ZHANG

METHOD OF DYNAMIC SELECTION OF SATELLITE NAVIGATION SYSTEM IN THE AUTONOMOUS MODE OF POSITIONING

Popov S.G., Zaborovsky V.S., Kurochkin L.M., Sharagin M.P., Zhang L. **Method of Dynamic Selection of Satellite Navigation System in the Autonomous Mode of Positioning.**

Abstract. Today, the list of applications that require accurate operational positioning is constantly growing. These tasks include: tasks of managing groups of Autonomous mobile robots, geodetic tasks of high-precision positioning, navigation and monitoring tasks in intelligent transport systems. Satellite navigation systems are a data source for operational positioning in such tasks. Today, global and local satellite navigation systems are actively used: GPS, GLONASS, BeiDou, Galileo. They are characterized by different completeness of satellite constellation deployment, which determines the accuracy of operational positioning in a particular geographical point, which depends on number of satellites available for observation, as well as the characteristics of the receiver, landscape features, weather conditions and the possibility of using differential corrections. The widespread use of differential corrections at the moment is not possible due to the fact that number of stable operating reference stations is limited - the Earth is covered by them unevenly; reliable data networks necessary for the transmission of differential corrections are also not deployed everywhere; budget versions of single-channel receivers of the navigation signal are widely used, which do not allow the use of differential corrections. In this case, there is a problem of operational choice of the system or a combination of satellite positioning systems, providing the most accurate navigation data. This paper presents a comparison of static and dynamic methods for selecting a system or a combination of satellite positioning systems that provide the most accurate definition of the object's own coordinates when using a single-channel receiver of navigation signals in offline mode. The choice is made on the basis of statistical analysis of data obtained from satellite positioning systems. During the analysis, the results of post-processing of data obtained from satellite navigation systems and refined with the use of differential corrections of navigation data were compared.

Keywords: Navigation of Autonomous Mobile Objects, Statistical Analysis of Navigation Data, Methods of Satellite Positioning System Selection.

Popov Sergey Gennad'evich — Ph.D., Associate Professor, Associate Professor, Telematics Department of Institute of Applied Mathematics and Mechanics, St. Petersburg and Peter the Great St. Petersburg Polytechnic University. Research interests: DBMS, robotics, satellite navigation systems. The number of publications — 70. popovserge@spbstu.ru; 29, Polytechnicheskaya, 195251, St.Petersburg, Russian Federation; office phone: +7(921)9613493.

Zaborovsky Vladimir Seregeevich — Ph.D., Dr.Sci., Professor, Professor, Telematics Department of Institute of Applied Mathematics and Mechanics, St. Petersburg and Peter the Great St. Petersburg Polytechnic University. Research interests: cyberphysics, quantum computing. The number of publications — 135. vlad@neva.ru; 29, Polytechnicheskaya, 195251, St.Petersburg, Russian Federation; office phone: +7(921)9398954.

Kurochkin Leonid Mikhailovich — Ph.D., Associate Professor, Associate Professor, Telematics Department of Institute of Applied Mathematics and Mechanics, St. Petersburg and

Peter the Great St. Petersburg Polytechnic University. Research interests: robotics, intelligent transport networks. The number of publications — 30. kurochkinl@spbstu.ru; 29, Polytechnicheskaya, 195251, St.Petersburg, Russian Federation; office phone: +7(921)3465767.

Sharagin Maksim Pavlovich — Ph.D. student, Telematics Department of Institute of Applied Mathematics and Mechanics, St. Petersburg and Peter the Great St. Petersburg Polytechnic University. Research interests: machine learning, robotics. The number of publications — 3. msharagin@gmail.com; 29, Политехническая, 195251, Санкт-Петербург, Russian Federation; office phone: +7(812)5526521.

Zhang Lei — Ph.D., Professor, School of Computer Science and Software Engineering, East China Normal University (ECNU). Research interests: satellite navigation systems, robotics. The number of publications — 18. lzhang@ce.ecnu.edu.cn; 500, Dongchuan Rd., Minhang district, 200241, Шанхай, China; office phone: +86-18818223050.

Acknowledgment. State order, basic part: 2.9198.2017 / 8.9 Sustainable management of a heterogeneous grouping of cyber-physical objects in the conditions of space-time uncertainty.

References

1. Schüler T. On ground-based GPS tropospheric delay estimation. Doctor's Thesis, Studiengang Geodsie und Geoinformation. 2001. vol. 73. 364 p.
2. Khutorova O.G. et al. [Sensiing of the Structure of the Radio Wave Refractivity in the Troposphere by a Network of Satellite Navigation System Receivers in the City Of Kazan]. *Izvestiya vysshih uchebnyh zavedenij. Radiofizika – Radiophysics and Quantum Electronics.* 2011. vol. 54. no. 1. pp. 1–8. (In Russ.).
3. Popov S.G., Popov M.V. [Research methods to improve the accuracy of navigation]. *Materialy nauchnogo foruma s mezhdunarodnym uchastiem "Nedelya nauki SPbPU"* [Materials of the scientific forum with international participation "Science Week SPbPU"]. 2016. pp. 8–10. (In Russ.).
4. Adrados C. et al. Global Positioning System (GPS) location accuracy improvement due to Selective Availability removal. *Comptes Rendus Biologies*. 2012. vol. 325. no. 2. pp. 165–170.
5. Siejka Z. Validation of the accuracy and convergence time of real time kinematic results using a single galileo navigation system. *Sensors.* 2018. vol. 18(8). pp. 2412.
6. Takács B., Siki Z., Markovits-Somogyi R. Extension of RTKLIB for the Calculation and Validation of Protection Levels. International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences. 2017. vol. 42. 6 p.
7. Suldi A.M., Samad A.M.A., Hashim H. Feasible study on establishment of a GPS reference station. 2012 IEEE 8th International Colloquium on Signal Processing and its Applications. 2012. pp. 362–366.
8. Makarenko G.K., Aleshechkin A.M. [Study on filtering algorithms to determine the coordinates of an object by the signals of satellite navigation systems]. *Doklady Tomskogo gosudarstvennogo universiteta sistem upravleniya i radioehlektroniki – Proceedings of TUSUR journal.* 2012. vol. 2-2(26). pp. 15–18. (In Russ.).
9. Gavrilov A.V. [Using the Kalman filter to solve the Problem of Refining the Coordinates UAV]. *Sovremennye problemy nauki i obrazovaniya – Modern problems of science and education.* 2015. vol. 1-1. pp. 1784–1784. (In Russ.).
10. Wans S.C. et al. Study on an intelligent fault-tolerant technique for multiple satellite configured navigation under highly dynamic conditions. Science China Information Sciences. 2011. vol. 54. pp. 529–541.

11. Hu J. et al. An extended Kalman filter and back propagation neural network algorithm positioning method based on anti-lock brake sensor and global navigation satellite system information. *Sensors*. vol. 18. no. 19. pp. 2753.
12. Ferrara N.G. et al. A new implementation of narrowband interference detection, characterization, and mitigation technique for a software-defined multi-GNSS receiver. *GPS Solutions*. 2018. vol. 22(4). pp. 106.
13. Skrypnik O.N., Nechaev E.E., Arefev R.O. [Construction and Analysis of GPS Accuracy Fields on the Basis of Hardware-Software Means NI GPS Simulation Toolkit]. *Nauchnyj vestnik Moskovskogo gosudarstvennogo tekhnicheskogo universiteta grazhdanskoj aviacii – Civil Aviation High TEVHNOLOGIES.* 2014. vol. 209. pp. 5–12. (In Russ.).
14. Dvorkin V.V., Karutin S.N. [Precise Positioning According to GNSS Signal]. *Sibirskij zhurnal nauki i tekhnologij – Siberian Journal of Science and Technology.* 2013. vol. 6(52). pp. 70–76. (In Russ.).
15. Guo J. et al. Multi-GNSS precise point positioning for precision agriculture. *Precision Agriculture.* 2018. vol. 19(5). pp. 895–911.
16. Li X. et al. Accuracy and reliability of multi-GNSS real-time precise positioning: GPS, GLONASS, BeiDou, and Galileo. *Journal of Geodesy.* 2015. vol. 89. no. 6. pp. 607–635.
17. Serapinas B.B. *Global'nye sistemy pozicionirovaniya* [Global Positioning Systems]. M.: IFK "Katalog". 2002. 105 p. (In Russ.).
18. Li X. et al. Precise Positionig with Current Multi-Constellation Global Navigation Satellite Systems: GPS, GLONASS, Galileo and BeiDou. *Scientific Reports*. 2015. vol. 5. pp. 8328.
19. Xiaolei Y., Yongrong S., Jianye L., Jianfeng M. Fast Algorithm of Selecting Satellites for Multiple Satellite Integrated Navigation Engineering. 2009 WRI World Congress on Computer Science and Information Engineering. 2009. vol. 5. pp. 121–125.
20. The International GNSS Service. Available at: http://www.igs.org/ (accessed: 14.09.18).
21. Kurochkin L.M., Kurochkin M.A., Popov S.G., Popov M.V. [Result of experimental studies of positioning accuracy using various satellite navigation systems]. *Nauchno-tekhnicheskie vedomosti SPbGPU. Informatika. Telekommunikacii. Upravlenie – St. Petersburg State Polytechnical University Journal. Computer Science. Telecommunication and Control Systems.* 2018. Issue 10. vol. 4. pp. 79–88. (In Russ.).

УДК 004.89 **DOI** 10.15622/sp.18.2.326-353

О.К. Головнин, А.А. Столбова

ВЕЙВЛЕТ-АНАЛИЗ КАК ИНСТРУМЕНТ ИССЛЕДОВАНИЯ ХАРАКТЕРИСТИК ДОРОЖНОГО ДВИЖЕНИЯ ДЛЯ ИНТЕЛЛЕКТУАЛЬНЫХ ТРАНСПОРТНЫХ СИСТЕМ В УСЛОВИЯХ НЕДОСТАЮЩИХ ДАННЫХ

Головнин О.К., Столбова А.А. **Вейвлет-анализ как инструмент исследования характеристик дорожного движения для интеллектуальных транспортных систем в условиях недостающих данных.**

Аннотация. Мероприятия по получению достоверной информации о текущем состоянии транспортных потоков являются необходимыми для реализации эффективных методов управления, предлагаемых современными интеллектуальными транспортными системами. Часто встречающейся проблемой при получении характеристик транспортных потоков с технических устройств является потеря исходных данных, которая приводит к необходимости решения задачи анализа неэквидистантных временных рядов. Эффективным подходом к исследованию неэквидистантных данных выступает спектральный анализ, требующий приведения неэквидистантного процесса к равномерному виду, например, восстановлением пропущенных отсчетов, что ведет к появлению погрешности датирования.

Для анализа и интерпретации нестационарных неэквидистантных временных рядов, полученных из систем мониторинга транспортных потоков, предлагается использовать метод вейвлет-преобразования с подстройкой интервалов дискретизации, результатом которого является частотно-временная развертка с равномерным представлением. Вейвлет-анализ применен к макроскопическим характеристикам транспортного потока, описывающим динамическое состояние транспортной сети в масштабе города или области.

Программное обеспечение, реализующее предложенный метод вейвлет-анализа характеристик транспортных потоков, разработано с использованием атрибутно-ориентированного подхода на фреймворке интеллектуальной транспортной геоинформационной системы ITSGIS. Интеграция разработанного программного обеспечения с интеллектуальной транспортной системой выполняется на трех уровнях: уровень данных — получение исходных данных от систем мониторинга; уровень бизнес-логики — представление обработанных данных для сервисов интеллектуальной транспортной системы; уровень представления пользователю — встраивание визуальных компонентов в пользовательские интерфейсы ITSGIS.

Вейвлет-анализ характеристик транспортных потоков проведен с использованием вейвлетов Морле на примере трех различных по интенсивности и скорости движения участков автодорог в городе Орхус (Дания). В качестве набора данных для анализа выступает недельный интервал с понедельника по воскресенье. Выполнен анализ данных о средней скорости, числе транспортных средств и среднем времени прохождения участка улично-дорожной сети. Построены и проанализированы вейвлет-спектры и скейлограммы, выявлены общие зависимости в частотном расположении экстремумов, выявлены различия в спектральной мощности.

Разработанное программное обеспечение, внедренное в интеллектуальную транспортную систему ITSGIS, проходит экспериментальную апробацию при решении практических задач государственных и муниципальных служб на территории России.

Ключевые слова: транспортный поток, вейвлет, интеллектуальная транспортная система, спектральный анализ, частотный анализ, ИТС.

1. Введение. Мероприятия по получению достоверной информации о текущем состоянии транспортных потоков (ТП) являются необходимыми для реализации эффективных методов управления, пред-

лагаемых современными интеллектуальными транспортными системами [1]. С точки зрения мониторинга, характеристикам ТП присущи нестабильность, многообразие и практическая сложность получения [2]. Эффективность мониторинга характеристик ТП может быть повышена за счет применения средств автоматизации процессов сбора, хранения, планирования и анализа информации [3].

Активно используются и развиваются методы и средства мониторинга характеристик ТП, которые используют данные, полученные с помощью петлевых датчиков [4], фото- и видеокамер [5], спутниковых навигационных систем [6], операторов сотовой связи [7], дистанционного зондирования Земли [8]. Существуют методы мониторинга и прогнозирования характеристик ТП на основе гибридных подходов, например спектрально-статистического [9] и пространственно-временного [10]. Методы прогнозирования характеристик ТП, применяющие к статистическим данным нейросетевой анализ [11] и выполняющие адаптации макроскопических моделей ТП [12], ставят своей целью снижение влияния неполноты информации о текущем состоянии ТП на прогнозные значения за счет предыдущих периодов.

Применение сетей Петри в интеллектуальных транспортных системах показывает высокую эффективность в случае неравномерного или высокого транспортного спроса, но качество решений на их основе тесно связано с качеством используемых динамических моделей ТП [13]. Традиционная теория сетей и систем массового обслуживания не позволяет построить точную модель ТП, поскольку ТП, движущийся по улично-дорожной сети, не является пуассоновским, что приводит к появлению моделей, в которых ТП имеет распределение Эрланга [14]. Такие модели применяются для потоков в компьютерных сетях, но вследствие особенностей ТП, состоящего из различных транспортных средств, не могут быть использованы в предзаторовых и заторовых состояниях. Однако отдельные методы теории массового обслуживания с успехом применяются для решения проблемы минимизации транспортных задержек [15].

Разрабатываются модели, источником данных для которых выступают социальные сети, например интегрированная модель прогнозирования скорости движения [16], однако неопределенность местоположения данных не позволяет применять их для управления движением. Методы глубокого машинного обучения в прогнозировании различных характеристик ТП показывают хорошие результаты [17], но требуют накопленного объема исходных достоверных данных.

Повышение достоверности данных о ТП достигается за счет применения моделей прогнозирования временных рядов [18] и статистических методов [19]. Наиболее эффективные результаты в отношении скорости расчетов и точности получаемых значений показывают

методы спектрального анализа [20]. Однако частотные методы спектрального анализа не позволяют определить время существования частоты в исследуемом процессе, что приводит к ограниченным возможностям при анализе процессов, нестационарных по частоте. Вейвлет-анализ относится к частотно-временным методам и позволяет реализовать анализ времени существования частоты в процессе, является одним из активно развивающихся методов спектрального анализа нестационарных процессов [21, 22] и прогнозирования [23].

Подходы, основанные на вейвлет-анализе, используются в прогнозировании временных рядов, являющихся волатильными и гетероскедастичными [24, 25]. Вейвлеты успешно применяются для прогнозирования и классификации в нейронных сетях [26-28], для подавления шумов инерционных датчиков при управлении движением, оценке поведения водителей и мониторинге состояния автомобильных дорог [29], прогнозирования показателей ТП, включая прогноз средней скорости ТП [30]. В работе [31] рассмотрена гибридная модель прогнозирования характеристик ТП на основе декомпозиции мод, учитывающая свойственные им характеристики.

Стоит отметить, что часто встречающейся проблемой при анализе ТП является потеря данных, которая ведет к неэквидистантности исходных данных. Например, в работе [32] описывается подход к решению проблемы пропущенных значений при анализе данных с петлевых датчиков. В этом случае при анализе неэквидистантный процесс приводится к равномерному виду — такой подход является простым, не требует разработки новых алгоритмов, однако приводит к появлению погрешности датирования, следовательно, проблемы анализа неэквидистантных процессов решены не в полной мере.

Таким образом, цель работы — разработка метода и программного обеспечения для вейвлет-анализа характеристик ТП в частотной и временной областях без восстановления пропущенных отсчетов. Программное обеспечение должно масштабироваться и тиражироваться для возможности использования на практике в составе интеллектуальной транспортной системы.

2. Метод вейвлет-анализа характеристик транспортных потоков. Данные о характеристиках ТП, получаемые в общем случае из ненадежных систем мониторинга, представляют собой временной ряд:

$$\begin{gathered}\left\{x_i, \Delta t_i\right\}^{i=1\ldots N}; \\ \Delta t_i = t_{i+1} - t_i.\end{gathered} \tag{1}$$

где i — номер отсчета, x_i — значения временного ряда t_i — время отсчета.

К типовым рядам данных, получаемых в результате мониторинга характеристик ТП, относятся ряды с пропусками наблюдений [33]:

$$\begin{cases} x_i = x_i\left(t_i\right); \\ t_i = \sum\limits_{k=1}^{i} Y_k \cdot \Delta t_0, \end{cases} \tag{2}$$

где Y_i — случайная величина, распределенная по сдвинутому на единицу закону Паскаля, Δt_0 — интервал принудительной дискретизации.

Так как в случае с пропусками наблюдений интервал дискретизации временных рядов является случайной величиной, то получаемый ряд характеристик ТП относится к неэквидистантным. Для анализа и интерпретации нестационарных неэквидистантных временных рядов, полученных из систем мониторинга ТП, предлагается использовать метод вейвлет-преобразования с частотно-временной разверткой с равномерным представлением:

1. Выбрать интервал принудительной дискретизации временного ряда Δt_0.

2. С учетом полученного интервала восстановить массив сдвигов *b*, соответствующий равномерному временному ряду.

3. При расчете вейвлет-коэффициентов $W\left(a,b\right)$ на каждом шаге сдвига *b* необходимо пересчитать вейвлет ψ с новыми неравномерными интервалами дискретизации, соответствующими интервалам временного ряда $t_{i+1} - t_i$.

На шаге 1 метода используется минимально возможное значение интервала принудительной дискретизации временного ряда $\Delta t_0 = \min\limits_k \Delta t_k$, поскольку наиболее вероятная причина неэквидистантности в данных о ТП — пропуски наблюдений (2).

Применяемое непрерывное вейвлет-преобразование имеет следующий вид:

$$W(a,b) = \frac{1}{\sqrt{a}} \int\limits_{-\infty}^{\infty} f(t)\psi\left(\frac{t-b}{a}\right)dt, \tag{3}$$

где $f\left(t\right)$ — исследуемый процесс; $\psi\left(t\right)$ — выбранный вейвлет; $a \neq 0$ — параметр масштаба; $b \geq 0$ — параметр сдвига.

Численно-аналитический подход к вычислению вейвлет-преобразования не позволяет повысить точность и скорость вычисления, поскольку результирующее выражение содержит неберущийся интеграл. Применение классического метода Симпсона при вычислении вейвлет-преобразования рядов с пропусками наблюдений является невозможным ввиду того, что он предполагает равномерность исходного ряда. Интерполяция подынтегральной функции полиномом в форме Ньютона является довольно ресурсоемкой процедурой, поскольку потребует итерационного расчета коэффициентов вейвлет-преобразования на каждом шаге. В связи с этим для вычисления используется метод трапеций, а выражение для оценки коэффициентов вейвлет-преобразования имеет следующий вид [34]:

$$W(a,b)=\frac{1}{\sqrt{a}}\left(x_0\psi\left(\frac{t_0-b}{a}\right)\frac{(t_1-t_0)}{2}+\sum_{i=1}^{N-2}x_i\psi\left(\frac{t_i-b}{a}\right)\left(\frac{t_{i+1}-t_{i-1}}{2}\right)+\right.$$
$$\left.+x_{N-1}\psi\left(\frac{t_{N-1}-b}{a}\right)\frac{(t_{N-1}-t_{N-2})}{2}\right), \quad (4)$$

где N — число отсчетов исследуемого временного ряда.

Подобно алгоритму, примененному в [35], предлагается использовать вейвлеты Морле для анализа характеристик ТП:

$$\psi(t)=\exp(-ikt)\exp\left(-\frac{t^2}{2\sigma^2}\right). \quad (5)$$

Вейвлет Морле обладает преимуществом перед другими базисными вейвлетами с точки зрения анализа характеристик ТП — позволяет влиять на выбор ширины окна и доминантной частоты, отвечающей за избирательность вейвлета, что дает возможность настроить функцию для получения наиболее точных результатов как по частоте, так и по времени.

Для анализа результатов применим вейвлет-спектр, описывающий распределение энергии по масштабам:

$$S(a_i,b_j)=\left|W(a_i,b_j)\right|^2. \quad (6)$$

Также используем скейлограмму, которая имеет следующий вид [36]:

$$Sg\left(a_i, b_j\right) = \frac{1}{N_b} \sum_{j=0}^{N_b-1} S\left(a_i, b_j\right). \quad (7)$$

Применим вейвлет-анализ к макроскопическим характеристикам транспортного потока, описывающим динамическое состояние транспортной сети в масштабе города или области, — средней скорости $v(t)$, интенсивности $I(t)$ и плотности $k(t)$:

$$v(t) = \frac{I(t)}{k(t)}; \quad (8)$$

$$I(t) = \frac{\partial Q(t)}{\partial t}, \quad (9)$$

где $Q(t)$ — количество автомобилей на участке улично-дорожной сети в момент времени t.

3. Реализация программного обеспечения. Программное обеспечение, реализующее предложенный метод вейвлет-анализа характеристик ТП, разработано с использованием атрибутно-ориентированного подхода [37] на фреймворке ITSGIS [38], предназначенном для построения интеллектуальных транспортных геоинформационных систем.

Архитектура программного обеспечения приведена на рисунке 1. Интеграция с интеллектуальной транспортной системой выполняется на трех уровнях: уровень данных — получение исходных данных от систем мониторинга; уровень бизнес-логики — представление обработанных данных для сервисов интеллектуальной транспортной системы; уровень представления пользователю — встраивание визуальных компонентов в пользовательские интерфейсы ITSGIS.

Разработанное программное обеспечение для вейвлет-анализа характеристик ТП выполняет следующие функции:

– извлечение последовательностей данных с равномерной и неравномерной дискретизацией из различных источников данных: из файлов формата XML, JSON, CSV, из баз данных и онтологических баз знаний;

– получение спектральных характеристик процесса;

– вычисление вейвлет-функций и коэффициентов вейвлет-преобразования;

– построение вейвлет-спектров и скейлограмм;

– вычисление погрешности преобразований.

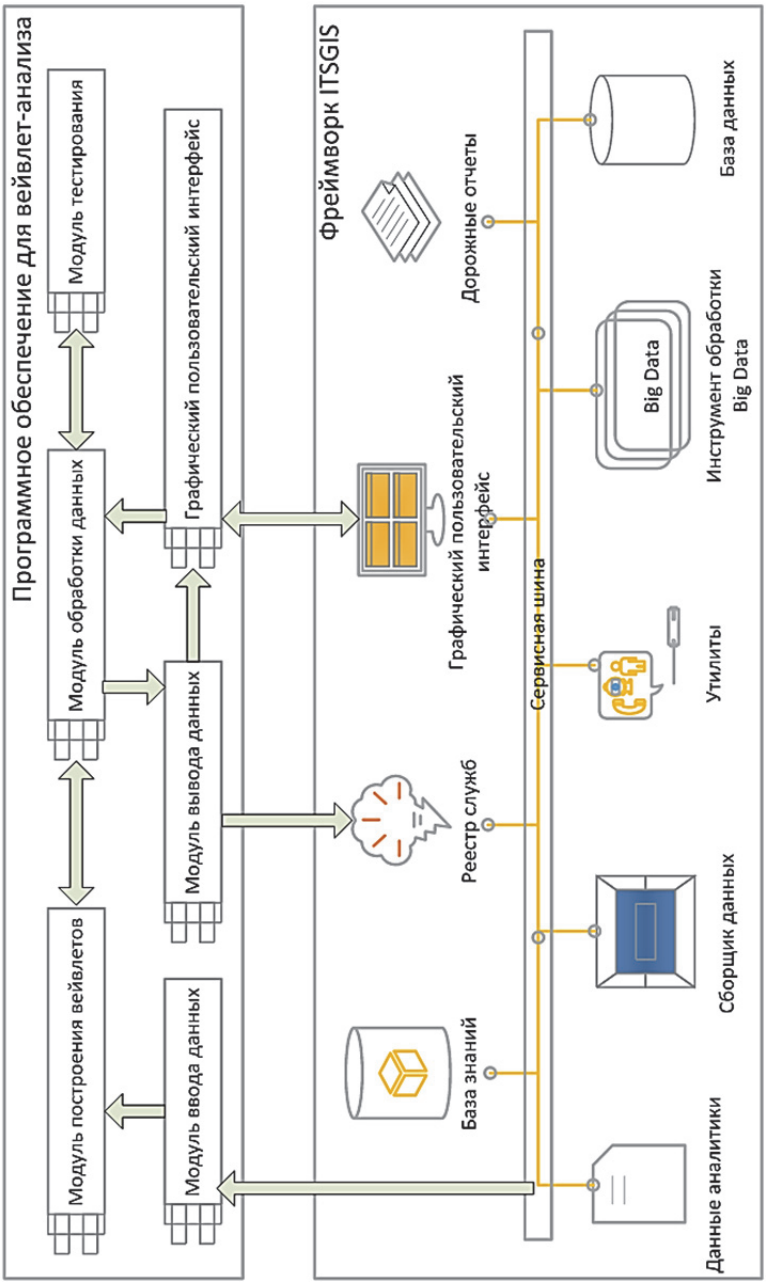


Рис. 1. Архитектура программного обеспечения

Схема реализованного алгоритма расчета коэффициентов вейвлет-преобразования приведена на рисунке 2.

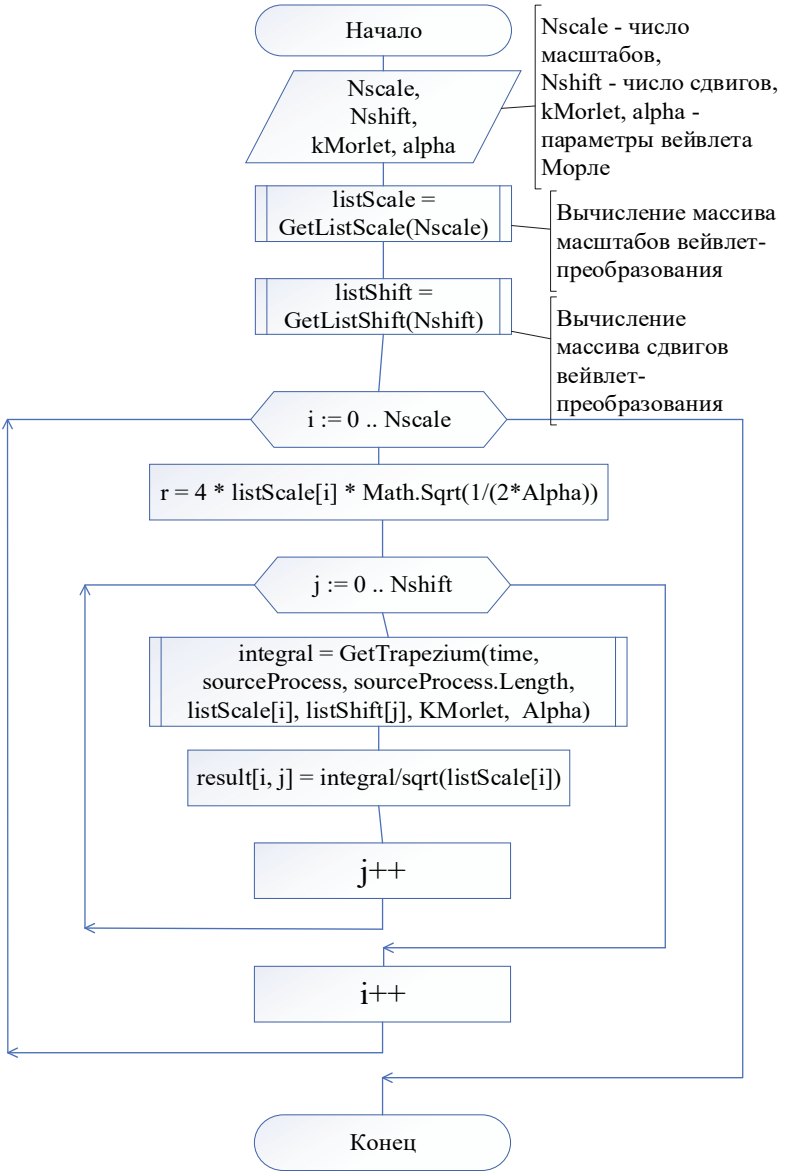


Рис. 2. Схема алгоритма вычисления коэффициентов вейвлет-преобразования

В приведенном алгоритме используются:

– *GetListScale(Nscale)* — функция, возвращающая массив масштабов вейвлет-преобразования;

– *GetListShift(Nshift)* — функция, возвращающая массив сдвигов вейвлет-преобразования;

– *r* — ширина вейвлета;

– *GetTrapezium(time, sourceProcess, sourceProcess.Length, a, b, KMorlet, Alpha)* — функция, возвращающая значение интеграла, вычисленного методом трапеций. Входными параметрами функции являются временные метки, значения и число отсчетов исходного процесса, текущие значения масштаба и сдвига, параметры вейвлета Морле.

Фрагменты кода на языке программирования C#, реализующие алгоритм вычисления коэффициентов вейвлет-преобразования, представлены в листинге 1.

```
/// <summary>
/// Вычисление вейвлет-преобразования.
/// </summary>
/// <param name="sourceProcess"> Значения исходного процесса. </param>
/// <param name="time"> Временные отсчеты исходного процесса. </param>
/// <param name="listScale"> Список масштабов. </param>
/// <param name="listShift"> Список сдвигов. </param>
/// <returns> Коэффициенты вейвлет-преобразования. </returns>
double[,] GetWaveletTransform(double[] sourceProcess, double[] time, double[] listScale,
                              double[] listShift)
{
    double[,] result = new double[Nscale, Nshift];
    for (int i = 0; i < Nscale; i++)
    {
        // Вычисление ширины вейвлета.
        double r = (4 * listScale[i] * Math.Sqrt(1 / (2 * Alpha)));

        for (int j = 0; j < Nshift; j++)
        {
            // Вызов функции вычисления интеграла.
            double integral = GetTrapezium(time, sourceProcess,
            sourceProcess.Length, listScale[i], listShift[j], KMorlet,
            Alpha);

            // Вычисление коэффициента вейвлет-преобразования.
            result[i, j] = Math.Abs(integral / Math.Sqrt(listScale[i]));
        }
    }
    return result;
}
```

Листинг 1. Вычисление коэффициентов вейвлет-преобразования

Данные о ТП — атрибутные данные, представляющие собой семантику объектов, процессов и явлений транспортной инфраструктуры, имеют пространственно-временную привязку. Таким образом, программное обеспечение должно иметь возможность манипулирования данными, имеющими различные атрибутные, топологические и функциональные связи. Уровень хранения и обработки данных реализуется на системах управления базами данных реляционного типа, поддерживающих геометрические примитивы OGC в представлении WKB/WKT. Уровень бизнес-логики реализуется как приложение WCF, функционирующее в виде службы Windows или IIS на выделенном сервере приложений. Уровень представления пользователю реализуется как настольное приложение. Программное обеспечение реализовано для .NET Framework 4.5 на языке C#. Графический интерфейс пользователя построен на основе WinForms, при расширенной визуализации используется технология OpenGL. Взаимодействие с сервисами интеллектуальной транспортной системы выполнено через протокол SOAP/XML. Для интеграции с различными реляционными источниками исходных данных используется технология объектно-реляционного отображения NHibernate.

Данные о ТП и объектах транспортной инфраструктуры представляются в виде доменных объектов, для загрузки которых применен паттерн проектирования «Загрузка по требованию». На каждый метод WCF-сервиса с помощью механизмов метапрограммирования прикрепляется разработанный суррогатный селектор, обеспечивающий выбор средства сериализации в зависимости от типа передаваемых данных. Сериализация объектов осуществляется рекурсивно в формат XML, геометрические примитивы и топологические отношения сериализуются в бинарное представление WKB. Получение данных обеспечивается WCF-сервисом, принимающим запросы на загрузку требуемых данных.

Для уменьшения объема передаваемых данных о ТП в WCF-сервис к конечной точке добавлено поведение компрессии данных, обеспечивающее ZIP-сжатие сериализованного XML-представления данных в WKB-код, который встраивается внутрь SOAP-сообщения. Детализованные данные о ТП могут быть представлены укрупненными низкодетализованными данными, сохраняющими основные характеристики ТП. Для таких случаев в уровень бизнес-логики добавлены механизмы симплификации.

4. Анализ характеристик транспортных потоков.

4.1. Подготовка исходных данных. Исходные данные о ТП для анализа получены из базы данных, подготовленной CityPulse для си-

стем класса «Умный город» [39]. Исходные данные представлены в формате CSV для города Орхус (Дания) [40]. Каждое измерение характеристик ТП выполняется через 5 минут. Отсутствие эквидистантности в исходных данных происходит из-за пропусков измерений. Каждая запись содержит сведения о ТП: среднее время прохождения участка улично-дорожной сети, средняя скорость, время измерения, число транспортных средств.

В качестве одного набора данных для анализа выступает недельный интервал с понедельника по воскресенье. Анализируются характеристики: средняя скорость, число транспортных средств, среднее время прохождения участка. Для каждого ряда данных проводится операция центрирования.

Проведен вейвлет-анализ неэквидистантных данных о характеристиках транспортных потоков для 3 участков дорог: с высокой интенсивностью движения (Nordjyske Motorvej), со средней интенсивностью движения (Randersvej), с низкой интенсивностью движения (Søftenvej). Характеристики исследуемых участков дорог приведены в таблице 1, на рисунке 3 показано их расположение на карте.

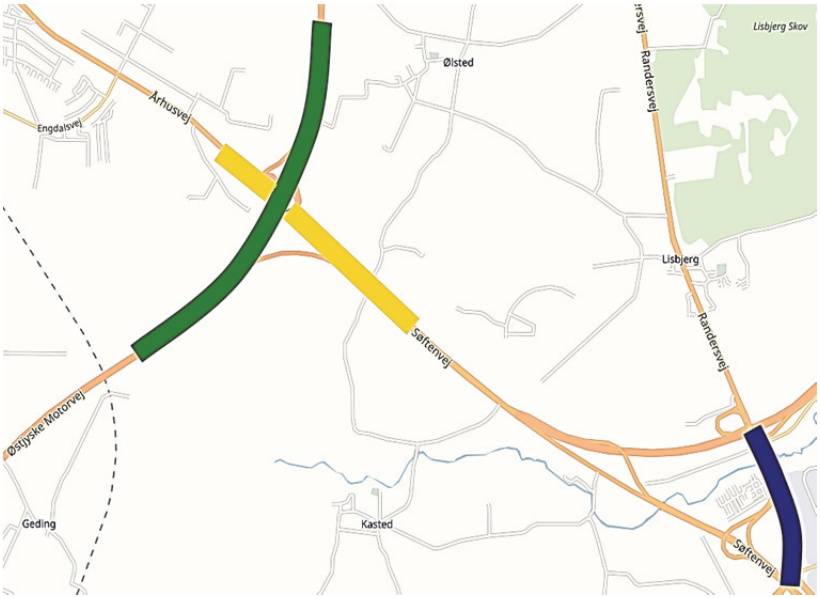


Рис. 3. Расположение исследуемых участков автодорог

Таблица 1. Исследуемые участки автодорог

Характеристика	Участок улично-дорожной сети		
	Nordjyske Motorvej	Randersvej	Søftenvej
Тип участка	Крупная автострада	Крупная автострада	Шоссе
Протяженность, м	2335	1195	2061
Скорость свободного движения, км/ч	112	81	51
Широта начала участка, Долгота начала участка	56.23489, 10.12501	56.21071, 10.17302	56.21508, 10.13978
Широта конца участка, Долгота конца участка	56.21740, 10.10702	56.20391, 10.17512	56.22579, 10.11658

4.2. Анализ средней скорости транспортного потока. Исходный ряд данных о средней скорости ТП приведен на рисунке 4, на рисунке 5 приведен центрированный ряд (на примере участка Randersvej).

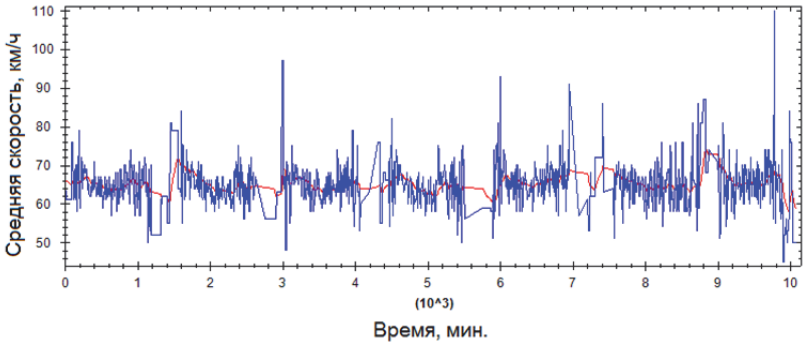


Рис. 4. Исходный ряд данных средней скорости

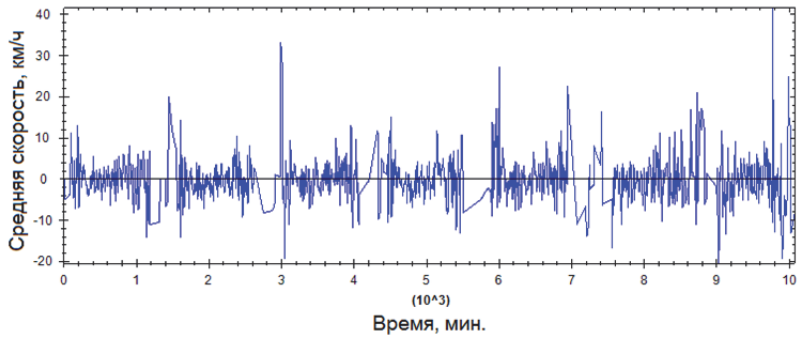


Рис. 5. Центрированный ряд данных средней скорости

Далее на рассчитанных по (6) графиках вейвлет-спектра по оси X откладывается время в минутах, а по оси Y — частота в Гц. Чем больше значение спектра, тем светлее рисунок. На скейлограммах, рассчитанных по (7), по оси X откладывается частота в Гц, а по оси Y — нормированная мощность. При расчетах использовались вейвлеты Морле (5).

Для участка Nordjyske Motorvej рассчитанный вейвлет-спектр для ряда данных о средней скорости ТП приведен на рисунке 6. Увеличение спектральной плотности заметно в понедельник, среду, четверг и пятницу, что характерно для автодороги такого класса.

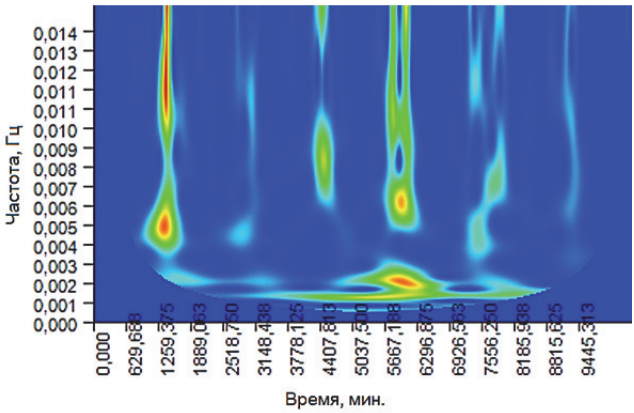


Рис. 6. Вейвлет-спектр для Nordjyske Motorvej (средняя скорость)

Для участка Randersvej вейвлет-спектр для ряда данных о средней скорости ТП приведен на рисунке 7. Заметное увеличение спектральной плотности отмечено только в пятницу.

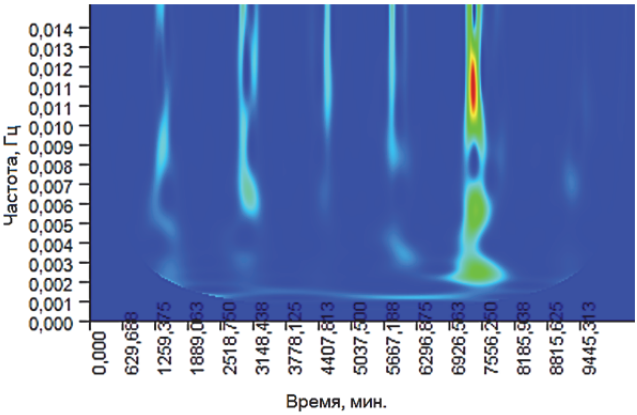


Рис. 7. Вейвлет-спектр для Randersvej (средняя скорость)

Для участка Søftenvej вейвлет-спектр для ряда данных о средней скорости приведен на рисунке 8. Увеличение спектральной плотности заметно во вторник и в пятницу.

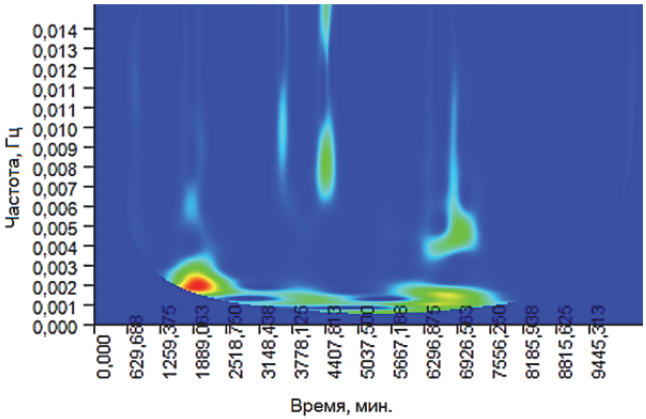


Рис. 8. Вейвлет-спектр для Søftenvej (средняя скорость)

Вейвлет-спектры показывают, что существует 2 диапазона частот для каждого дня недели: высокий 0.004–0.016 Гц и низкий 0.001–0.003 Гц. В пределах этих диапазонов происходит увеличение спектральной плотности. Для автодороги с большим скоростным режимом задействованы оба диапазона, для дороги с низким скоростным режимом — в основном диапазон низких частот. Диапазон высоких частот задействуется только в дни с наименьшей интенсивностью движения.

Построенные скейлограммы для трех участков дорог объединены на одном рисунке 9.

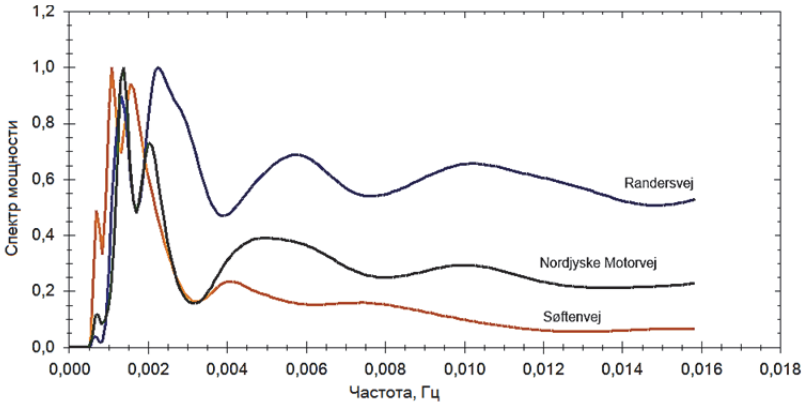


Рис. 9. Скейлограммы (средняя скорость)

Скейлограммы показывают, что для недельного интервала исследуемого временного ряда по средней скорости характерны 5 особых точек — локальные максимумы частоты в интервалах 0.0006–0.0006 Гц, 0.0009–0.0012 Гц, 0.0012–0.0021 Гц, 0.0039–0.0056 Гц, 0.0073–0.0101 Гц. Таким образом, можно сделать вывод о том, что общая закономерность временного ряда скоростей соблюдается независимо от уровня интенсивности ТП на участке дороги, направления движения их типа дороги. Расположение первых трех наибольших экстремумов в одной области низких частот свидетельствует о преобладании низкочастотных составляющих во всех временных рядах. Следовательно, участки дороги с различной интенсивностью движения и скоростным режимом имеют общий вид скейлограммы для средней скорости ТП, но вариации изменения средней скорости ТП различны.

4.3. Анализ числа транспортных средств на участке дороги. Исходный ряд данных для числа транспортных средств приведен на рисунке 10, на рисунке 11 приведен центрированный ряд (на примере участка Randersvej).

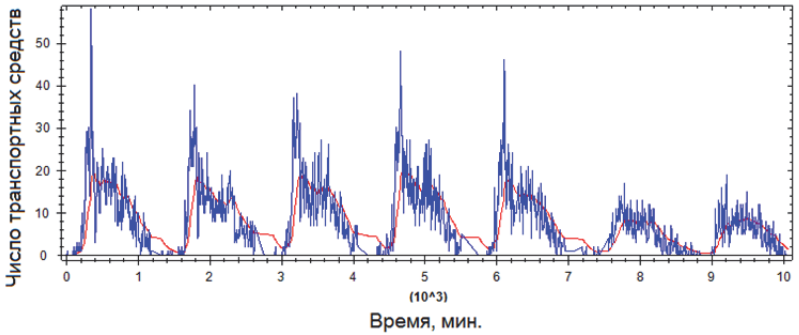


Рис. 10. Исходный ряд данных для числа транспортных средств

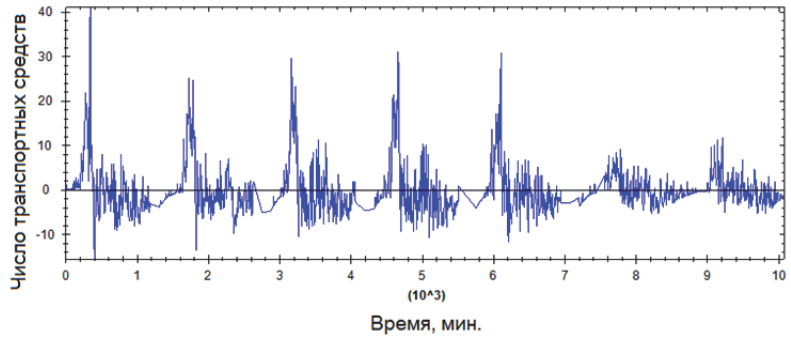


Рис. 11. Центрированный ряд данных для числа транспортных средств

Для участка Nordjyske Motorvej вейвлет-спектр для ряда данных о числе транспортных средств приведен на рисунке 12. Заметны увеличения спектральной плотности в высоком диапазоне частот во вторник, среду, четверг и пятницу.

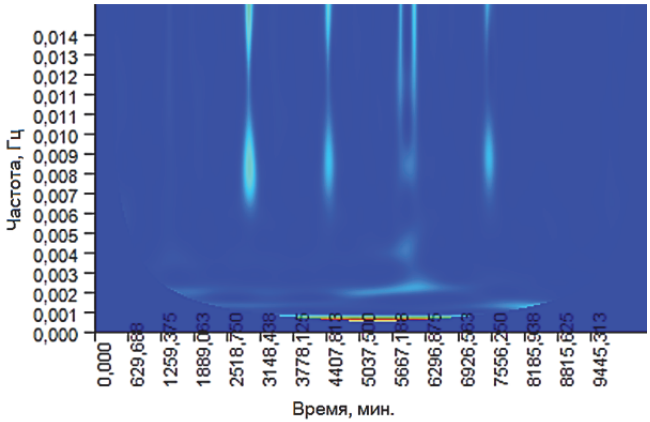


Рис. 12. Вейвлет-спектр для Nordjyske Motorvej (число транспортных средств)

Для участка Randersvej вейвлет-спектр для ряда данных о числе транспортных средств приведен на рисунке 13. Увеличение спектральной плотности в низком диапазоне частот заметно с понедельника по четверг.

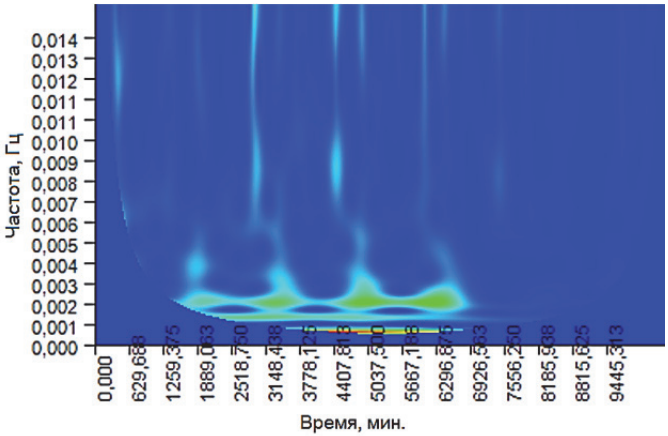


Рис. 13. Вейвлет-спектр для Randersvej (число транспортных средств)

Для участка Søftenvej вейвлет-спектр для ряда данных о числе транспортных средств приведен на рисунке 14. Заметны увеличения

спектральной плотности в среднем и высоком диапазонах частот в понедельник, вторник, среду и пятницу.

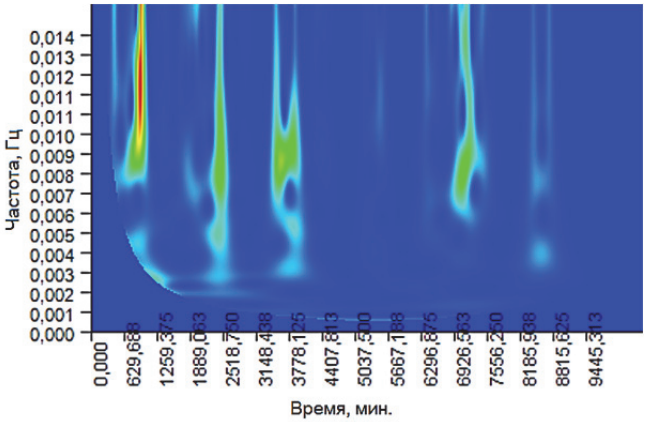


Рис. 14. Вейвлет-спектр для Søftenvej (число транспортных средств)

Скейлограммы для рядов данных по числу транспортных средств для трех участков дорог объединены на одном рисунке 15.

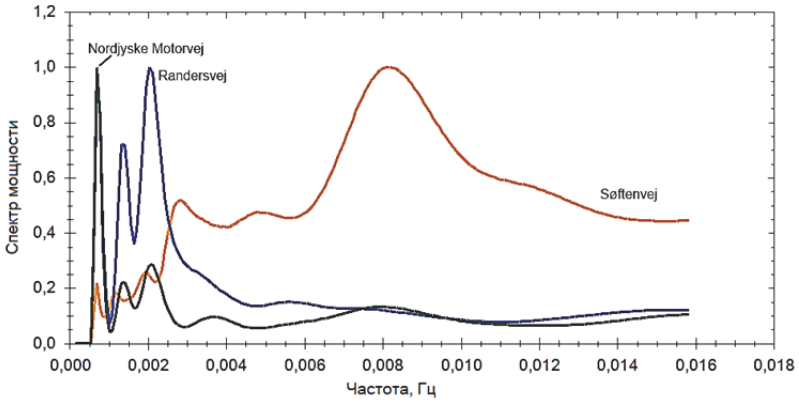


Рис. 15. Скейлограммы (число транспортных средств)

Частотное расположение первых трех экстремумов приходится на диапазоны частот 0.0005–0.0006 Гц, 0.0010–0.0013 Гц и 0.0018–0.0020 Гц и практически совпадают, что позволяет сделать вывод об одинаковом характере изменения интенсивности транспортных потоков в низкочастотном диапазоне, но разная мощность показывает на

отличия в значениях характеристик. Скейлограмма для участка дороги Søftenvej при увеличении частот сильно отличается и имеет экстремум в среднем диапазоне частот на 0,0080 Гц, на что оказывает влияние низкий скоростной режим и невысокая пропускная способность автодороги. Таким образом, интенсивность движения и скоростной режим на участке автодороги влияют на общий вид скейлограммы для числа транспортных средств и позволяют выявить наиболее значимые с точки зрения интенсивности ТП частотные составляющие.

4.4. Анализ среднего времени прохождения участка дороги. Исходный ряд данных для среднего времени прохождения участка приведен на рисунке 16, на рисунке 17 приведен центрированный ряд (на примере участка Randersvej).

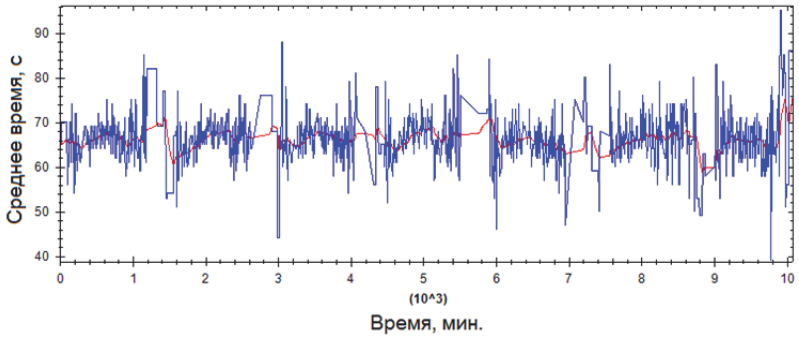


Рис. 16. Исходный ряд данных для среднего времени прохождения

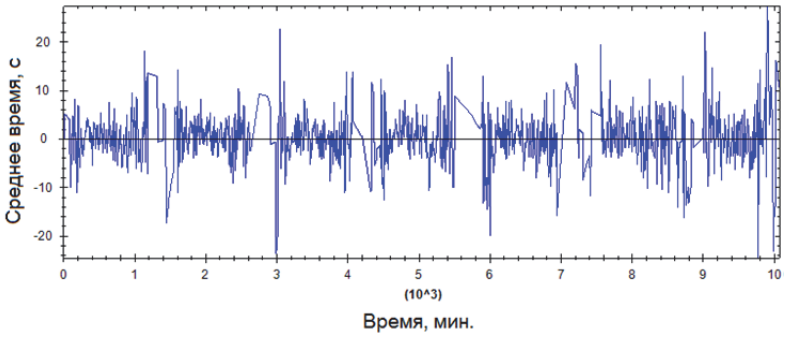


Рис. 17. Центрированный ряд данных для среднего времени прохождения

Для участка Nordjyske Motorvej вейвлет-спектр для ряда данных о среднем времени прохождения участка улично-дорожной сети приведен на рисунке 18. Заметны увеличения спектральной плотности в высоком диапазоне частот во вторник, среду, субботу и воскресенье.

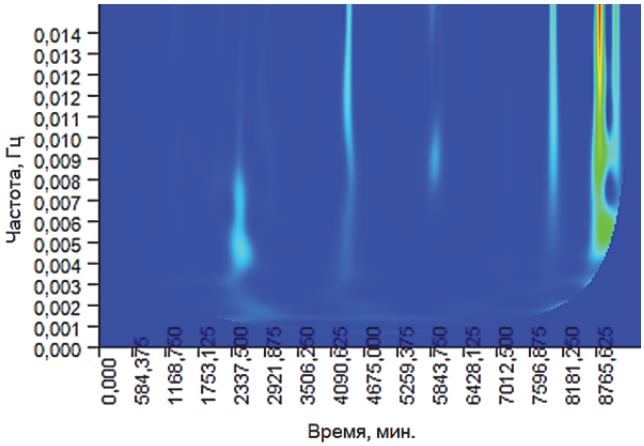


Рис. 18. Вейвлет-спектр для Nordjyske Motorvej (среднее время прохождения)

Для участка Randersvej вейвлет-спектр для ряда данных о среднем времени прохождения участка улично-дорожной сети приведен на рисунке 19. Увеличение спектральной плотности во всех диапазонах частот заметно с четверга по воскресенье.

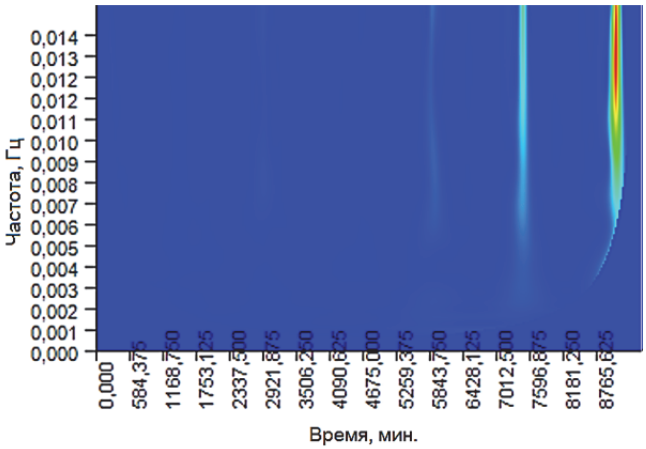


Рис. 19. Вейвлет-спектр для Randersvej (среднее время прохождения)

Для участка Søftenvej вейвлет-спектр для ряда данных о среднем времени прохождения участка улично-дорожной сети приведен на рисунке 20. Заметно увеличение спектральной плотности в субботу.

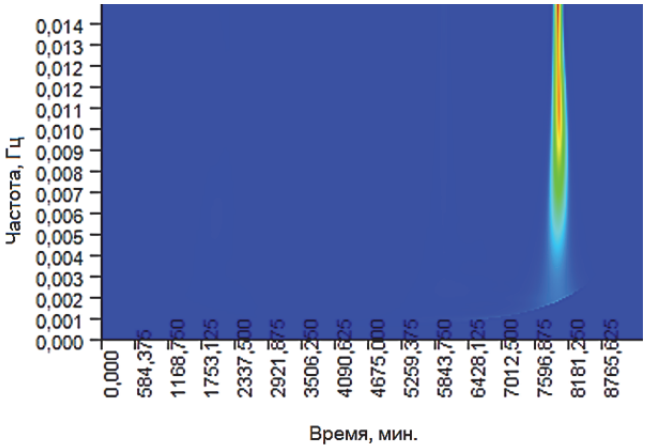


Рис. 20. Вейвлет-спектр для Søftenvej (среднее время прохождения)

Скейлограммы для рядов данных о среднем времени прохождения для трех участков дорог объединены на одном рисунке 21.

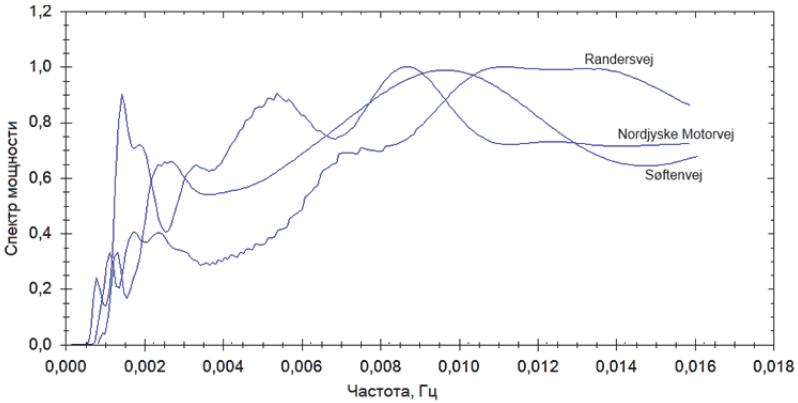


Рис. 21. Скейлограммы (среднее время прохождения)

Анализ скейлограмм, построенных для рядов данных о среднем времени прохождения участка, показывает отсутствие общих закономерностей для этих участков. Для автодороги с высокой интенсивностью ТП характерно появление экстремумов в зависимости от суточной неравномерности дорожного движения, автодороги со средней и низкой интенсивностью ТП практически не подвержены влиянию неравномерности.

5. Анализ интенсивности транспортных потоков в ITSGIS. На рисунках 22-23 приведены результаты анализа характеристик ТП в интеллектуальной транспортной геоинформационной системе ITSGIS по среднесуточным годовым данным, собранным за 2017 год в городском округе Самара. Выделение зон (полигональных участков улично-дорожной сети) и особых точек характеризуется значением интенсивности ТП, что позволяет выявлять зоны и точки напряжения в ТП путем нахождения экстремумов на карте интенсивности: светлее — низкая интенсивность, темнее — высокая интенсивность.

Интеллектуальные транспортные системы формируют различные управляющие воздействия в зависимости от зоны управления. В городских условиях наиболее необходимыми зонами для управления являются перекрестки на магистральных улицах (рисунок 22), где анализ характеристик ТП без восстановления пропущенных отсчетов позволит организовать эффективное управление даже в случае отсутствия или неисправности датчиков на некоторых перекрестках.

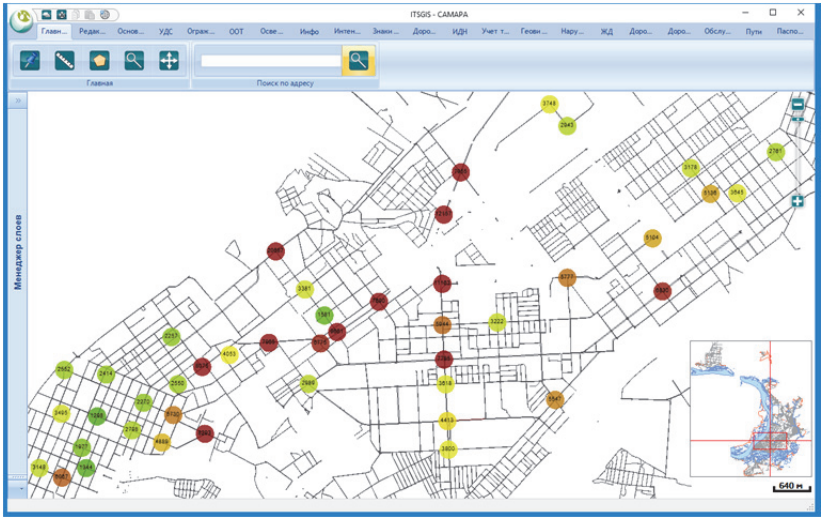


Рис. 22. Картограмма интенсивности транспортных потоков

Детализация до уровня перекрестка (рисунок 23) предоставит интеллектуальной транспортной геоинформационной системе возможность управления пофазным разъездом транспортных средств в зависимости от характеристик ТП, прибывающих к перекрестку.

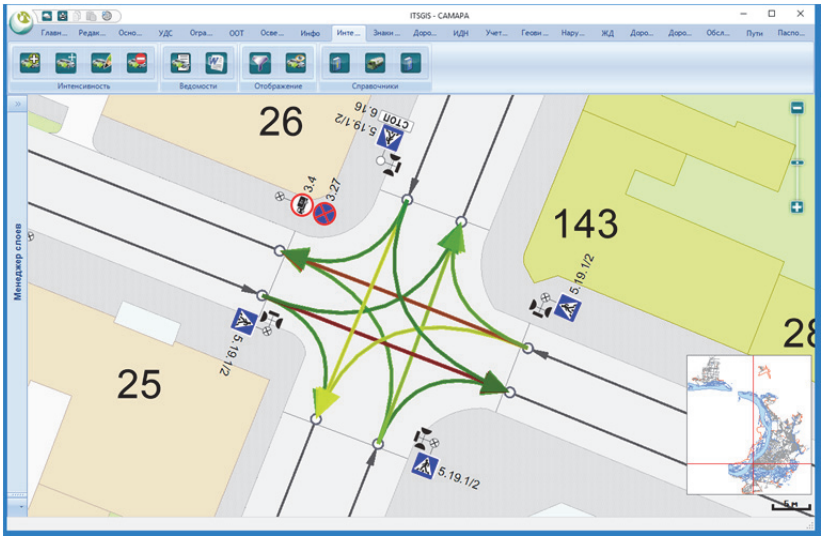


Рис. 23. Отображение интенсивности транспортных потоков на уровне перекрестка

6. Заключение. Предложен метод вейвлет-анализа характеристик ТП, который учитывает неэквидистантность данных и позволяет строить частотно-временную развертку с равномерным представлением без восстановления пропущенных отсчетов с подстройкой интервалов дискретизации. Разработанный метод реализован в виде программного обеспечения, встраиваемого в интеллектуальную транспортную систему ITSGIS.

Метод вейвлет-анализа применен при анализе характеристик ТП на примере трех различных по интенсивности и скорости движения участков автодорог в городе Орхус (Дания). Построены и проанализированы вейвлет-спектры и скейлограммы, выявлены общие зависимости в частотном расположении экстремумов, выявлены различия в спектральной мощности для различных по своим характеристикам участков автодорог.

Разработанное программное обеспечение, реализующее предлагаемый подход к анализу характеристик ТП и внедренное в интеллектуальную транспортную систему ITSGIS, проходит экспериментальную апробацию при решении практических задач государственных и муниципальных служб на территории Российской Федерации в городах и городских округах: Самара, Саранск, Тольятти, Рязань, Сургут (ХМАО–Югра), Владимир, Соль-Илецк, Новокуйбышевск, Жигулевск, Чапаевск, Трехгорный; в муниципальных районах Ульяновской и Самарской областей.

Литература

1. *Zhang R., Newman S., Ortolani M., Silvestri S.* A Network Tomography Approach for Traffic Monitoring in Smart Cities // IEEE Transactions on Intelligent Transportation Systems. 2018. vol. 19. no. 7. pp. 2268–2278.
2. *Taylor M.A., Bonsall P.W.* Understanding traffic systems: data analysis and presentation. 2nd edn. // London: Routledge. 2017. 443 p.
3. *Jain N.K., Saini R.K., Mittal P.* A Review on Traffic Monitoring System Techniques // Soft Computing: Theories and Applications. 2019. pp. 569–577.
4. *Askari H. et al.* A hybridized electromagnetic-triboelectric self-powered sensor for traffic monitoring: concept, modelling, and optimization // Nano Energy. 2017. vol. 32. pp. 105–116.
5. *Sahgal D. Ramesh A., Parida M.* Real-Time Vehicle Queue Detection at Urban Traffic Intersection using Image Processing // International Journal of Engineering Science and Generic Research. 2018. vol. 4. no. 2. pp. 12–15.
6. *Liu Z., Jiang S., Zhou P., Li M.* A participatory urban traffic monitoring system: the power of bus riders // IEEE Transactions on Intelligent Transportation Systems. 2017. vol. 18. no. 10. pp. 2851–2864.
7. *Bellavista P., Caselli F., Corradi A., Foschini L.* Cooperative Vehicular Traffic Monitoring in Realistic Low Penetration Scenarios: The COLOMBO Experience // Sensors. 2018. vol. 18. no. 3. pp. 822.
8. *Михеева Т.И., Федосеев А.А., Михеев С.В., Головнин О.К.* Метод синтеза тематического слоя объектов транспортной сети на основе материалов космической съемки // Информационный технологии. 2017. vol. 23. no. 11. pp. 808–816.
9. *ZhangY., Zhang Y., Haghani A.* A hybrid short-term traffic flow forecasting method based on spectral analysis and statistical volatility model // Transportation Research Part C: Emerging Technologies. 2014. vol. 43. pp. 65–78.
10. *Jiang Y. et al.* Spatio-temporal propagation of traffic jams in urban traffic networks. arXiv preprint 1705.08269. 2017.
11. *Moretti F., Pizzuti S., Panzieri S., Annunziato M.* Urban traffic flow forecasting through statistical and neural network bagging ensemble hybrid modeling // Neurocomputing. 2015. vol. 67. pp. 3–7.
12. *Zeroual A., Harrou F., Sun Y., Messai N.* Monitoring road traffic congestion using a macroscopic traffic model and a statistical monitoring scheme // Sustainable cities and society. 2017. vol. 35. pp. 494–510.
13. *Fu H. et al.* Modeling and integrated control of macroscopic heterogeneous traffic flow in large scale urban network using coloured Petri net // 98th TRB Annual Meeting: Compendium of Papers. 2019. pp. 19–04885.
14. *Yu L.* Queuing theory with heavy tails and network traffic modeling. URL: https://hal.archives-ouvertes.fr/hal-01891760 (дата обращения: 30.10.2018).
15. *Babicheva T.S.* The use of queuing theory at research and optimization of traffic on the signal-controlled road intersections // Procedia Computer Science. 2015. vol. 55. pp. 469–478.
16. *Lin L. et al.* Road traffic speed prediction: a probabilistic model fusing multi-source data // Proceedings of IEEE Transactions on Knowledge and Data Engineering. 2018. vol. 30. no. 7. pp. 1310–1323.
17. *Liu Z., Li Z., Wu K., Li M.* Urban Traffic Prediction from Mobility Data Using Deep Learning // IEEE Network. 2018. vol. 32. no. 4. pp. 40–46.
18. *Wang Y.D. et al.* Compression algorithm of road traffic data in time series based on temporal correlation // IET Intelligent Transport Systems. 2017. vol. 12. no. 3. pp. 177–185.

19. *Crawford F., Watling D.P., Connors R.D.* A statistical method for estimating predictable differences between daily traffic flow profiles // Transportation Research Part B: Methodological. 2017. vol. 95. pp. 196–213.
20. *Tchrakian T.T., Basu B., O'Mahony M.* Real-time traffic flow forecasting using spectral analysis // IEEE Transactions on Intelligent Transportation Systems. 2012. vol. 13. no. 2. pp. 519–526.
21. *Addison P.* The Illustrated Wavelet Transform Handbook // Boca Raton: CRC Press. 2017. 464 p.
22. *Bhattacharyya A., Singh L., Pachori R.B.* Fourier–Bessel series expansion based empirical wavelet transform for analysis of non-stationary signals // Digital Signal Processing. 2018. vol. 78. pp. 85–196.
23. *Yang S., Liu J.* Time Series Forecasting based on High-Order Fuzzy Cognitive Maps and Wavelet Transform // IEEE Transactions on Fuzzy Systems. 2018. vol. 26. no. 6. pp. 3391–3402.
24. *Wang J., Liu W.* Wavelet estimations for heteroscedastic super smooth errors // Communications in Statistics-Theory and Methods. 2018. pp. 1–21.
25. *Zeng X., Wang J.* Wavelet density deconvolution estimations with heteroscedastic measurement errors // Statistics & Probability Letters. 2018. vol. 134. pp. 79 – 85.
26. *Cheng Y., Zhang Y., Hu J., Li L.* Mining for similarities in urban traffic flow using wavelets // International Conference on Intelligent Transportation Systems Conference (ITSC). 2007. pp. 119–124.
27. *Wang J., Wang Z., Li J., Wu J.* Multilevel Wavelet Decomposition Network for Interpretable Time Series Analysis // Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD' 2018). 2018. pp. 2437–2446.
28. *Tian F., Ming W.T., Yun W.* Application of Wavelet Fuzzy Neural Network in Real Time Traffic Flow Forecasting // Proceedings of IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC). 2018. pp. 1452–1455.
29. *El-Wakeel A.S., Noureldin A., Hassanein H.S., Zorba N.* Utilization of Wavelet Packet Sensor De-noising for Accurate Positioning in Intelligent Road Services // Proceedings of the 14th International Wireless Communications & Mobile Computing Conference (IWCMC). 2018. pp. 1231–1236.
30. *Xiangxue W., Lunhui X.* Wavelet-based short-term forecasting with improved threshold recognition for urban expressway traffic conditions // IET Intelligent Transport Systems. 2018. vol. 12. no. 6. pp. 463–473.
31. *Zheng Z., Pan L., Pholsena K.* Mode Decomposition Based Hybrid Model for Traffic Flow Prediction // Proceedings of IEEE Third International Conference on Data Science in Cyberspace (DSC). 2018. pp. 521–526.
32. *Chen X. et al.* Kernel sparse representation with hybrid regularization for on-road traffic sensor data imputation // Sensors. 2018. vol. 18. no. 9. pp. 2884.
33. *Прохоров С.А.* Прикладной анализ неэквидистантных временных рядов // Самарский государственный аэрокосмический университет. 2001. 375 с.
34. *Прохоров С.А., Столбова А.А.* Программный комплекс анализа неэквидистантных временных рядов на основе непрерывного вейвлет-преобразования // Программные продукты и системы. 2017. Т. 30. № 4. С. 668–671.
35. *Khaymovich A.I., Prokhorov S.A., Stolbova A.A., Kondratyev A.I.* A model of milling process based on Morlet wavelets decomposition of vibroacoustic signals // International Conference Information Technology and Nanotechnology (ITNT). 2017. vol. 1904. pp. 135–140.

36. *Cannarile F., Baraldi P., Colombo P., Zio E.* A Novel Method for Sensor Data Validation based on the analysis of Wavelet Transform Scalograms // International Journal of Prognostics and Health Management, Prognostics and Health Management Society. 2018. vol. 9. no. 1. pp. 002.
37. *Golovnin O.K., Mikheeva T.I.* Attribute-driven network-centric urban transport process control system modeling // Journal of Physics: Conference Series. 2018. vol. 1096. no 1. pp. 012199.
38. ITSGIS Homepage. URL: http://www.itsgis.ru (дата обращения: 30.10.2018).
39. *Tönjes R. et al.* Real Time IoT Stream Processing and Large-scale Data Analytics for Smart City Applications // European Conference on Networks and Communications, poster session. 2014. 5 p.
40. CityPulse Dataset Collection. URL: http://iot.ee.surrey.ac.uk:8080/datasets.html (дата обращения: 30.10.2018).

Головнин Олег Константинович — канд. техн. наук, доцент, кафедра информационных систем и технологий, Самарский национальный исследовательский университет имени академика С.П. Королева (Самарский университет). Область научных интересов: системная инженерия, управление сложными системами, интеллектуальные транспортные системы, геоинформационные системы. Число научных публикаций — 150. golovnin@bk.ru; Московское ш., 34, 443086, Самара, Российская Федерация; р.т.: +7(846)267-46-72.

Столбова Анастасия Александровна — канд. техн. наук, доцент, кафедра информационных систем и технологий, Самарский национальный исследовательский университет имени академика С.П. Королева (Самарский университет). Область научных интересов: вейвлет-преобразование, неэквидистантные временные ряды. Число научных публикаций — 20. anastasiya.stolbova@bk.ru; Московское ш., 34, 443086, Самара, Российская Федерация; р.т.: +7(846)267-46-72.

DOI 10.15622/sp.18.2.326-353

O.K. GOLOVNIN, A.A. STOLBOVA

WAVELET ANALYSIS AS A TOOL FOR STUDYING THE ROAD TRAFFIC CHARACTERISTICS IN THE CONTEXT OF INTELLIGENT TRANSPORT SYSTEMS WITH INCOMPLETE DATA

Golovnin O.K., Stolbova A.A. **Wavelet Analysis as a Tool for Studying the Road Traffic Characteristics in the Context of Intelligent Transport Systems with Incomplete Data.**

Abstract. A frequent problem of traffic flow characteristics acquisition is data loss, which leads to uneven time series analysis. An effective approach to uneven data analysis is the spectral analysis, which requires obtaining process with a constant sampling interval, for example, by restoring missing data, which leads to the appearance of dating error. Thus, the main purpose of this study is to develop a method and software for wavelet analysis of traffic flow characteristics without restoring the missing data.

To analyze and interpret non-stationary uneven time series obtained from traffic monitoring systems, we propose the wavelet transformation method with adjustment of the sampling intervals, which results in a time-frequency domain with a constant sampling interval. Wavelet analysis is applied to the macroscopic traffic flow characteristics.

We developed the software for traffic flow wavelet analysis on the "ITSGIS" intelligent transport geo-information framework using the attribute-oriented approach.

Wavelet analysis of traffic flows characteristics using Morlet wavelets was accomplished for data analysis of the city of Aarhus, Denmark. Wavelet spectra and scalograms were constructed and analyzed, general dependencies in the frequency distribution of extremes, and differences in spectral power were revealed.

The developed software is being experimentally tested in solving practical problems of municipalities and road agencies in Russia.

Keywords: Traffic Flow, Wavelet, Intelligent Transport System, Spectral Analysis, Frequency Analysis, ITS.

Golovnin Oleg Konstantinovich — Ph.D., Associate Professor, Information Systems and Technologies Department, Samara National Research University. Research interests: system engineering, system control and management, intelligent transport systems, geographic information systems. The number of publications — 150. golovnin@bk.ru; 34, Moskovskoye sh., 443086, Samara, Russian Federation; office phone: +7(846)267-46-72.

Stolbova Anastasia Aleksandrovna — Ph.D., Associate Professor, Information Systems and Technologies Department, Samara National Research University. Research interests: wavelet transform, uneven time series. The number of publications — 20. anastasiya.stolbova@bk.ru; 34, Moskovskoye sh., 443086, Samara, Russian Federation; office phone: +7(846)267-46-72.

References

1. Zhang R., Newman S., Ortolani M., Silvestri S. A Network Tomography Approach for Traffic Monitoring in Smart Cities. *IEEE Transactions on Intelligent Transportation* Systems. 2018. pp. 2268–2278.
2. Taylor M.A., Bonsall P.W. Understanding traffic systems: data analysis and presentation. 2nd edn. London: Routledge. 2017. 443 p.
3. Jain N.K., Saini R.K., Mittal P. A Review on Traffic Monitoring System Techniques. Soft Computing: Theories and Applications. 2019. pp. 569–577.
4. Askari H. et al. A hybridized electromagnetic-triboelectric self-powered sensor for traffic monitoring: concept, modelling, and optimization. *Nano Energy*. 2017. vol. 32. pp. 105–116.

5. Sahgal D. Ramesh A., Parida M. Real-Time Vehicle Queue Detection at Urban Traffic Intersection using Image Processing. *International Journal of Engineering Science and Generic Research.* 2018. vol. 4. no. 2. pp. 12–15.
6. Liu Z., Jiang S., Zhou P., Li M. A participatory urban traffic monitoring system: the power of bus riders. *IEEE Transactions on Intelligent Transportation Systems.* 2017. vol. 18. no. 10. pp. 2851–2864.
7. Bellavista P., Caselli F., Corradi A., Foschini L. Cooperative Vehicular Traffic Monitoring in Realistic Low Penetration Scenarios: The COLOMBO Experience. *Sensors.* 2018. vol. 18. no. 3. pp. 822.
8. Mikheeva T.I., Fedoseev A.A., Mikheev S.V., Golovnin O.K. [Method of Transport Net Thematic Layer Synthesis via Remotely Sensed Imagery]. *Informatsionnye tekhnologii – Information Technologies.* 2017. vol. 23. no. 11. pp. 808–816. (In Russ.).
9. Zhang,Y., Zhang Y., Haghani A. A hybrid short-term traffic flow forecasting method based on spectral analysis and statistical volatility model. *Transportation Research Part C: Emerging Technologies.* 2014. vol. 43. pp. 65–78.
10. Jiang Y. et al. Spatio-temporal propagation of traffic jams in urban traffic networks. arXiv preprint 1705.08269. 2017.
11. Moretti F., Pizzuti S., Panzieri S., Annunziato M. Urban traffic flow forecasting through statistical and neural network bagging ensemble hybrid modeling. *Neurocomputing.* 2015. vol. 67. pp. 3–7.
12. Zeroual A., Harrou F., Sun Y., Messai N. Monitoring road traffic congestion using a macroscopic traffic model and a statistical monitoring scheme. *Sustainable cities and society*. 2017. vol. 35. pp. 494–510.
13. Fu H. et al. Modeling and integrated control of macroscopic heterogeneous traffic flow in large scale urban network using coloured Petri net. 98th TRB Annual Meeting: Compendium of Papers. 2019. pp. 19–04885.
14. Yu L. Queuing theory with heavy tails and network traffic modeling. Available at: https://hal.archives-ouvertes.fr/hal-01891760 (accessed: 30.10.2018).
15. Babicheva T.S. The use of queuing theory at research and optimization of traffic on the signal-controlled road intersections. *Procedia Computer Science*. 2015. vol. 55. pp. 469–478.
16. Lin L. et al. Road traffic speed prediction: a probabilistic model fusing multi-source data. *Proceedings of IEEE Transactions on Knowledge and Data Engineering*. 2018. vol. 30. no. 7. pp. 1310–1323.
17. Liu Z., Li Z., Wu K., Li M. Urban Traffic Prediction from Mobility Data Using Deep Learning. *IEEE Network.* 2018. vol. 32. no. 4. pp. 40–46.
18. Wang Y.D. et al. Compression algorithm of road traffic data in time series based on temporal correlation. *IET Intelligent Transport Systems*. 2017. vol. 12. no. 3. pp 177–185.
19. Crawford F., Watling D.P., Connors R.D. A statistical method for estimating predictable differences between daily traffic flow profiles. *Transportation Research Part B: Methodological*. 2017. vol. 95. pp. 196–213.
20. Tchrakian T.T., Basu B., O'Mahony M. Real-time traffic flow forecasting using spectral analysis. *IEEE Transactions on Intelligent Transportation Systems.* 2012. vol. 13. no. 2. pp. 519–526.
21. Addison P. The Illustrated Wavelet Transform Handbook. Boca Raton: CRC Press. 2017. 464 p.
22. Bhattacharyya A., Singh L., Pachori R.B. Fourier–Bessel series expansion based empirical wavelet transform for analysis of non-stationary signals. *Digital Signal Processing*. 2018. vol. 78. pp. 85–196.

23. Yang S., Liu J. Time Series Forecasting based on High-Order Fuzzy Cognitive Maps and Wavelet Transform. *IEEE Transactions on Fuzzy Systems*. 2018. pp. 3391–3402.
24. Wang J., Liu W. Wavelet estimations for heteroscedastic super smooth errors. *Communications in Statistics-Theory and Methods*. 2018. pp. 1–21.
25. Zeng X., Wang J. Wavelet density deconvolution estimations with heteroscedastic measurement errors. *Statistics & Probability Letters.* 2018. vol. 134. pp. 79–85.
26. Cheng Y., Zhang Y., Hu J., Li L. Mining for similarities in urban traffic flow using wavelets. International Conference on Intelligent Transportation Systems Conference (ITSC). 2007. pp. 119–124.
27. Wang J., Wang Z., Li J., Wu J. Multilevel Wavelet Decomposition Network for Interpretable Time Series Analysis. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD' 2018). 2018. pp. 2437–2446.
28. Tian F., Ming W.T., Yun W. Application of Wavelet Fuzzy Neural Network in Real Time Traffic Flow Forecasting. Proceedings of IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC). 2018. pp. 1452–1455.
29. El-Wakeel A.S., Noureldin A., Hassanein H.S., Zorba N. Utilization of Wavelet Packet Sensor De-noising for Accurate Positioning in Intelligent Road Services. Proceedings of the 14th International Wireless Communications & Mobile Computing Conference (IWCMC). 2018. pp. 1231–1236.
30. Xiangxue W., Lunhui X. Wavelet-based short-term forecasting with improved threshold recognition for urban expressway traffic conditions. *IET Intelligent Transport Systems*. 2018. vol. 12. no. 6. pp. 463–473.
31. Zheng Z., Pan L., Pholsena K. Mode Decomposition Based Hybrid Model for Traffic Flow Prediction. Proceedings of IEEE Third International Conference on Data Science in Cyberspace (DSC). 2018. pp. 521–526.
32. Chen X. et al. Kernel sparse representation with hybrid regularization for on-road traffic sensor data imputation. *Sensors*. 2018. vol. 18. no. 9. pp. 2884.
33. Prokhorov S.A. *Prikladnoj analiz nejekvidistantnyh vremennyh rjadov* [Applied analysis of nonuniform time series]. Samara: Samara state aerospace university. 2001. 375 p. (In Russ.).
34. Prokhorov S.A., Stolbova A.A. [A software package for nonuniform time series analysis based on continuous wavelet transformation]. *Programmnye produkty i sistemy. Software & Systems.* 2017. vol. 30. no. 4. pp. 668–671. (In Russ.).
35. Khaymovich A.I., Prokhorov S.A., Stolbova A.A., Kondratyev A.I. A model of milling process based on Morlet wavelets decomposition of vibroacoustic signals. International Conference Information Technology and Nanotechnology (ITNT). 2017. vol. 1904. pp. 135–140.
36. Cannarile F., Baraldi P., Colombo P., Zio E. A Novel Method for Sensor Data Validation based on the analysis of Wavelet Transform Scalograms. *International Journal of Prognostics and Health Management, Prognostics and Health Management Society*. 2018. vol. 9. no. 1. pp.002.
37. Golovnin O.K., Mikheeva T.I. Attribute-driven network-centric urban transport process control system modeling. *Journal of Physics: Conference Series.* 2018. vol. 1096. no 1. pp. 012199.
38. ITSGIS Homepage. Available at: http://www.itsgis.ru (accessed: 30.10.2018).
39. Tönjes R. et al. Real Time IoT Stream Processing and Large-scale Data Analytics for Smart City Applications. European Conference on Networks and Communications, poster session. 2014. 5 p.
40. CityPulse Dataset Collection. Available at: http://iot.ee.surrey.ac.uk:8080/datasets.html (accessed: 30.10.2018).

УДК 656, 004.8, 007.5, 51-74, 510.67 **DOI** 10.15622/sp.18.2.354-389

Я.А. СЕЛИВЕРСТОВ, В.И. ЧИГУР, А.М. САЗАНОВ, С.А. СЕЛИВЕРСТОВ, А.С. СВИСТУНОВА

РАЗРАБОТКА СИСТЕМЫ ДЛЯ ТОНОВОГО АНАЛИЗА ОТЗЫВОВ ПОЛЬЗОВАТЕЛЕЙ ПОРТАЛА «AUTOSTRADA.INFO/RU»

Селиверстов Я.А., Чигур В.И., Сазанов А.М, Селиверстов С.А., Свистунова А.С. **Разработка системы для тонового анализа отзывов пользователей портала «AUTOSTRADA.INFO/RU».**

Аннотация. Социальные сети (Вконтакте, Facebook), тематические сообщества в сетях микроблогинга (Twitter), ресурсы для путешественников (TripAdvisor) и транспортные порталы (Autostrada) являются источником актуальной и оперативной информации о дорожно-транспортной обстановке, качестве предоставляемых транспортных услуг и степени удовлетворенности пассажиров уровнем транспортного обслуживания. Однако существующие системы транспортного мониторинга не содержат программных инструментов, способных осуществлять сбор и анализ дорожно-транспортной информации в среде Интернет. В настоящей работе рассматривается задача построения системы автоматического извлечения и классификации дорожно-транспортной информации с транспортных интернет-порталов и апробация разработанной системы для анализа транспортных сетей Крыма и города Севастополя. Для решения этой задачи проанализированы библиотеки с открытым исходным кодом для тематического сбора и исследования данных. Разработан алгоритм для извлечения и анализа текстов. Осуществлена разработка краулера с использованием пакета Scrapy на языке Python3 и собраны отзывы пользователей с портала http://autostrada.info/ru о состоянии транспортной системы Крыма и города Севастополя. Для лемматизации текстов и векторного преобразования текстов были рассмотрены методы tf, idf, tf-idf и их реализация в библиотеке Scikit-Learn: CountVectorizer и TF-IDF Vectorizer. Для обработки текстов были рассмотрены методы Bag-of-Words и n-gram. В ходе разработки модели классификатора рассмотрены наивный байесовский алгоритм (MultinomialNB) и модель линейного классификатора с оптимизацией стохастического градиентного спуска (SGDClassifier). В качестве обучающей выборки использовался корпус объемом 225 тысяч размеченных текстов с ресурса Twitter. Проведено обучение классификатора, в ходе которого использовалась стратегия кросс-валидации и метод ShuffleSplit. Проведено тестирование и сравнение результатов тоновой классификации. По результатам валидации лучшей оказалась линейная модель со схемой n-грамм [1, 3] и векторизатором TF-IDF. В ходе апробации разработанной системы был проведен сбор и анализ отзывов, относящихся к качеству транспортных сетей республики Крым и города Севастополя. Сделаны выводы и определены перспективы дальнейшего функционального развития разрабатываемого инструментария.

Ключевые слова: автоматический анализ текстов, краулеры, классификация текстов, интеллектуальные транспортные системы, машинное обучение, TF-IDF, наивный байесовский алгоритм, линейный классификатор, анализ тональности.

1. Введение. Стремительное развитие мобильных и облачных технологий, перевод логистической, потребительской, коммуникационной и расчетно-денежной деятельностей в

информационно-сетевое пространство открывает новые пути развития интеллектуальных транспортных систем (ИТС) и систем транспортного мониторинга.

Работа современных ИТС [1, 2] строится на данных, получаемых с систем видео-мониторинга [3], а также информации о местоположении пользователей и транспортных средств [4], которая собирается с использованием мобильных устройств, поддерживающих GPS/WiFi/Lte/WPAN стандарты передачи данных [5, 6].

Стоимость систем видео-мониторинга сравнительно высока, поэтому их размещают только на особо загруженных участках улично-дорожных сетей крупных городов и мегаполисов. Улично-дорожные сети небольших городов и поселков, а также региональные и областные транспортные сети остаются не охваченными системами видео-мониторинга, и как следствие, информация о дорожно-транспортной обстановке на них отсутствует.

Мобильные телефоны с доступом в Интернет имеются, как правило, у каждого водителя. В случае обнаружения проблемных участков на дороге или дорожно-транспортных происшествий водитель способен зафиксировать эту информацию в виде отзыва на специализированном интернет-портале.

Таким образом, одним из источников разнородной информации, относящейся к сфере транспорта, может быть web-пространство.

Транспортные данные в web-пространстве, как правило, структурированы и разбросаны по тематическим интернет-ресурсам.

К таковым относят: отраслевые сайты (http://autostrada.info/ru), тематические интернет сообщества (https://www.worldoftrucks.com/en/), группы в социальных сетях (Вконтакте, Facebook) и сетях микроблогинга (Twitter), а также чаты и форумы.

Информация на транспортных web-порталах и тематических интернет-сообществах формируется в виде отзывов непосредственно самими пользователями, поэтому для ее сбора не требуется больших затрат. Тема web-портала или интернет сообщества определяет характер размещаемой информации, например, если тематика группы «пробки», то, как правило, размещаемые пользователями отзывы содержат сведения о пробках и заторах на дорогах. Если же тематика группы «поборы на дорогах», то размещаемые пользователями отзывы содержат сведения о недобросовестной работе сотрудников весового контроля или служителей правопорядка.

Таким образом, каждой теме может быть поставлена в соответствие некоторая характеристика или фактор, оказывающий влияние на транспортный процесс и дорожные условия. Такое

структурирование информации упрощает процесс составления тематических корпусов в области транспорта, что, в свою очередь, позволяет строить более глубокие системы классификации транспортных данных и выявлять на их основе новые управляющие воздействия.

Таким образом, использование систем извлечения и анализа дорожно-транспортной информации из web-пространства в качестве систем транспортного мониторинга [7-9] открывает новые каналы поступления транспортной информации, способной повысить информированность участников дорожного движения о состоянии транспортных сетей и условий дорожного движения.

2. Анализ предметной области. Проанализируем последние публикации, в которых рассматриваются методы извлечения и анализа текстов, относящихся к транспортной сфере.

В работе [10] рассматриваются методы построения автоматического классификатора для анализа вопросов и ответов путешественников на сайте TripAdvisor в разделе Q-A форума «О поиске оптимальных маршрутов в разных городах». Целью данного исследования являлось построение вопросно-ответной системы. В работе [11] представлена методология сбора данных из социальной сети Twitter с использованием данных геотегированной службы Twitter. Собранные данные содержат информацию о моделях мобильности и поведенческих характеристиках пользователей. В работе [12] из сообщений социальной сети Twitter вычленяется информация о качестве транспортного обслуживания пассажиров городского транспорта, пробках, сбоях в расписании, минировании вокзалов и транспортных средств, авариях и ДТП — данная информация использовалась для повышения качества транспортного обслуживания в период чемпионата мира по футболу. В работе [13] разрабатывается программный модуль для ИТС на основе нечеткой онтологии, на основе правил логического вывода, семантического анализа и анализа тональностей текстов из социальных сетей (Twitter, Facebook), городских транспортных порталов, сайтов для путешественников и туристов. Данная система эффективно извлекает отзывы и сообщения, связанные с особенностями городской среды (например, автобусные и железнодорожные вокзалы, мосты, парки, рестораны, аэропорты, медицинские центры и гостиницы) и транспортными инцидентами (например, столкновениями, плохими дорогами, скоплениями и пробками). В работе [14] разрабатывается метод автоматической обработки информации о путешествиях из туристических блогов, которые

определены как туристические журналы, написанные блоггерами в дневниковой форме. В работе показано, что блоги о путешествиях — это полезный источник транспортной информации для туристов. В работе [15] разрабатываются эффективные методы сбора, извлечения и передачи данных из социальных сетей (Twitter) для информационного обеспечения интеллектуальных систем управления дорожным движением, продвинутых систем поддержки путешественников (support advanced traveler information systems) и транспортных диспетчерских центров. В работе [16] исследуется вклад семантического и семантико-синтаксического видов анализа на эффективность решения прикладных задач обработки текстов: вопросно-ответного поиска и извлечения определений из научных публикаций в области транспорта. В работе [17] рассматриваются методы, используемые для обнаружения экстремистских текстов из Интернета. В статье [18] рассматривается система для сбора, обработки и фильтрации сообщений пользователей, связанных с дорожно-транспортными происшествиями и авариями. Сообщения получены из Twitter с использованием REST API в режиме реального времени. В процессе адаптивного сбора данных формируется корпус текстов, состоящий из важных ключевых слов и их комбинаций, которые могут подразумевать дорожные происшествия. Затем сообщение преобразуется в бинарный вектор в пространстве признаков и классифицируется как относящийся к транспортному происшествию или нет. Все сообщения, относящиеся к транспортному происшествию, геокодируются для определения их местоположения и далее классифицируются в одну из пяти категорий инцидентов. Данная система была успешно протестирована в двух регионах: Питтсбурге и Филадельфии. В работе [19] разрабатывается система для автоматического сбора актуальной и ценной информации из Twitter, связанной с транспортом и транспортными услугами, во время проведения футбольных матчей. В работе успешно протестированы методы автоматического сбора и идентификации сообщений на выборках свыше 3.7 миллионов сообщений. В работе [20] анализируются общественные системы обмена велосипедами в Испании через анализ настроений в социальных сетях с учетом мнений жителей и посетителей, также выявляются положительные и отрицательные факторы и определяется их потенциальное влияние на качество туристических и транспортных услуг. В процессе анализа были обработаны данные из 46 систем обмена велосипедами за период 2010-2016 годов. Результаты исследования показывают, что

инфраструктура туризма и транспорта, включающая велосипедные дорожки, станции и обеспечение eBike, нуждается в скоординированном планировании, поскольку она неразрывно связана с уровнем туристического и транспортного обслуживания. Социальные сети в рамках данного исследования показали себя как достоверный источник транспортной информации. В статье [21] рассматривается система сбора и анализа отзывов туристов и путешественников с сайта TripAdvisor.com с учетом их географического местоположения, а также статистических и территориальных данных. Методы анализа текстов применяются для оценки восприятия туристами положительных факторов (мест, событий, достопримечательностей), которые могут использоваться в качестве инструментов поддержки планирования поездок. Исследование проведено на примере Хорватии. Результаты исследования раскрывают ценность и взаимодополняемость данных, связанных с социальными сетями, с официальной статистикой планирования транспорта и туризма.

Анализ предметной области показал, что передовые системы для извлечения и анализа тематических текстов активно внедряются в системы городского транспортного мониторинга и системы поддержки туристической и транспортной мобильности.

3. Постановка задачи. Целью настоящей статьи является разработка и тестирование системы, способной анализировать тексты с транспортных web-порталов. В качестве интернет-ресурса выбран web-портал http://autostrada.info/ru, в качестве методов контент-анализа — анализ тональности; в качестве шкалы классификации полярности документа — бинарная шкала с двумя классами оценок: положительные и отрицательные; в качестве оцениваемого объекта выбраны транспортные сети республики Крым и города Севастополь.

Предполагается выполнить следующий перечень работ:

1) Разработать схему алгоритма для извлечения и анализа текстов.

2) Программно реализовать функционал алгоритма для сбора текстов по дорожно-транспортной проблематике.

3) Протестировать разработанную программу и собрать тексты с сайта http://autostrada.info/ru.

4) Сформировать корпуса текстов для последующего обучения классификатора.

5) Разработать тоновый классификатор.

6) Обучить классификатор и оценить его работу.

4. Анализ фреймворков для получения данных из Web. Одна из главных задач тематического веб-краулера — поиск и добавление в коллекцию документов наиболее значимых информационных источников, что обеспечивает создание коллекции высокого качества [22]. Уже существует широкий ассортимент известных библиотек. Это позволяет не писать с нуля новые поисковые боты [23].

Анализ библиотек с открытым исходным кодом из списка TOP-50 определил наиболее функциональные для нашей системы фреймворки.

На основе анализа, представленного в таблице 1, под такие характеристики подходит фреймворк Scrapy.

Scrapy — одна из наиболее популярных и производительных библиотек Python для получения данных с веб-страниц. Фреймворк Scrapy является сфокусированным, легко устанавливается, поддерживает выгрузку данных в форматах JSON, XML, CSV.

Таблица 1. Анализ библиотек (фреймворков) для получения данных из web

Название	Описание	Источник
Heritrix	Гибкий, расширяемый, надежный и масштабируемый фреймворк, написанный на Java и способный получать, архивировать и анализировать тексты. Heretrix работает в распределенной среде с помощью хеширования URL хостов.	[18, 19]
Nutch	Представляет собой инкрементный, параллельный, распределенный, кроссплатформенный модульный фреймворк для построения поисковых систем, написанный на java. Поддерживает граф связей узлов, различные фильтры и нормализацию URL.	
Scrapy	Расширяемый, сфокусированный, параллельный, кроссплатформенный и гибкий фреймворк-библиотека для Python. Легко устанавливается, поддерживает выгрузку данных в форматах JSON, XML, CSV. Широко используется для веб-скрайбинга, не имеет встроенных функций для работы в распределенной среде.	

5. Разработка алгоритма для извлечения и анализа тематических текстов. Построение системы для извлечения и анализа тематических текстов начинается с разработки обобщенного алгоритма.

Алгоритм в общем виде состоит из процедур, представленных в таблице 2, а схема алгоритма представлена на рисунке 1.

Таблица 2. Общий вид алгоритма для извлечения и анализа тематических текстов

Процедура алгоритма	Наименование процедуры
1	Формирование очереди ссылок, подаваемых на вход краулера
2	Список источников добавляются в очередь обхода краулера
3	Клаулер сканирует страницу из очереди
4	Краулер скачивает интересующий его веб-документ в базу данных
5	Проводится очистка веб-документа от «мусора»
6	Производится сохранение очищенного текста в базу данных
7	Подготовка коллекций, ручная разметка текстов и построение корпуса театических текстом
8	Запуск классификатора тональности
9	Обучение классификатора на различных корпусах текстов
10	Оценка классификатора тональности

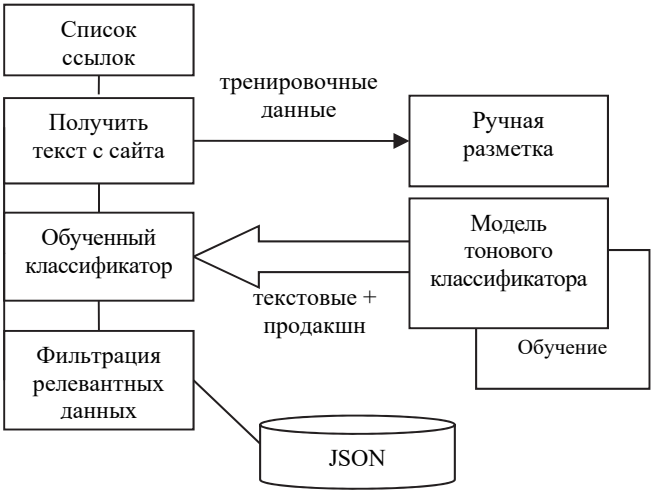


Рис. 1. Схема системы для извлечения и анализа тематических текстов

6. Разработка краулер-модуля. На первом этапе исследования разрабатывался краулер-модуль. Краулер-модуль выполняет процедуры 1-4 алгоритма (таблица 2), а именно: 1) формирует очередь ссылок; 2) добавляет список источников в очередь обхода;

3) сканирует страницу из очереди; 4) скачивает интересующий его web-документ в базу данных.

Часть листинга программы краулер-модуля представлена на листинге 1.

```
importscrapy
classRoadSpider(scrapy.Spider):
name = 'road_spider'
start_urls = [
'http://autostrada.info/ru/reviews/page/1/',
]
def parse(self, response):
for review in response.css('div.col-md-12.reviewBlock'):
        tmp = review.css('p.comment.break-word::text').extract_first()
        tmp1 = review.css('a.label.label-code::text').extract_first()
        tmp2 = review.css('a.highwayLabel::text').extract_first()
        tmp = tmp.replace('\r\n', ' ')
        tmp = tmp.replace('\n', '')
        dd = {
                'title': tmp1 + ' ' + tmp2,
                        'subtitle': review.css('div.col-sm-8.b-rate.hidden-xs
                    b::text').extract_first(),
                'date': review.css('strong.reviewDate::text').extract_first(),
                'rate': review.css('span.b-stars::attr(title)').extract_first(),
      'description': tmp,
                }
        try:
                        dd['date'] = dd['date'].replace('\t', '')
                dd['date'] = dd['date'].replace('\n', '')
                dd['date'] = dd['date'].replace('\u0433.', '')
        except:
                pass
        yielddd
```

Листинг 1. Часть программы краулер-модуля

В процессе работы краулера с сайта http://autostrada.info/ru извлекаются мнения пользователей в текстовом виде.

В результате работы краулер-модуля был собран корпус, содержащий 1130 текстов за период с 01 марта 2009 года по 1 ноября 2018 года с сайта http://autostrada.info/ru. Рассмотрим несколько примеров текстов корпуса и того, что в них содержится.

На рисунке 2 представлен пример отзыва с сайта http://autostrada.info/ru о состоянии участка трассы, пролегающий между Феодосией и Керчью.

Рис. 2. Отзыв о состоянии дорог на сайте http://autostrada.info/ru

Структура отзывов на сайте autostrada представлена на рисунке 3.

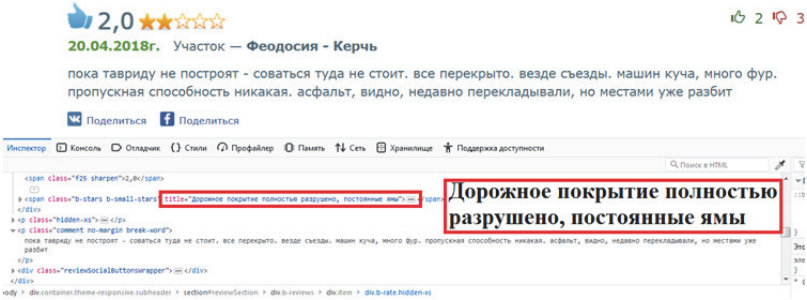


Рис. 3. Структура отзыва на сайте http://autostrada.info/ru

Извлеченный текст записывается в базу данных с указанием атрибутов: date (дата и время создания отзыва), description (описание ситуации), subtitle (наименование трасы), title (кодифицированное название трассы) и url (адрес отзыва в интернет).

Например, для отзыва, представленного на рисунке 4, атрибуты имеют вид: *date: «14.07.2018 14:18»; description: «Асфальт новый, но уже много участков с колеёй. Дорожные строители параллельно, рядом, строят магистраль «Таврида», поэтому много объездов строящихся мостов и развязок, а так-же участков с ограничением скорости. В добавок, временными ограждениями проезжая часть заужена так, что со встречной машиной разъёзжаешься «впритирку». И ограждениями-же закрыт доступ к обочинам. Поэтому, если кто остановился в потоке – сразу пробка, ведь дорога очень перегружена, в том числе и самосвалами и техникой строителей дороги»; Subtitle: «Феодосия-Керчь»; title: «М-17 Херсон – Джанкой – Керчь».*

```
{ 'date': '14.07.2018 14:18 ' ,
  'description': 'Асфальт новый, но уже много участков с колеёй. Дорожные строители  '
                 ' параллельно, рядом, строят магистраль "Таврида", поэтому много объездов  '
                 ' строящихся мостов и развязок, а так-же участков с ограничением скорости. В '
                 'добавок, временными ограждениями проезжая часть заужена так, что со '
                 'встречной машиной разъёзжаешься "впритирку". И ограждениями-же закрыт '
                 'доступ к обочинам. Поэтому, если кто остановился в потоке -- сразу пробка, ведь '
                 'дорога очень перегружена, в том числе и самосвалами и техникой строителей '
                 'дороги. ' ,
  'subtitle': 'Феодосия - Керчь',
  'title': 'М-17 Херсон – Джанкой – Керчь',
  'rate': 'Дорожное покрытие полностью разрушено, постоянные ямы' ,
  'url': http://autostrada.info/ua/highway/M-17 '}
```

Рис. 4. База данных с текстами по дорожно-транспортной проблематике

Далее осуществляются процедуры 5 и 6 алгоритма (см. таблица 2). Все собранные краулером отзывы группируются в единый текст и подвергаются процедуре предобработки: слова приводятся к нижнему регистру, затем отсеиваются все вспомогательные символы, такие как знаки препинания и стоп-слова.

Далее с помощью библиотеки pymorhy2 слова приводятся к нормальной форме. На следующем этапе осуществляется векторизация [24] и производится лексический анализ текста.

7. Лексический анализ и векторизация текста. Перед тем как использовать машинное обучение на текстовых документах необходимо перевести текстовое содержимое в числовой вектор признаков с учетом tf, idf и tf-idf [25, 26]. Векторизатор строит словарь индексов признаков.

В более сложных моделях [27] используют алгоритмы семантико-синтаксического анализа [28] и токенизации [29] с возможностью настройки тонового анализа [30].

Для обработки текстов целесообразно использовать два метода: CountVectorizer и TFIDFVectorizer [31]. Ниже будет дано обоснование использования данных методов. Оба метода используют модель Bag of Words [32]. Листинг программы, выполняющей лексический анализ, векторизацию и индексирование текста, представлен на листинге 2.

```
fromsklearn.feature_extraction.text import CountVectorizer
fromsklearn.feature_extraction.text import TfidfVectorizer
fromsklearn.grid_search import GridSearchCV
fromsklearn.cross_validation import ShuffleSplit, cross_val_score
importpandasaspd
# считываемподготовленныйдатасет
        dataset = pd.read_csv('data/cleaned_data.csv', index_col=0).dropna()
```

```
# массив n-граммных схем, которые будут использоваться в работе
# например, (1, 3) означает униграммы + биграммы + триграммы
ngram_schemes = [(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6)]
forngram_scheme in ngram_schemes:
print('N-gram Scheme:', ngram_scheme)
count_vectorizer = CountVectorizer(analyzer = "word", ngram_range=ngram_scheme)
tfidf_vectorizer = TfidfVectorizer(analyzer = "word", ngram_range=ngram_scheme)
vectorizers = [count_vectorizer, tfidf_vectorizer]
vectorizers_names = ['Count Vectorizer', 'TF-IDF Vectorizer']
for i in range(len(vectorizers)):
        print(vectorizers_names[i])
        vectorizer = vectorizers[i]
        X = vectorizer.fit_transform(dataset['text'])
        y = dataset['label']
        cv = ShuffleSplit(len(y), n_iter=5, test_size=0.3, random_state=0)
```

Листинг 2. Лексический анализ и векторизация текста

Рассмотрим более подробно используемые выше методы и обоснуем их выбор.

7.1. Метод Bag of Words. Математическая модель Bag of Words (перевод с англ. — мешок слов) — это модель обработки текста, при котором слова выбираются в случайном порядке.

Модель Bag of Words [33] позволяет перейти к компактному представлению документа, в котором любое слово $w_t \in V$ словаря V в документе d_i имеет количество вхождений равное n_t, следовательно, любой документ d_i может быть представлен вектором в виде [32]:

$$\overline{d}_i = \left(n_1\left(w_1\right) + n_{\ t}\left(w_t\right) + \cdots + n_m\left(w_m\right)\right), \tag{1}$$

где m — количество слов в документе d_i.

Как правило, выделяют два основных типа атрибутов:

1. Частотные атрибуты — когда каждое значение в векторе $\overline{d}$ соответствует количеству вхождений признаков (слов) в документ d; тогда $n_i(d) \in (0;+\infty)$;

2. Бинарные (наличия/отсутствия) атрибуты — когда каждое значение в векторе $\overline{d}$ бинарное (true/false или 0/1) и отражает факт присутствия признака w_i в документе d, тогда $n_i(d) \in \{0;1\}$.

Алгоритм построения модели следующий: 1) составляется словарь терминов из всех слов, встречающихся в тексте, при этом из

текста предварительно исключаются все знаки препинания, числа и «стоп-слова»; 2) для каждого документа определяется вектор, каждая компонента которого соответствует термину из словаря, а ее значение определяется числом, характеризующим сколько раз это слово встретилось в тексте. Размерность вектора соответствует мощности словаря.

Такой подход довольно распространен и прост в реализации, но он не избавлен от недостатков. Например, отзыв «трасса не очень хорошая» имеет негативную тональность, однако, если рассматривать каждое слово по отдельности, невозможно будет это определить. Кроме того, модель, вероятно, «выучит», что слово «хороший» имеет положительную тональность, но в данном случае это не то, что требуется.

Проблема определения смыслового окраса текста может быть решена с помощью метода n-gram. Обычно для таких задач используют схемы с униграмными, биграмными или тригамными признаками и их совместные комбинации независимо друг от друга.

7.2. Метод *n*-gram. Математическая модель n-gram — это модель представления текстов в виде набора последовательностей, состоящих из N слов. Различают следующие модели n-gram: 1 слово — униграммы, при которой определяется вероятность $P(w_i)$ появления i-го слова (w_i) в тексте; 2 слова — биграммы, при которой определяется вероятность появления пар слов $P(w_i|w_{i-1})$ в тексте, 3 слова — триграммы, при которой определяется вероятность появления троек слов $P(w_i|w_{i-2}, w_{i-1})$ в тексте [34, 35].

Таким образом, задача сводится к определению вероятности появления цепочки слов $V_m = (w_1, w_2, \ldots, w_t)$ в некотором тексте d_m.

Вероятность $P(w_1, w_2, \ldots, w_t)$ можно представить в виде произведения условных вероятностей входящих в нее n-gram [27]:

$$P(w_1, w_2, \ldots, w_t) = \prod_{i=1}^{t} P(w_i | w_1, w_2, \ldots, w_{i-1}), \quad (2)$$

или аппроксимируя $P(w)$ при ограниченном контексте длинной (n-l), согласно [34]:

$$P(w_1, w_2, \ldots, w_t) \cong \prod_{i=1}^{t} P(w_i | w_{i-n+1}, w_{i-n+2}, \ldots, w_{i-1}). \quad (3)$$

Вероятность появления n-gram вычисляется согласно [34]:

$$P\left(w_i \middle| w_{i-n+1},\ldots,w_{i-1}\right) = \frac{C\left(w_{i-n+1},\ldots,w_i\right)}{C\left(w_{i-n+1},\ldots,w_{i-1}\right)}, \quad (4)$$

где C — количество появлений последовательности слов в обучающем корпусе.

В процессе подбора лучших параметров для модели рассматриваются схемы n-gram от 1 до 5. Как правило, в длинных документах среднее количество словоупотреблений будет выше, чем в коротких, даже если они посвящены одной теме. Чтобы избежать этих потенциальных несоответствий, достаточно разделить количество употреблений каждого слова в документе на общее количество слов. Этот признак называется «частота термина» или Term Frequency [29].

7.3. Мера Term Frequency. Частота термина — отношение числа вхождений некоторого слова к общему числу слов документа, при которой оценивается важность слова w_i в пределах отдельного документа [25]:

$$tf\left(w_i;d\right) = \frac{n_t}{\sum_i n_i}. \quad (5)$$

Следующим уточняющим параметром словоупотреблений является мера обратной частоты документа (Inverse Document Frequency).

7.4. Мера Inverse Document Frequency. Обратная частота документа **idf** — инверсия частоты, с которой некоторое слово встречается в документах коллекции [25]:

$$idf\left(w_i;D\right) = \log\frac{D}{\left|\left\{d_i \in D \middle| w_i \in d_i\right\}\right|}, \quad (6)$$

где $\left|D\right|$ — число документов в коллекции; $\left|\left\{d_i \in D \middle| w_i \in d_i\right\}\right|$ — число документов из коллекции D в которой встречается слово w_i.

Использование меры idf позволяет снизить вес широкоупотребительных слов, которые являются менее информативным, чем те, которые используются только в небольшой части.

Примером низко информативных слов могут служить служебные слова, артикли, предлоги, союзы.

На последнем этапе вычисляется ключевая характеристика tf-idf, определяющая перечень уникальных слов однозначно определяющих данный документ.

7.5. Мера Term Frequency-Inverse Document Frequency. Частота термина-обратная частота документа $tf-idf$ — статистическая мера, которая используется для оценки важности слова в контексте документа, являющегося частью коллекции документов или корпуса [32]:

$$tf-idf = tf_{i,j} \ln\left(\frac{N}{df_i}\right), \tag{7}$$

где $tf_{i,j}$ — отношение количества вхождений слова к общему числу терминов документа, df_i — число документов из коллекции, в которых встречается слово, N — число документов в коллекции.

Таким образом, вес некоторого слова пропорционален количеству употреблений этого слова в документе и обратно пропорционален частоте употребления слова в других документах коллекции. Следовательно, если слово часто встречается в каком-либо документе и при этом встречается редко в других документах, то это слово имеет большую значимость для анализируемого документа.

Следовательно, именно меру $tf-idf$ целесообразно использовать в разработке системы извлечения и анализа тематических текстов. В программном исполнении $tf-idf$ реализован в библиотеке Scikit-Learn в виде стандартного метода векторизатора TF-IDF Vectorizer [31].

Рассмотрев основные методы анализа текстов, перейдем к построению классификатора тональности.

8. Разработка тонового классификатора. Для построения модели тонового классификатора рассмотрим и сравним две наиболее используемые модели классификации: наивный байесовский классификатор и линейный классификатор на основе стахостического градиента. Как известно, существуют и другие часто используемые методы, которые являются эффективными для различных задач классификации, например метод опорных векторов (SVM) или сверточные нейронные сети (CNN). Данные методы будут рассмотрены в рамках следующих исследований.

В программном исполнении наивный байесовский классификатор реализован в библиотеке Scikit-Learn в виде

стандартного метода MultinomialNB, а линейный классификатор на основе стахостического градиента в виде — SGDClassifier [31].

Листинг программы тонового классификатора на основе стандартных методов MultinomialNB и SGDClassifier классификаторов представлен на листинге 3.

```
# Наивный байес
clf = MultinomialNB()
NB_result = cross_val_score(clf, X, y, cv=cv).mean()
# Линейный классификатор
clf = SGDClassifier()
parameters = {
'loss': ('log', 'hinge'),
'penalty': ['none', 'l1', 'l2', 'elasticnet'],
'alpha': [0.001, 0.0001, 0.00001, 0.000001]
}
gs_clf = GridSearchCV(clf, parameters, cv=cv, n_jobs=-1)
gs_clf = gs_clf.fit(X, y)
L_result = gs_clf.best_score_
```

Листинг 3. Листинг программы тонового классификатора

Рассмотрим более подробно модели отобранных классификаторов и обоснуем их выбор.

8.1. Наивный байесовский классификатор. Существуют два подхода к наивному байесовскому классификатору — мультиномиальный и многомерный, которые дают разные результаты.

Определим, какой из подходов лучше использовать для классификации текстов в данном случае [29].

Многомерная модель (https://docplayer.ru/45424867-Naivnyy-bayesovskiy-klassifikator.html): пусть $V = \{w_t\}_{t=1}^{|V|}$ — словарь; тогда документ d_i — это вектор длины $|V|$, состоящий из битов B_{it}; $B_{it} = 1$, если слово w_t встречается в документе d_i.

Правдоподобие принадлежности d_i классу c_j рассчитывается согласно [36]:

$$p\left(d_i \middle| c_j\right) = \prod_{t=1}^{|V|}\left(B_{it}\, p\left(w_t \middle| c_j\right) + \left(1 - B_{it}\right)\left(1 - p\left(w_t \middle| c_j\right)\right)\right). \quad (8)$$

Для обучения такого классификатора нужно получить вероятности $p\left(w_t \middle| c_j\right)$. Рассмотрим процесс обучения.

Обучение: пусть дан набор документов $D = \{d_i\}_{i=1}^{|V|}$, которые распределены по классам c_j, дан словарь $V = \{w_t\}_{t=1}^{|V|}$ и заданы биты документов B_{it}.

Тогда можно подсчитать оценки вероятностей того, что-то или иное слово встречается в том или ином классе [36]:

$$p(w_t|c_j) = \frac{1 + \sum_{i=1}^{|D|} B_{it} p(c_j|d_i)}{2 + \sum_{i=1}^{|D|} p(c_j|d_i)}. \quad (9)$$

Априорные вероятности классов рассчитываются в соответствии с [36]:

$$p(c_j) = \frac{1}{|D|} \sum_{i=1}^{|D|} p(c_j|d_j). \quad (10)$$

Тогда классификация будет происходить в соответствии с [36]:

$$c = \arg\max_j p(c_j) p(d_j|c_j) =$$

$$= \arg\max_j \left(\frac{1}{|D|} \sum_{i=1}^{|D|} p(c_j|d_i) \right) \prod_{t=1}^{|V|} \left(B_{it} p(w_t|c_j) + (1 - B_{it})(1 - p(w_t|c_j)) \right) = \quad (11)$$

$$= \arg\max_j \left(\log \left(\sum_{i=1}^{|D|} p(c_j|d_i) \right) + \sum_{t=1}^{|V|} \log \left(B_{it} p(w_t|c_j) + (1 - B_{it})(1 - p(w_t|c_j)) \right) \right).$$

В мультиномиальной модели документ — это последовательность слов, отобранных методом «Bag of Words» [37]. Данный метод был рассмотрен выше. Для подсчета правдоподобия документа требуется перемножить вероятности того, что из «мешка» были «вытащены» те самые слова, которые встретились в документе.

Наивное предположение заключается в том, что из «мешка» «вытаскиваются» разные слова независимо друг от друга.

Мультиномиальная модель: пусть $V = \{w_t\}_{t=1}^{|V|}$ — словарь, тогда документ d_i — это вектор длины $|d_i|$, состоящий из слов, каждое из которых «вынуто» из словаря с вероятностью $p(w_t|c_j)$.

Правдоподобие принадлежности d_i классу c_j имеет вид [36]:

$$p\left(d_i\middle|c_j\right)=p\left(\left|d_i\right|\right)\left|d_i\right|!\prod_{t=1}^{|V|}\frac{1}{N_{it}!}p\left(w_t\middle|c_j\right)^{N_{it}}, \tag{12}$$

где N_{it} — количество вхождений w_t в d_i.

Для обучения такого классификатора требуется определить вероятности $p\left(w_t\middle|c_j\right)$. Далее рассматривается процесс обучения.

Обучение: пусть дан набор документов $D=\left\{d_i\right\}_{i=1}^{|D|}$, которые уже распределены по классам c_j (возможно, даже вероятностно распределены), дан словарь $V=\left\{w_t\right\}_{t=1}^{|V|}$, и значение вхождения N_{it} известно.

Тогда, в соответствии с (13), можно подсчитать оптимальные оценки вероятностей, что то или иное слово встречается в том или ином классе [36]:

$$p\left(w_t\middle|c_j\right)=\frac{1+\sum_{i=1}^{|D|}N_{it}p\left(c_j\middle|d_i\right)}{|V|+\sum_{s=1}^{|V|}\sum_{i=1}^{|D|}N_{is}p\left(c_j\middle|d_i\right)}. \tag{13}$$

Априорные вероятности классов рассчитываются согласно [36]:

$$p\left(c_j\right)=\frac{1}{|D|}\sum_{i=1}^{|D|}p\left(c_j\middle|d_i\right). \tag{14}$$

Тогда классификация будет происходить в соответствии с [36]:

$$\begin{aligned} c&=\arg\max_j p\left(c_j\right)p\left(d_j\middle|c_j\right)= \\ &=\arg\max_j\left(\frac{1}{|D|}\sum_{i=1}^{|D|}p\left(c_j\middle|d_i\right)\right)\prod_{t=1}^{|V|}\frac{1}{N_{it}!}p\left(w_t\middle|c_j\right)^{N_{it}}= \\ &=\arg\max_j\left(\log\left(\sum_{i=1}^{|D|}p\left(c_j\middle|d_i\right)\right)+\sum_{t=1}^{|V|}N_{it}\log p\left(w_t\middle|c_j\right)\right). \end{aligned} \tag{15}$$

Многомерная модель дает лучшую оценку предсказания на текстах с объемом, не превышающем 100 слов. Когда размер текстов составляет несколько тысяч слов, то лучшие результаты дает мультиномиальная модель. Таким образом, выбор MultinomialNB классификатора является обоснованным. Далее приведено его сравнение с линейным SGDClassifier классификатором.

8.2. Линейный классификатор с обучением стохастического градиентного спуска. Основная идея линейного классификатора заключается в том, что признаковое пространство может быть разделено гиперплоскостью на две полуплоскости, в каждой из которых прогнозируется одно из двух значений целевого класса.

Пусть вектор $\vec{x}$ представляет собой входные данные, а на выходе классификатора вычисляется показатель $\vec{y}$ по формуле [38]:

$$y = f\left(\vec{w} \cdot \vec{x}\right) = f\left(\sum_i w_i x_i\right), \tag{16}$$

где $\vec{x}$ — нормализованный вектор из частот слов в документе; $\vec{w}$ — действительный вектор весов той же размерности, что и признаковое пространство; f — функция преобразования скалярного произведения.

В ряде случаев задачи текстовой классификации, включающие в себя более одного класса, сводятся к нескольким задачам бинарной классификации [37]. В этом случае метки целевого класса $D_C \subset D$ обозначаются «1» (положительные примеры), а нецелевого «-1» (отрицательные примеры), а функция принадлежности $\mathrm{y}:\mathrm{D}\to\{\mathrm{l};-\mathrm{l}\}$ может быть представлена в виде [39]:

$$y = \begin{cases} +1, & x \in D_C \\ -1, & x \notin D_C \end{cases}. \tag{17}$$

Естественной интерпретацией для y в дискретном случае будет разделяющая гиперплоскость между различными классами. Для ускорения этого метода используется метод стохастического градиентного спуска: на каждой итерации спуск осуществляется с учетом одного случайно выбранного документа $d \in D$.

Значения весов вектора $\vec{w}$ определяются в ходе обучения на тестовых выборках [39].

Обучение: пусть $y^*: X \to Y$ — целевая зависимость, известная только на объектах обучающей выборки $X^l = \left(x_i, y_i\right)_{i=1}^l, y_i = y^*\left(x_i\right)$.

Требуется найти вектор весов $\boldsymbol{w}$, при котором алгоритм $a(x,w)$ аппроксимирует целевую зависимость $y^*(x_i)$.

Подобная задача сводится к поиску вектора $\boldsymbol{w}$ и доставляющего минимум функционалу [38]:

$$Q(w)=\sum_{i=1} L\left(a(x_i,w),y_i\right)\to \min_w, \tag{18}$$

где $L(a,y)$ — заданная функция потерь, характеризующая величину ошибки ответа $a(x,w)$ при правильном ответе y.

Применение для минимизации $Q(w)$ метод градиентного спуска. В этом методе выбирается некоторое начальное приближение для вектора весов w, затем запускается итерационный процесс, на каждом шаге которого вектор w изменяется в направлении и наиболее быстрого убывания функционала Q.

Это направление противоположно вектору градиента [38]:

$$\nabla Q(w)=\left(\frac{\partial Q(w)}{\partial w_j}\right)_{j=1}^{n};\ w:=w-\eta\nabla Q(w), \tag{19}$$

где $\eta>0$ — величина шага в направлении антиградиента, называемая также темпом обучения (learning rate).

Предполагая, что функция потерь L и функция активации f дифференцируемы, распишем градиент в виде [38]:

$$w:=w-\eta\sum_{i=1} L'_a\left(a(x_i,w),y_i\right)f'\left(\langle w,x_i\rangle\right)x_i. \tag{20}$$

Каждый прецедент (x_i,y_i) вносит аддитивный вклад в изменение вектора w, но вектор w изменяется только после перебора всех l объектов.

Для повышения скорости сходимости данного процесса прецеденты выбираются случайным образом (x_i,y_i), для каждого делается градиентный шаг и сразу обновляется вектор весов [38]:

$$w:=w-\eta L'_a\left(a(x_i,w),y_i\right)f'\left(\langle w,x_i\rangle\right)x_i. \tag{21}$$

Инициализация весов может производиться различными способами. Необходимо брать небольшие случайные значения, например: $w_j := random\left(-\frac{1}{2n}, \frac{1}{2n}\right)$.

Критерий останова такого алгоритма основан на приблизительной оценке функционала Q методом экспоненциальной скользящей средней. Точное значение потребовало бы вычисления l скалярных произведений $\langle w, x_i \rangle$, что довольно накладно. Когда градиентный метод подходит к окрестности минимума, оценка скользящего среднего стабилизируется и приближается к точному значению функционала Q.

Метод стохастического градиента хорошо приспособлен для динамического обучения, когда обучающие объекты поступают потоком, и надо быстро обновлять вектор весов при появлении каждого нового объекта.

Метод позволяет настраивать веса на избыточно больших выборках за счет того, что случайной подвыборки может оказаться достаточно для обучения. Допускаются различные стратегии обучения. В случае большой выборки или динамического потока можно вообще не сохранять обучающие объекты. В случае малой выборки можно повторно предъявлять для обучения одни и те же объекты, что способствует повышению качества классификации.

9. Обучение и тестирование тонового классификатора. Для обучения классификатора использовалась готовая выборка (http://study.mokoron.com/), состоящая приблизительно из 225 тысяч размеченных текстов, имеющих положительный и отрицательная окрас. В ходе тестирования качество классификации было максимизировано посредством перебора различных сочетаний классификаторов, методов векторизации, схем n-грамм и других параметров. Байесовский классификатор не нуждается в подборе параметров, а параметры линейной модели специально подбирались по сетке. В качестве сетки был использован словарь, в котором ключи представляли собой названия параметров, а значения состояли из наборов, которым требуется проверка.

Далее строилось декартово произведение на этих параметрах и по полученным точкам со всеми возможными наборами измерялось качество классификации. Этот процесс называется поиском по сетке.

В ходе тестирования были рассмотрены: вид функции потерь, вид регуляризации и множитель альфа перед регуляризацией.

В качестве стратегии кросс-валидации применялся метод ShuffleSplit из библиотеки scikit-learn, производилось 5 итераций и в тестовую выборку отсекалось 30 процентов данных. Результаты последних трех итераций представлены на рисунке 5.

```
N-gram Scheme: (1, 1)
Count Vectorizer
NB: 0.636833277424
Linear: 0.667829587387
Linear Parameters: {'alpha': 0.0001, 'penalty': 'elasticnet', 'loss': 'log'}

TF-IDF Vectorizer
NB: 0.583092921838
Linear: 0.698989898989
Linear Parameters: {'alpha': 1e-05, 'penalty': 'elasticnet', 'loss': 'log'}

N-gram Scheme: (1, 2)
Count Vectorizer
NB: 0.681784636028
Linear: 0.705333780611
Linear Parameters: {'alpha': 0.001, 'penalty': 'l2', 'loss': 'hinge'}

TF-IDF Vectorizer
NB: 0.608587722241
Linear: 0.717343173432
Linear Parameters: {'alpha': 1e-05, 'penalty': 'elasticnet', 'loss': 'log'}

N-gram Scheme: (1, 3)
Count Vectorizer
NB: 0.692787655149
Linear: 0.714793693391
Linear Parameters: {'alpha': 0.001, 'penalty': 'l2', 'loss': 'log'}

TF-IDF Vectorizer
NB: 0.633143240523
Linear: 0.719490103992
Linear Parameters: {'alpha': 0.0001, 'penalty': 'l2', 'loss': 'hinge'}

N-gram Scheme: (1, 4)
Count Vectorizer
NB: 0.69533713519
Linear: 0.719154646892
Linear Parameters: {'alpha': 0.001, 'penalty': 'l2', 'loss': 'hinge'}

TF-IDF Vectorizer
NB: 0.650788326065
Linear: 0.719490103992
Linear Parameters: {'alpha': 0.0001, 'penalty': 'l2', 'loss': 'hinge'}

N-gram Scheme: (1, 5)
Count Vectorizer
NB: 0.690640724589
Linear: 0.715062059712
Linear Parameters: {'alpha': 0.001, 'penalty': 'l2', 'loss': 'hinge'}

TF-IDF Vectorizer
NB: 0.660986246226
Linear: 0.718483730292
Linear Parameters: {'alpha': 0.0001, 'penalty': 'l2', 'loss': 'hinge'}
```

Рис. 5. Выбор модели классификатора

По результатам валидации лучшей оказалась линейная модель с триграммной схемой (1, 3), векторизатором TF-IDF и параметрами: penalty — l2 (функция штрафа L2-регуляризация, которая штрафует весовые значения добавлением суммы их квадратов к ошибке);

alpha — 0.000001(константа, которая умножает член регуляризации); loss — log (функция потерь в виде логистической регрессии). Ее результат составил ≈ 0.72.

Качество классификации превышает 70%, что говорит о правильном подборе релевантных обучающих выборок.

Теперь перейдем к практической реализации и тестовой эксплуатации системы для извлечения и анализа дорожно-транспортной информации с сайта http://autostrada.info/ru о состоянии транспортных сетей республики Крым и города Севастополь.

10. Практическая реализация. С использованием построенного тонового классификатора был проведен анализ отзывов пользователей портала http://autostrada.info/ru и осуществлена оценка качества транспортных сетей республики Крым и города Севастополь за период с 2011–2018 годы.

В результате анализа классификации 1130 отзывов были получены две выборки: 493 положительных отзыва и 637 отрицательных. Результаты анализа в таблице 3.

Таблица 3. Результаты автоматической классификации трасс по отзывам

Номер трассы	Наименование трассы	Участок трассы	Количество положительных отзывов	Количество отрицательных отзывов
			493	*637*
Н-05	Красноперекопск – Симферополь	-	25	8
Н-06	Симферополь – Севастополь	-	27	4
Н-19	Ялта – Севастополь	-	19	1
М-18	Харьков – Симферополь – Ялта	Новомосковск – Запорожье	49	199
М-18	Харьков – Симферополь – Ялта	Харьков – Новомосковск	39	111
М-18	Харьков – Симферополь – Ялта	Запорожье – Мелитополь	112	145
М-18	Харьков – Симферополь – Ялта	Симферополь – Ялта	14	2
М-18	Харьков – Симферополь – Ялта	Джанкой – Симферополь	13	6

Продолжение таблицы 3.

Номер трассы	Наименование трассы	Участок трассы	Количество положительных отзывов	Количество отрицательных отзывов
			493	*637*
М-18	Харьков – Симферополь – Ялта	Мелитополь – Джанкой	30	18
М-18	Харьков – Симферополь – Ялта	-	38	57
М-17	Херсон – Джанкой – Керчь	Херсон – Армянск	32	24
М-17	Херсон – Джанкой – Керчь	Джанкой – Феодосия	9	15
М-17	Херсон – Джанкой – Керчь	Красноперекопск – Феодосия	1	2
М-17	Херсон – Джанкой – Керчь	Феодосия – Керчь	5	5
М-17	Херсон – Джанкой – Керчь	Херсон – Джанкой	4	3
М-17	Херсон – Джанкой – Керчь	Красноперекопск – Джанкой	4	3
М-17	Херсон – Джанкой – Керчь	-	7	10
Р-23	Симферополь – Феодосия	-	18	6
Р-25	Симферополь – Евпатория	-	11	10
Р-27	Севастополь – Инкерман	-	10	-
Р-29	Алушта – Феодосия	Коктебель – Феодосия	5	-
Р-29	Алушта – Феодосия	-	11	2
Р-34	Ялта – Алушта	-	2	-
Р-35	Грушевка – Судак	-	5	4
Р-58	Окружная Севастополя	-	3	2

Наличие положительных и отрицательных отзывов пользователей в границах одного и того же участка трассы характеризуют оценки различных важных для водителей параметров.

Для наглядности результатов исследования приведем размеченную карту дорог Крыма и города Севастополь, соответствующую положительным (зеленый цвет) и отрицательным (желтый цвет) отзывам (см. рисунок 6).

Рис. 6. Размеченная в соответствии с отзывами карта дорог Крыма и города Севастополь

Примеры положительных и отрицательных отзывов, полученных при классификации, представлены в таблице 4.

Таблица 4. Примеры классифицированных отзывов на положительные и отрицательные (трассы Крыма и города Севастополя)

Номер трассы	Трасса	Положительные	Отрицательные
М-17	Херсон – Джанкой – Керчь (Феодосия – Керчь)	Отличная дорога, хотя на карте написано очень плохая).	Ехать только днем, местами очень плохо.
М-18	'Харьков – Симферополь – Ялта' (Запорожье – Мелитополь)	Дорога хорошая. Средняя скорость 80-100 км/ч на обычной легковушке. Ям нет, все залатаны. Ехать можно не напрягаясь. Учтите, что даже ночью трасса загружена, много грузовиков и людей едущий на отдых.	Крайне не рекомендую ехать! Состояние ужасное! Ямы, наплывы, латки, колейность, стиральная доска. Объезжайте!

Продолжение таблицы 4.

Номер трассы	Трасса	Положительные	Отрицательные
А-146	Краснодар – Верхнебаканский (Абинск – Верхнебаканский)	Дорога отличная, есть незначительные пробои на мостах.	Просто ужас, ехать затруднительно и опасно. Через каждые 3-5 км на обочине памятники, напоминающие о ДТП.
Н-05	Красноперекопск – Симферополь	Трасса хорошая.	Дорога до границы не очень. Наплывы. Ехать днём. После границы ещё хуже.
Н-06	Симферополь – Севастополь	Всегда была в хорошем состоянии. Никаких проблем.	Последнее время много аварий, и качество дороги играет не последнюю роль
Н-19	Ялта – Севастополь	Отличная дорога, ям нет.	Из за жары регулярно куча ДТП.
Р-23	Симферополь – Феодосия	Хорошая дорога.	Не очень.
Р-25	Симферополь – Евпатория	Дорога хорошая, ям нет, трещин мало.	Дороги нет. Она отсутствует вообще. Только по территории города Саки остался кусок который сделали. Все остальное просто яма. Любой Фома пусть сам проедет по ней 100 метров с закрытыми глазами на скорости 80 км и останется вообще без колес.
Р-27	Севастополь – Инкерман	Отличная двухполосная дорога, без намеков на ямы.	-

Продолжение таблицы 4.

Номер трассы	Трасса	Положительные	Отрицательные
Р-29	Алушта – Феодосия (Коктебель - Феодосия)	Горная дорога. Ехать комфортно. Ямы иногда на повороте, а так хорошее покрытие. Ехать не быстро, но машин не много. Очень красивые виды.	Между Судаком и Коктебелем ужасная горная дорога.
Р-34	Ялта – Алушта	Хорошая дорога.	-
Р-35	Грушевка – Судак	Нормальная трасса, с нормальным покрытием, есть небольшой участок через горы возле Судака, дальше ровная дорога.	Ужасная по состоянию дорога.

Каждый отзыв содержал следующие атрибуты: номер и наименование трассы, наименование уточненного участка трассы, дату и время регистрации отзыва и отзыв о качестве дороги. Примеры положительного и отрицательного отзывов представлены на рисунках 7 и 8.

```
841 =>
array (
  'title' => 'Н-19 Ялта – Севастополь',
  'subtitle' => NULL,
  'date' => '02.11.2015 20:18',
  'rate' => 'Дорога с идеальным или близким к идеальному покрытием',
  'description' => 'Дорожное покрытие хорошее, трещины есть, но мало. ',
```

Рис. 7. Пример положительного отзыва

```
217 =>
array (
  'title' => 'М-18 Харьков – Симферополь – Ялта',
  'subtitle' => 'Харьков – Новомосковск',
  'date' => '19.05.2018 22:12',
  'rate' => 'Дорожное покрытие полностью разрушено, постоянные ямы',
  'description' => 'Дорога в ужасном состоянии на отрезке Харьков – Днепр,
                    старайтесь строить маршрут через М-29 ',
```

Рис. 8. Пример отрицательного отзыва

По результатам анализа отзывов можно сделать вывод о том, что качество транспортных сетей Крыма с 2016 года постепенно

улучшается. На сентябрь 2018 около 60% дорожно-транспортной инфраструктуры республики Крым и города Севастополь все еще требуют ремонта. Оставшиеся 40 % находятся в удовлетворительном состоянии. Так, например, требуют ремонта и расширения дороги регионального значения: Р-29 Алушта – Феодосия, Р-23 Симферополь – Феодосия, Р-35 Грушевка – Судак, Р-25 Симферополь – Евпатория, Р-260 Таврида, Р-58 окружная дорога Севастополя и другие. Также необходимо увеличение снегоуборочной и ремонтной техники и повышение качества работы дорожных служб, так как в период с января 2017 – март 2017 негативные отзывы пользователей содержали информацию о заторах и ДТП по причине ухудшения качества дорожного полотна.

В дальнейшем планируется реализовать глубокую классификацию отзывов по тематическим группам, таким как: пробки, ДТП, ремонт, гололед и снежные заторы, ямы и выбоины, пропажа людей, штрафы и другое.

В рамках следующего этапа планируется также сравнить методы bag-of-words и tf-idf с методом векторного представления слов word2vec, который в ряде работ [41, 42] показал лучшие результаты. Также планируется рассмотреть методы тематической классификации текстов, такие как свёрточные нейронные сети (CNN) [43, 44], метод опорных векторов (SVM) [45, 46] и др.

11. Заключение. В ходе тестирования работы классификатора проведен анализ отзывов пользователей, относящихся к качеству транспортных сетей республики Крым и города Севастополь. Классификатор позволил разделить отзывы на положительные и отрицательные, и выявить проблемные участки транспортных сетей Крыма.

Подобные системы, основанные на анализе отзывов пользователей, позволят устанавливать причинно-следственные связи транспортно-логистической активности населения [40], формировать кодифицированные библиотеки шаблонов транспортного поведения [47], выполнять среднесрочное и долгосрочное прогнозирование процессов транспортной мобильности, формировать новые [48] и расширять существующие критерии и параметры управления транспортными потоками [49], выходя за рамки циклов светофорного регулирования и типовых схем прокладки маршрутов.

Использование систем оперативного анализа разнородных данных web-контента в составе интеллектуальных транспортных систем является эффективной и обоснованной технологией сегодняшнего времени.

Авторский коллектив благодарит администрацию сайта autostrada.info/ru за предоставленное разрешение на обработку и анализ текстовой информации.

Литература

1. *Seliverstov Y.A. et al.* Development of management principles of urban traffic under conditions of information uncertainty // Conference on Creativity in Intelligent Technologies and Data Science. 2017. pp. 399–418.
2. *Искандеров Ю.М.* Интеллектуальные транспортные системы: возможности и особенности применения // Мир дорог. 2013. № 68. С. 38–39.
3. *Искандеров Ю.М.* Использование инструментария семантических графов с оболочками при создании интеллектуальных транспортных систем // Международная научно-практическая конференция «Интеллектуальные системы на транспорте». 2011. С. 75–82.
4. *Искандеров Ю.М.* Построение модели интегрированной информационной системы транспортной логистики на основе мультиагентных технологий // Сборник статей Международной научно-практической конференции «Новая экономика и основные направления ее формирования». 2016. С. 62–69.
5. *Искандеров Ю.М., Ласкин М.Б., Лебедев И.С.* Особенности моделирования транспортно-технологических процессов в цепях поставок // Восьмая Всероссийская научно-практическая конференция «Имитационное моделирование. Теория и практика» (ИММОД-2017). 2017. С. 110–113.
6. *Свистунова А.С., Чумак А.С.* Интеллектуализация информационного обеспечения процесса перевозки негабаритных грузов // XVII Международная научно-практическая конференция «Логистика: современные тенденции развития». 2018. С. 76–79.
7. *Seliverstov Y.A. et al.* The method of selecting a preferred route based on subjective criteria // 2017 IEEE II International Conference on Control in Technical Systems (CTS). 2017. pp. 126–130.
8. *Seliverstov Ya.A. et al.* Intelligent systems preventing road traffic accidents in megalopolises in order to evaluate // 2017 20th IEEE International Conference on Soft Computing and Measurements (SCM). 2017. pp. 489–492.
9. *Малыгин И.Г., Комашинский В.И., Афонин П.Н.* Системный подход к построению когнитивных транспортных систем и сетей // Научно-аналитический журнал «Вестник Санкт-Петербургского университета Государственной противопожарной службы МЧС России». 2015. № 4. С. 68–73.
10. *Gal-Tzur A., Rechavi A., Beimel D., Freund S.* An improved methodology for extracting information required for transport related decisions from Q & A forums: A case study of TripAdvisor // Travel Behaviour and Society. 2018. vol. 10. pp. 1–9.
11. *Chaniotakis E., Antoniou C.* Use of Geotagged Social Media in Urban Settings: Empirical Evidence on its Potential from Twitter // 2015 IEEE 18th International Conference on Intelligent Transportation Systems. 2015. pp. 214–219.
12. *Kuflik T. et al.* Automating a framework to extract and analyse transport related social media content: The potential and the challenges // Transportation Research Part C: Emerging Technologies. 2017. vol. 77. pp. 275–291.
13. *Ali F. et al.* Fuzzy Ontology-based Sentiment Analysis of Transportation and City Feature Reviews for Safe Traveling // Transportation Research Part C: Emerging Technologies. 2017. vol. 77. pp. 33–48.
14. *Nanba H. et al.* Automatic compilation of travel information from automatically identified travel blogs // Proceedings of the ACL-IJCNLP 2009 Conference Short Papers. 2009. pp. 205–208.

15. *Zhang Z. et al.* Final Report. Mining Transportation Information from Social Media for Planned and Unplanned Events // Transportation Informatics, University Transportation Center. 2016. 68 p.
16. *Тихомиров И.А. и др.* Инструменты анализа научно-технологических заделов России // Труды Института системного анализа Российской академии наук. 2016. Т. 66. № 3. С. 98–104.
17. *Ананьева М.И.* О проблеме выявления экстремистской направленности в текстах // Вестник Новосибирского государственного университета. Серия: Информационные технологии. 2016. Т. 14. № 4. С. 5–13.
18. *Gu Y., Qian Z.S., Chen F.* From Twitter to detector: Real-time traffic incident detection using Social Media data // Transportation research part C: emerging technologies. 2016. vol. 67. pp. 321–342.
19. *Kuflik T. et al.* Automating a framework to extract and analyse transport related Social Media content: The potential and the challenges // Transportation Research Part C: Emerging Technologies. 2017. vol. 77. pp. 275–291.
20. *Serna A., Gerrikagoitia J.K., Bernabe U., Ruiz T.* A Method to Assess Sustainable Mobility for Sustainable Tourism: The Case of the Public Bike Systems // Information and Communication Technologies in Tourism. 2017. pp. 727–739.
21. *Serna A., Gasparovic S.* Transport analysis approach based on big data and text mining analysis from social media // Transportation Research Procedia. 2018. vol. 33. pp. 291–298.
22. *Блеканов И.С., Бондаренко Д.С.* Оценка эффективности методов поиска тематических сообществ в веб-пространстве // Научно-технические ведомости Санкт-Петербургского государственного политехнического университета. Информатика. Телекоммуникации. Управление. 2010. № 5(108). С. 18–24.
23. *Печников А.А., Сотенко Е.М.* Программы-краулеры для сбора данных о представительских сайтах заданной предметной области – аналитический обзор // Современные наукоемкие технологии. 2017. № 2. С. 58–62.
24. *Отраднов К.К., Раев В.К.* Экспериментальное исследование эффективности методик векторизации текстовых документов и алгоритмов их кластеризации // Вестник Рязанского государственного радиотехнического университета. 2018. № 64. С. 73–84.
25. *Михайлов Д.В., Козлов А.П., Емельянов Г.М.* Выделение знаний и языковых форм их выражения на множестве тематических текстов: подход на основе меры TF-IDF // Компьютерная оптика. 2015. Т. 39. № 3. С. 429–438.
26. *Ghaddar B., Naoum-Sawaya J.* High dimensional data classification and feature selection using support vector machines // European Journal of Operational Research. 2018. vol. 265. no. 3. pp. 993–1004.
27. *Rabiner L., Juang B.* Fundamentals of Speech Recognition // Prentice Hall. 1993. 507 p.
28. *Шелманов А.О. и др.* Семантико-синтаксический анализ текстов в задачах вопросно-ответного поиска и извлечения определений // Искусственный интеллект и принятие решений. 2016. № 4. С. 47–61.
29. *Кузнецов А.Н., Вышемирский Д.А.* Об одном подходе к решению задачи токенизации при анализе больших массивов пользовательских паролей // Безопасность информационных технологий. 2017. № 2. С. 50–60.
30. *Рубцова Ю.В.* Построение корпуса текстов для настройки тонового классификатора // Программные продукты и системы. 2015. № 1. С. 72–78.
31. *Мюллер А., Гвидо С.* Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными // Альфа-книга. 2017. 393 с.

32. Карякина А.А., Ботов Д.С. Анализ текстов для прогнозирования оттока клиентов Интернет-Провайдера // Челябинский физико-математический журнал. 2018. Т. 3. № 2. С. 227–236.
33. *Нугуманова А.Б., Бессмертный И.А., Пецина П., Байбурин Е.М.* Обогащение модели Bag of Words семантическими связями для повышения качества классификации текстов предметной области // Программные продукты и системы. 2016. № 2. С. 89–99.
34. *Кипяткова И.С.* Программно-алгоритмическое обеспечение создания синтаксическо-статистической модели русского языка по текстовому корпусу // Труды СПИИРАН. 2013. № 1(24). С. 332–348.
35. *Петровский М.И., Глазкова В.В.* Алгоритмы машинного обучения для задачи анализа и рубрикации электронных документов // Вычислительные методы и программирование: новые вычислительные технологии. 2007. Т. 8. № 2. С. 57–69.
36. *Сизов А.А., Николенко С.И.* Наивный Байесовский классификатор. DOCPLAYER. URL: https://docplayer.ru/45424867-Naivnyy-bayesovskiy-klassifikator.html. (дата обращения: 25.01.2019).
37. *Воронцов К.В.* Вероятностное тематическое моделированиею URL: http://www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf. (дата обращения: 25.01.2019).
38. *Воронцов К.В.* Лекции по линейным алгоритмам классификации. URL: http://www.machinelearning.ru/wiki/images/6/68/voron-ML-Lin.pdf. (дата обращения: 25.01.2019).
39. *Шаграев А.Г., Фальк В.Н.* Линейные классификаторы в задаче классификации текстов // Вестник Московского энергетического института. 2013. № 4. С. 204–208.
40. *Селиверстов Я.А., Селиверстов С.А.* Использование систем класса ГАТЛОСЭМИ для упреждения причин возникновения ДТП и неблагоприятных социальных исходов в «умном городе» // Научно-технические ведомости Санкт-Петербургского государственного политехнического университета. Информатика. Телекоммуникации. Управление. 2016. № 1(236). С. 65–81.
41. *Ботов Д.С., Кленин Ю.Д., Николаев И.Е.* Извлечение информации с использованием нейросетевых моделей языка на примере анализа вакансий в системах онлайн-рекрутмента // Вестник Югорского государственного университета. 2018. № 3(50). С. 37–48.
42. *Kim D., Seo D., Cho S., Kang P.* Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec // Information Sciences. 2019. vol. 477. pp. 15–29.
43. *Liao S. et al.* CNN for situations understanding based on sentiment analysis of twitter data // Procedia Computer Science. 2017. vol. 111. pp. 376–381.
44. *Lee G. et al.* Sentiment classification with word localization based on weakly supervised learning with a convolutional neural network // Knowledge-Based Systems. 2018. vol. 152. pp. 70–82.
45. *Deng Y., Sander A., Faulstich L., Denecke K.* Towards automatic encoding of medical procedures using convolutional neural networks and autoencoders // Artificial Intelligence in Medicine. 2019. vol. 93. pp. 29–42.
46. *Alimova I.S., Tutubalina E.V.* Entity-level classification of adverse drug reactions: a comparison of neural network models // Proceedings of the Institute for System Programming of the RAS. 2018. vol. 30. no. 5. pp. 177–196.
47. *Селиверстов Я.А., Селиверстов С.А.* Формальное построение цепочек транспортной активности городского населения // Научно-технические ведомости СПбГПУ. Информатика. Телекоммуникации. Управление. 2015. № 4(224). С. 91–104.

48. *Селиверстов С.А., Селиверстов Я.А.* Обзор показателей транспортной обеспеченности мегаполиса // Вестник гражданских инженеров. 2015. № 5(52). С. 237–247.
49. *Селиверстов С.А., Селиверстов Я.А.* О методе оценки эффективности организации процесса дорожного движения мегаполиса // Вестник транспорта Поволжья. 2015. № 2(50). С. 91–96.

Селиверстов Ярослав Александрович — канд. техн. наук, старший научный сотрудник, лаборатория интеллектуальных транспортных систем, Федеральное государственное бюджетное учреждение науки Институт проблем транспорта им. Н.С. Соломенко Российской академии наук (ИПТ РАН); магистрант, кафедра компьютерных систем и программных технологий института компьютерных наук и технологий, Санкт-Петербургский политехнический университет Петра Великого. Область научных интересов: интеллектуальные транспортные системы, машинное обучение, интеллектуальный анализ данных, компьютерное моделирование транспортных систем. Число научных публикаций — 69. seliverstov-yr@mail.ru; 12-я линия В.О., 13, 199178, Санкт-Петербург, Российская Федерация; р.т.: +7(812) 321-95-68.

Чигур Виктория Игоревна — студентка, факультет прикладной математики, Санкт-Петербургский государственный университет (СПбГУ). Область научных интересов: интеллектуальный анализ данных, машинное обучение, большие данные, компьютерная безопасность, системное программирование. Число научных публикаций — 3. v.chigur67@gmail.com; Университетская набережная, 7–9, 199034, Санкт-Петербург, Российская Федерация; р.т.: +7(812)328–20–00.

Сазанов Арсений Михайлович — аспирант, Институт компьютерных наук и технологий, Санкт-Петербургский политехнический университет Петра Великого. Область научных интересов: интеллектуальные системы, нейронные сети, искусственный интеллект. Число научных публикаций — 2. arseny.sazanov@gmail.com; Политехническая, 21, 195251, Санкт-Петербург, Российская Федерация; р.т.: +7(812)297-16-28.

Селиверстов Святослав Александрович — канд. техн. наук, старший научный сотрудник, лаборатория интеллектуальных транспортных систем, Федеральное государственное бюджетное учреждение науки Институт проблем транспорта им. Н.С. Соломенко Российской академии наук (ИПТ РАН); магистрант, Санкт-Петербургский политехнический университет Петра Великого. Область научных интересов: интеллектуальные транспортные системы, машинное обучение, интеллектуальный анализ данных, компьютерное моделирование транспортных систем. Число научных публикаций — 67. seliverstov_s_a@mail.ru; 12-я линия В.О., 13, 199178, Санкт-Петербург, Российская Федерация; р.т.: +7(812)321-95-68.

Свистунова Александра Сергеевна — программист, лаборатория информационных технологий на транспорте, Федеральное государственное бюджетное учреждение науки Санкт-Петербургского института информатики и автоматизации Российской академии наук (СПИИРАН). Область научных интересов: интеллектуальные транспортные системы, интеллектуальный анализ данных. Число научных публикаций — 2. svistunova_alexandra@bk.ru; 39, 14 линия В.О., 199178, Санкт-Петербург, Российская Федерация; р.т.: +7(812) 328-34-11.

Поддержка исследований. Работа выполнена при поддержке гранта РФФИ № 18-410-920016 в рамках инициативного проекта, проводимого совместно с Правительством Севастополя на тему: «Исследование социально-экономических и экологических процессов города Севастополя с ростом индустриального, транспортно-транзитного и туристского потенциалов».

DOI 10.15622/sp.18.2.354-389

Y.A SELIVERSTOV, V.I. CHIGUR, A.M. SAZANOV, S.A. SELIVERSTOV, A.S. SVISTUNOVA

SENTIMENT ANALYSIS OF "AUTOSTRADA.INFO/RU" USERS' COMMENTS

Seliverstov Y.A, Chigur V.I., Sazanov A.M., Seliverstov S.A., Svistunova A.S. **Sentiment Analysis of "AUTOSTRADA.INFO/RU" Users' Comments.**

Abstract. As a result of the analysis, it was revealed that social networks (Vkontakte, Facebook), thematic communities in microblogging networks (Twitter), resources for travelers (TripAdvisor), transport portals (Autostrada) are a source of up-to-date and operational information about the traffic situation, the quality of transport services and passenger satisfaction with the quality of levels of transport services. However, the existing transport monitoring systems do not contain software tools capable of collecting and analyzing traffic information located in the Internet environment. This paper discusses the task of building a system for automatically retrieving and classifying road traffic information from transport Internet portals and testing the developed system for analyzing the transport networks of Crimea and the city of Sevastopol. To solve this problem, an analysis of open source libraries for thematic data collection and analysis was carried out. An algorithm for extracting and analyzing texts has been developed. A crawler was developed using the Scrapy package in Python3, and user feedback from the portal http://autostrada.info/ru was collected on the state of the transport system of Crimea and the city of Sevastopol. For texts lemmatization and vector text transformation, the tf, idf, tf-idf methods and their implementation in the Scikit-Learn library were considered: CountVectorizer and TF-IDF Vectorizer. For word processing, Bag-of-Words and n-gram methods were considered. During the development of the classifier model, the naive Bayes algorithm (MultinomialNB) and the linear classifier model with optimization of the stochastic gradient descent (SGDClassifier) were used. As a training sample, a corpus of 225,000 labeled texts from the Twitter resource was used. The classifier was trained, during which the cross-validation strategy and the ShuffleSplit method were used. Testing and comparison of the results of the pitch classification were carried out. According to the results of validation, the linear model with the n-gram scheme [1, 3] and the vectorizer TF-IDF turned out to be the best. During the approbation of the developed system, the collection and analysis of reviews related to the quality of transport networks of the Republic of Crimea and the city of Sevastopol were conducted. Conclusions are drawn and prospects for further functional development of the developed tools are defined.

Keywords: Automatic Text Analysis, Crowlers, Classification of Texts, Intelligent Transport Systems, Machine Training, TF-IDF, Naive Bayes Algorithm, Linear Classifier, Sentiment Analysis.

Seliverstov Yaroslav Aleksandrovich — Ph.D., Senior Researcher, Laboratory of Intelligent Transport Systems, Solomenko Institute of Transport Problems of the Russian academy of sciences; Master Student, Department of Computer Systems and Software Technologies of Institute of Computer Science and Technology, Peter the Great St.Petersburg Polytechnic University. Research interests: intelligent transport systems, machine learning, data mining, computer simulation of transport systems. The number of publications — 69. seliverstov-yr@mail.ru; 13, 12-th Line V.O., 199178, St. Petersburg, Russian Federation; office phone: +7(812) 321-95-68.

Chigur Viktoriya Igorevna — bachelor's student, Faculty of Applied Mathematics, St. Petersburg State University. Research interests: data mining, machine learning, big data, computer security, system programming. The number of publications — 3. v.chigur67@gmail.com; 7–9, University Embankment, 199034, , Russian Federation; office phone: +7(812)328–20–00.

Sazanov Arseniy Mikhailovich — Ph.D. Student, Institute of Computer Science and Technology, Peter the Great St.Petersburg Polytechnic University. Research interests: intellectual systems, neural networks, artificial intelligence. The number of publications — 2. arseny.sazanov@gmail.com; 21, Polytechnicheskaya, 195251, St. Petersburg, Russian Federation; office phone: +7(812)297-16-28.

Seliverstov Svyatoslav Aleksandrovich — Ph.D., Laboratory of Intelligent Transport Systems, Laboratory of Intelligent Transport Systems, Solomenko Institute of Transport Problems of the Russian academy of sciences; Master Student, Peter the Great St.Petersburg Polytechnic University. Research interests: intelligent transport systems, machine learning, data mining, computer simulation of transport systems. The number of publications — 67. seliverstov_s_a@mail.ru; 13, 12-th Line V.O., 199178, St. Petersburg, Russian Federation; office phone: +7(812)321-95-68.

Svistunova Aliaksandra Sergeevna — programmer, Transport Information Technologies Laboratory, St. Petersburg Institute of Informatics and Automation of the Russian Academy of Sciences (SPIIRAS). Research interests: intelligent transport systems, data mining. The number of publications — 2. svistunova_alexandra@bk.ru; 39, 14-th Line V.O., 199178, St. Petersburg, Russian Federation; office phone: +7(812) 328-34-11.

Acknowledgements. The scientific research was supported by the Russian Foundation for Basic Research within the framework of the project № 18-410-920016 p_a "Research of socio-economic and ecological processes of Sevastopol with the growth of industrial, traffic, transit and tourist potentials".

References

1. Seliverstov Y.A. et al. Development of management principles of urban traffic under conditions of information uncertainty. Conference on Creativity in Intelligent Technologies and Data Science. 2017. pp. 399–418.
2. Iskanderov Yu.M. [Intellectual transport systems: possibilities and features of application]. *Mir dorog – World of roads.* 2013. vol. 68. pp. 38–39. (In Russ.).
3. Iskanderov Yu.M. [Using the toolkit of semantic graphs with shells when creating intelligent transport systems]. *Mezhdunarodnaya nauchno-prakticheskaya konferenciya "Intellektual'nye sistemy na transporte"* [Proceedings of International Scientific-Practical Conference on Intelligent Transport Systems]. 2011. pp. 75–82. (In Russ.).
4. Iskanderov Yu.M. [Building a model of an integrated transport logistics information system based on multi-agent technologies]. *Sbornik statej Mezhdunarodnoj nauchno-prakticheskoj konferencii "Novaya ehkonomika i osnovnye napravleniya ee formirovaniya"* [Proceedings of International Scientific-Practical Conference on New Economy and the Main Directions of its Formation]. 2016. pp. 62–69. (In Russ.).
5. Iskanderov Yu.M., Laskin M.B., Lebedev I.S. [Features of modeling of transport and technological processes in supply chains]. *Vos'maya Vserossijskaya nauchno-prakticheskaya konferenciya "Imitacionnoe modelirovanie. Teoriya i praktika"* [8th All-Russian Scientific-Practical Conference "Simulation. Theory and Practice" (IMMOD-2017)]. 2017. pp. 110–113. (In Russ.).

6. Svistunova A.S., Chumak A.S. [Intellectualization of information support of the process of transportation of oversized cargo]. *XVII Mezhdunarodnaya nauchno-prakticheskaya konferenciya "Logistika: sovremennye tendencii razvitiya"* [XVII International Scientific-Practical Conference "Logistics: Modern Development Trends"]. 2018. pp. 76–79. (In Russ.).
7. Seliverstov Y.A. et al. The method of selecting a preferred route based on subjective criteria. 2017 IEEE II International Conference on Control in Technical Systems (CTS). 2017. pp. 126–130.
8. Seliverstov Ya.A. et al. Intelligent systems preventing road traffic accidents in megalopolises in order to evaluate. 2017 20th IEEE International Conference on Soft Computing and Measurements (SCM). 2017. pp. 489–492.
9. Malygin I.G., Komashinskiy V.I., Afonin P.N. [System approach to the construction of cognitive transport systems and networks]. *Nauchno-analiticheskij zhurnal "Vestnik Sankt-Peterburgskogo universiteta Gosudarstvennoj protivopozharnoj sluzhby MCHS Rossii" – Scientific and analytical journal Bulletin of the St. Petersburg University of the State Fire Service EMERCOM of Russia.* 2015. vol. 4. pp. 68–73. (In Russ.).
10. Gal-Tzur A., Rechavi A., Beimel D., Freund S. An improved methodology for extracting information required for transport related decisions from Q & A forums: A case study of TripAdvisor. *Travel Behaviour and Society.* 2018. vol. 10. pp. 1–9.
11. *Chaniotakis E., Antoniou C.* Use of Geotagged Social Media in Urban Settings: Empirical Evidence on its Potential from Twitter. 2015 IEEE 18th International Conference on Intelligent Transportation Systems. 2015. pp. 214-219.
12. Kuflik T. et al. Automating a framework to extract and analyse transport related social media content: The potential and the challenges. *Transportation Research Part C: Emerging Technologies.* 2017. vol. 77. pp. 275–291.
13. Ali F. et al. Fuzzy Ontology-based Sentiment Analysis of Transportation and City Feature Reviews for Safe Traveling. *Transportation Research Part C: Emerging Technologies.* 2017. vol. 77. pp. 33–48.
14. Nanba H. et al. Automatic compilation of travel information from automatically identified travel blogs. Proceedings of the ACL-IJCNLP 2009 Conference Short Papers. 2009. pp. 205–208.
15. Zhang Z. et al. Final Report. Mining Transportation Information from Social Media for Planned and Unplanned Events. Transportation Informatics, University Transportation Center. 2016. 68 p.
16. Tikhomirov I.A. et al. [Analysis tools for scientific and technological groundwork in Russia]. *Trudy Instituta sistemnogo analiza Rossijskoj akademii nauk – Proceedings of the Institute for System Analysis of the Russian Academy of Sciences*. 2016. vol. 66. no. 3. pp. 98–104. (In Russ.).
17. Anan'yeva M.I. [On the problem of identifying extremist orientation in the texts]. *Vestnik Novosibirskogo gosudarstvennogo universiteta. Seriya: Informacionnye tekhnologii – Bulletin of the Novosibirsk State University. Series: Information Technology.* 2016. vol. 14. no. 4. pp. 5–13. (In Russ.).
18. Gu Y., Qian Z.S., Chen F. From Twitter to detector: Real-time traffic incident detection using Social Media data. *Transportation research part C: emerging technologies*. 2016. vol. 67. pp. 321–342.
19. Kuflik T. et al. Automating a framework to extract and analyse transport related Social Media content: The potential and the challenges. *Transportation Research Part C: Emerging Technologies.* 2017. vol. 77. pp. 275–291.
20. Serna A., Gerrikagoitia J.K., Bernabe U., Ruiz T. A Method to Assess Sustainable Mobility for Sustainable Tourism: The Case of the Public Bike Systems. Information and Communication Technologies in Tourism. 2017. pp. 727–739.

21. Serna A., Gasparovic S. Transport analysis approach based on big data and text mining analysis from social media. *Transportation Research Procedia.* 2018. vol. 33. pp. 291–298.
22. Blekanov I.S., Bondarenko D.S. [Evaluation of the effectiveness of search methods for thematic communities in the web space]. *Nauchno-tekhnicheskie vedomosti Sankt-Peterburgskogo gosudarstvennogo politekhnicheskogo universiteta. Informatika. Telekommunikacii. Upravlenie – Scientific and Technical Gazette of St. Petersburg State Polytechnic University. Computer science. Telecommunications. Control.* 2010. vol. 5(108). pp. 18–24. (In Russ.).
23. Pechnikov A.A., Sotenko Ye.M. [Crawler programs for collecting data on representative sites of a given subject area – analytical review]. *Sovremennye naukoemkie tekhnologii – Modern high technologies.* 2017. vol. 2. pp. 58–62. (In Russ.).
24. Otradnov K.K., Rayev V.K. [Experimental study of the effectiveness of methods for vectorization of text documents and algorithms for their clustering]. *Vestnik Ryazanskogo gosudarstvennogo radiotekhnicheskogo universiteta – Vestnik of Ryazan State Radio Engineering University.* 2018. vol. 64. pp. 73–84. (In Russ.).
25. Mikhaylov D.V., Kozlov A.P., Yemel'yanov G.M. [The selection of knowledge and linguistic forms of their expression on the set of thematic texts: an approach based on the measure TF-IDF]. *Komp'yuternaya optika – Computer Oprics.* 2015. vol. 39. no. 3. pp. 429–438. (In Russ.).
26. Ghaddar B., Naoum-Sawaya J. High dimensional data classification and feature selection using support vector machines. *European Journal of Operational Research.* 2018. vol. 265. no. 3. pp. 993–1004.
27. Rabiner L., Juang B. Fundamentals of Speech Recognition.Prentice Hall. 1993. 507 p.
28. Shelmanov A.O. et al. [Semantic-syntactic analysis of texts in the tasks of question-answer search and extraction of definitions]. *Iskusstvennyj intellekt i prinyatie reshenij – Artificial Intelligence and Decision Making.* 2016. vol. 4. pp. 47–61. (In Russ.).
29. Kuznetsov A.N., Vyshemirskiy D.A. [On one approach to solving the tokenization problem when analyzing large arrays of user passwords]. *Bezopasnost' informacionnyh tekhnologij – Information Technology Security.* 2017. vol. 2. pp. 50–60. (In Russ.).
30. Rubtsova YU.V. [Building a corpus of texts to adjust the tone classifier]. *Programmnye produkty i sistemy – Software products and systems.* 2015. vol. 1. pp. 72–78. (In Russ.).
31. Myuller A., Gvido S. *Vvedeniye v mashinnoye obucheniye s pomoshch'yu Python. Rukovodstvo dlya spetsialistov po rabote s dannymi* [Introduction to machine learning using Python. A Guide for Data Specialists]. Al'fa-kniga 2017. 393 p. (In Russ.).
32. Karyakina A.A., Botov D.S. [Analysis of texts for forecasting the outflow of clients of the Internet provider]. *Chelyabinskij fiziko-matematicheskij zhurnal – Chelyabinsk Physics and Mathematics Journal.* 2018. vol. 3. no. 2. pp. 227–236. (In Russ.).
33. Nugumanova A.B., Bessmertnyy I.A., Petsina P., Bayburin Ye.M. [Enrichment of the Bag of Words model with semantic links to improve the quality of domain text classification]. *Programmnye produkty i sistemy – Software products and systems.* 2016. vol. 2. pp. 89–99. (In Russ.).
34. Kipyatkova I.S. [Software and algorithmic support for the creation of a syntactic-statistical model of the Russian language by text corpus]. *Trudy SPIIRAN – Proceedings of SPIIRAS*. 2013. vol. 1(24). pp. 332–348. (In Russ.).
35. Petrovskiy M.I., Glazkova V.V. [Machine learning algorithms for the task of analyzing and categorizing electronic documents]. *Vychislitel'nye metody i programmirovanie: novye vychislitel'nye tekhnologii – Numerical methods and programming: new computing technologies.* 2007. Issue 8. vol. 2. pp. 57–69. (In Russ.).

36. Sizov A.A., Nikolenko S.I. Naivnyy Bayyesovskiy klassifikator [Naive Bayes Classifier]. Available at: https://docplayer.ru/45424867-Naivnyy-bayesovskiy-klassifikator.html (accessed: 25.01.2019). (In Russ.).
37. K.V. Vorontsov. Veroyatnostnoye tematicheskoye modelirovaniye [Probabilistic thematic modeling]. Available at: http://www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf (accessed: 25.01.2019). (In Russ.).
38. Vorontsov K.V. Lektsii po lineynym algoritmam klassifikatsii [Lectures on linear classification algorithms]. Available at: http://www.machinelearning.ru/wiki/images/6/68/voron-ML-Lin.pdf (accessed: 25.01.2019). (In Russ.).
39. Shagrayev A.G., Fal'k V.N. [Linear classifiers in the task of classifying texts]. *Vestnik Moskovskogo ehnergeticheskogo instituta – Bulletin of the Moscow Power Engineering Institute.* 2013. vol. 4. pp. 204–208. (In Russ.).
40. Seliverstov Ya.A., Seliverstov S.A. [The use of GATLOSEMI class systems to pre-empt the causes of accidents and adverse social outcomes in the "smart city"]. *Nauchno-tekhnicheskie vedomosti Sankt-Peterburgskogo gosudarstvennogo politekhnicheskogo universiteta. Informatika. Telekommunikacii. Upravlenie – Scientific and technical statements of the St. Petersburg State Polytechnic University. Computer science. Telecommunications. Control.* 2016. № 1(236). pp. 65–81. (In Russ.).
41. Botov D.S., Klenin Yu.D., Nikolayev I.Ye. [Extraction of information using neural network language models on the example of the analysis of vacancies in online recruitment systems]. *Vestnik Yugorskogo gosudarstvennogo universiteta – Bulletin of the Yugra State University.* 2018. № 3(50). pp. 37–48. (In Russ.).
42. Kim D., Seo D., Cho S., Kang P. Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec. *Information Sciences.* 2019. vol. 477. pp. 15–29.
43. Liao S. et al. CNN for situations understanding based on sentiment analysis of twitter data. *Procedia Computer Science*. 2017. vol. 111. pp. 376–381.
44. Lee G. et al. Sentiment classification with word localization based on weakly supervised learning with a convolutional neural network. *Knowledge-Based Systems*. 2018. vol. 152. pp. 70–82.
45. Deng Y., Sander A., Faulstich L., Denecke K. Towards automatic encoding of medical procedures using convolutional neural networks and autoencoders. *Artificial Intelligence in Medicine*. 2019. vol. 93. pp. 29–42.
46. Alimova I.S., Tutubalina E.V. Entity-level classification of adverse drug reactions: a comparison of neural network models. *Proceedings of the Institute for System Programming of the RAS.* 2018. vol. 30. no. 5. pp. 177–196.
47. Seliverstov Ya.A., Seliverstov S.A. [Formal construction of transport activity chains of the urban population]. *Nauchno-tekhnicheskie vedomosti SPbGPU. Informatika. Telekommunikacii. Upravlenie – St. Petersburg State Polytechnic University Journal. Computer science. Telecommunications. Control*. 2015. vol. 4(224). pp. 91–104. (In Russ.).
48. Seliverstov S.A., Seliverstov Ya.A. [Review of metropolitan transport security indicators]. *Vestnik grazhdanskikh inzhenerov – Bulletin of Civil Engineers.* 2015. vol. 5(52). pp. 237–247. (In Russ.).
49. Seliverstov S.A., Seliverstov Ya.A. [On the method of evaluating the effectiveness of the organization of the traffic process of a megacity]. *Vestnik transporta Povolzh'ya – Bulletin of transport of the Volga region*. 2015. vol. 2(50). pp. 91–96. (In Russ.).

УДК 004.95 **DOI** 10.15622/sp.18.2.390-415

А.В. Воробьев, Г.Р. Воробьева, Н.И. Юсупова

КОНЦЕПЦИЯ ЕДИНОГО ПРОСТРАНСТВА ГЕОМАГНИТНЫХ ДАННЫХ

Воробьев А.В., Воробьева Г.Р., Юсупова Н.И. **Концепция единого пространства геомагнитных данных.**

Аннотация. Задача мониторинга параметров геомагнитного поля и его вариаций преимущественно решается сетью магнитных обсерваторий и вариационных станций, однако значимым препятствием при обработке и анализе получаемых таким образом данных наряду с их пространственной анизотропией являются пропуски (или полное отсутствие) достоверных значений и частичное несоответствие установленному формату. Неоднородность и аномальность данных исключает (существенно усложняет) возможность их автоматической интеграции и применения к ним инструментария для частотного анализа. Известные решения по интеграции разнородных геомагнитных данных базируются преимущественно на модели консолидации и лишь частично решают данную проблему. Получаемые в результате наборы данных, как правило, не соответствуют требованиям IAGA (International Association of Geomagnetism and Aeronomy — Международной ассоциации геомагнетизма и аэрономии), рекомендуемым к представлению результатов геомагнитных наблюдений. При этом пропуски во временных рядах устраняются известными средствами обработки геомагнитных данных путем исключения отсутствующих или аномальных значений из конечной выборки, что, очевидно, может привести как к потере актуальной информации о ходе изменения параметров геомагнитного поля и его вариаций, нарушению шага дискретизации, так и к неоднородности временного ряда. Предлагается подход к созданию единого пространства геомагнитных данных, основанный на комбинировании моделей консолидации и федерализации, включающий предварительную обработку исходных временных рядов с опционально доступной процедурой их восстановления и верификации, ориентированный на применение технологий облачных вычислений и иерархического формата с целью повышения вычислительной скорости обработки больших объемов данных и, как следствие, обеспечивающий получение пользователями более качественных и однородных данных.

Ключевые слова: геомагнитные данные, магнитные обсерватории, временные ряды, большие данные, единое информационное пространство, параллельные вычисления.

1. Введение. В настоящее время установлены и активно изучаются многочисленные негативные эффекты воздействия космической среды на объекты народного хозяйства, наиболее ярко проявляющиеся в периоды так называемых магнитных бурь. Особого внимания заслуживают такие эффекты, как:

– магнитное торможение искусственных спутников Земли;

– нарушение коротковолновой радиосвязи;

– дополнительные погрешности прецизионной магнитометрической аппаратуры;

– радиационное воздействие на биологические объекты, находящиеся в верхних слоях атмосферы;

– токовые наводки в трубопроводах, трансокеанских кабелях, системах автоматики высокоширотных железных дорог.

Актуальная в этой связи проблема многопараметрового мониторинга геомагнитного поля и его вариаций преимущественно решается посредством магнитных обсерваторий, аэромагнитных, гидромагнитных съемок, спутниковых и подземных скважинных наблюдений, а также с помощью портативных магнитометров различного принципа действия и динамического диапазона [1]. Каждый из способов имеет свои характерные преимущества и недостатки относительно других и используется для решения определенного круга задач как прикладного, так и фундаментального характеров. Однако, благодаря реализуемой концепции объединения магнитных обсерваторий в сети (INTERMAGNET [2, 3], IMAGE [4], AUTUMNX [5, 6] и пр. [7]) и открытому удаленному доступу к регистрируемым ими данным, именно они являются наиболее достоверным, распространенным и доступным для большинства ученых и специалистов методами наблюдения вариаций геомагнитного поля.

По этой и другим причинам (по состоянию на 2018 г.) свыше 30 известных широкой общественности сетей магнитных обсерваторий объединяют более 300 магнитных станций и обсерваторий, чье неравномерное распределение по поверхности Земли обусловливает относительный избыток геомагнитных данных в одних регионах планеты и их дефицит в других [1]. Однако, помимо проблем, сопряженных с геопространственным распределением магнитных обсерваторий, значимым препятствием на пути интеграции, обработки и анализа получаемой с их помощью информации являются несоблюдение единого формата и несогласованность представления результатов наблюдений параметров геомагнитного поля и его вариаций, а также различный период их регистрации. При этом разнородность геомагнитных данных обусловлена гетерогенностью их источников, одни из которых представлены магнитными обсерваториями, регистрирующими абсолютные значения параметров магнитного поля Земли, а другие — вариационными станциями, осуществляющими наблюдение за параметрами геомагнитных вариаций.

Предлагается одно из возможных решений задачи интеграции гетерогенных источников геомагнитных данных в единое информационное пространство, позволяющее потребителям получать полную и достоверную информацию о состоянии магнитного поля Земли в любой известный момент времени и в любой точке пространства. Приводится описание предложенной концепции, архитектуры единого пространства геомагнитных данных, а также инфокоммуникационных технологий, реализующих указанный подход.

2. Состояние вопроса. До начала 90-х годов XX века основным способом представления результатов наблюдений параметров магнитного поля Земли и его вариаций являлись магнитограммы на бумажном носителе [8, 9]. С развитием технологий вариационные станции и магнитные обсерватории перешли на цифровую систему регистрации и обработки геомагнитных данных, результатом которой стали ежесуточные файлы минутных значений вариаций геомагнитного поля с привязкой к абсолютным наблюдениям. И, наконец, эволюция инфокоммуникационных технологий позволила обеспечить открытость геомагнитных данных посредством стандартных сетевых протоколов и веб-ориентированных интерфейсов.

В итоге на сегодняшний день большинство магнитных обсерваторий и станций располагают размещенными в сети Интернет ресурсами, на которых доступны результаты многолетних наблюдений параметров геомагнитного поля и его вариаций. Так, к примеру, на вебресурсе сети INTERMAGNET (ftp://ftp.seismo.nrcan.gc.ca/intermagnet) размещены вариативные, предварительные, квазиокончательные и окончательные результаты поминутных и посекундных наблюдений параметров геомагнитного поля с 1991 года по настоящее время [2, 3]. Также в репозитории сети IMAGE (http://space.fmi.fi/image/www/index.php?page=home) доступны статистические таблицы и графики, отражающие результаты наблюдений компонент вектора геомагнитного поля (с шагом дискретизации 10-20 с или 1 мин) [4].

Однако и на современном уровне развития технологий существует ряд технических проблем, возникающих при обработке и анализе больших объемов гетерогенных геомагнитных данных, которые характерны и для других типов данных о состоянии окружающей среды [10, 11]. В первую очередь, магнитные сети, как правило, не соблюдают общепринятого формата представления результатов геомагнитных наблюдений, внося изменения в стандартный формат IAGA2002 [7] или используя собственные специфичные форматы, которые базируются на текстовом представлении табличных данных TSV (tab-separated values) [4]. Кроме того, наблюдаемые станциями и обсерваториями параметры геомагнитного поля и его вариаций отличаются: одни из них регистрируют направление магнитного поля, выраженное измеряемыми в градусах склонением (D) и наклонением (I), а другие — описывают полный вектор напряженности геомагнитного поля (F) на основе трех его компонент, измеряемых в нТл в различных системах координат [12-14]. И, наконец, может варьироваться шаг дискретизации регистрации геомагнитных данных, который принимает значения от полсекунды [6] до минуты [2, 4].

В определенном смысле усугубляет обозначенную проблему и тот факт, что регистрируемые станциями и обсерваториями геомагнитные данные содержат систематические пропуски (или полное от-

сутствие) достоверных значений за определенный период наработки IMO-станции (так, например, в период с 00:00 по 23:59 ч (UTC) 9 марта 2017 г. магнитной обсерваторией «Arti» пропущено 26 % значений временного ряда, магнитной обсерваторией «Guam» – 0.17 %, данные по обсерватории «Vostok» отсутствуют и т.д.). Следует также отметить и то, что геомагнитные данные, публикуемые INTERMAGNET и другими магнитными сетями, могут быть как очевидно недостоверными, так и не соответствовать формату IAGA-2002. Немаловажен и тот факт, что неполнота временных рядов (уже на теоретическом уровне) исключает возможность применения к ним математического и программно-алгоритмического инструментария частотного анализа и влияет на достоверность результатов сопряженных исследований, опирающихся на геомагнитные данные, получаемые посредством этих обсерваторий [1].

При этом информационные нужды потребителей геомагнитных данных не всегда ограничиваются масштабами одной магнитной сети и могут быть связаны с анализом и обработкой результатов наблюдений нескольких станций и обсерваторий, относящихся в том числе и к разным сетям. В этом случае потребитель сталкивается с очевидной технической проблемой, связанной с высокой трудоемкостью и затратами времени на поиск и получение необходимых данных. При этом все выявленные ранее проблемы геомагнитных данных (формат, структура, шаг дискретизации и т.д.) выходят на первый план и делают невозможным их автоматизированный анализ и обработку.

Попытка объединить результаты разнородных наблюдений параметров геомагнитного поля и его вариаций была предпринята в рамках проекта SuperMag (http://supermag.uib.no/index.html), инициированного в составе программы Electronic Geophysical Year (eGY, 2007–2008). Проект обеспечивает единый веб-ориентированный доступ к геомагнитным данным вариационных станций в единой координатной системе и унифицированных единицах измерения с заданным временным разрешением [14, 15]. В настоящее время ресурс предоставляет потребителям геомагнитные данные за период 1980–2010 годов, а количество доступных станций при этом варьируется от 90 до 165 в зависимости от анализируемого временного интервала.

В литературе [15] в качестве основного недостатка проекта SuperMag выделяют его ограниченность: в репозитории доступны только вариационные данные, в то время как отсутствуют абсолютные измерения, необходимые при исследовании динамики главного геомагнитного поля, а также практического применения для навигации, ориентации и геологии. В базе данных SuperMag также представлены данные, как правило, не менее семидневной давности, поэтому невозможно исполь-

зовать проект в качестве источника данных в режиме реального (или близком к нему) времени, в то время как такая необходимость часто возникает при решении различного рода прикладных задач.

Немаловажен и тот факт, что выбросы и пропущенные во временных рядах геомагнитных данных значения удаляются из набора предоставляемых потребителю данных, что может привести к потере критически важной информации, а также негативно сказывается на качестве и информативности систем визуализации данных. Еще один недостаток связан с тем, что поисковая система не отлажена и работает нестабильно, пропуская подавляющую часть запросов к станциям. И, наконец, проект SuperMag, как яркий пример консолидации данных, наследует его главный недостаток – избыточность, поскольку данные дублируются как в отдельных источниках, так и в едином репозитории.

Обозначенные проблемы интеграции результатов наблюдения параметров магнитного поля Земли и его вариаций, а также выявленные недостатки существующих подходов обуславливают актуальность научно-технической задачи, связанной с совершенствованием методов и средств распространения, обработки, анализа и визуализации больших объемов гетерогенных геомагнитных данных. Ее очевидным решением является объединение множества гетерогенных источников в единое пространство геомагнитных данных под управлением централизованного метода доступа, а также инструментария, обеспечивающего возможность их анализа и визуализации.

3. Постановка задачи. С целью расширения доступности результатов наблюдений параметров магнитного поля Земли и его вариаций необходимо разработать веб-ориентированную систему, реализующую инструментарий ведения и использования единого пространства геомагнитных данных как результата интеграции распределенных разнородных источников. Разработка информационной системы предполагает решение следующих задач:

1. Формулировка принципов объединения источников гетерогенных геомагнитных данных на основе моделей консолидации и федерализации для интеграции распределенных наборов разнородных данных.

2. Формализация структуры единого пространства геомагнитных данных, обеспечивающая объединение разнородных источников и повышение качества наборов данных за счет предварительной обработки для восстановления временных рядов и устранения аномальных значений в них.

3. Повышение вычислительной скорости получения и обработки больших объемов геомагнитных данных посредством создания виртуального вычислительного кластера, основанного на модели распределенных вычислений MapReduce и отличающегося тем,

что исходный выполняемый процесс выступает в качестве главного узла, а порождаемые им параллельные процессы являются рабочими узлами, реализующими изолированные запросы к гетерогенным источникам геомагнитных данных.

4. Принципы организации единого пространства геомагнитных данных. Под единым пространством геомагнитных данных, по аналогии с [16], будем понимать совокупность гетерогенных источников данных наблюдений за параметрами магнитного поля Земли и его вариаций, а также инфокоммуникационных технологий их интеграции, обработки, анализа и визуализации, функционирующих на основе единых принципов и обеспечивающих информационное взаимодействие поставщиков и потребителей геомагнитных данных, равно как и удовлетворение их информационных потребностей при решении прикладных и научно-исследовательских задач.

Согласно требованиям к структурной интеграции разнородных информационных ресурсов [17], единое пространство геомагнитных данных должно удовлетворять следующим основным принципам:

– прозрачности, согласно которому запросы потребителей геомагнитных данных не связаны с физическим расположением источников данных и формой представления информации в них;

– системности, что определяет необходимость формирования единого пространства как целостного образования, способного удовлетворить информационные потребности пользователей при решении прикладных и научно-исследовательских задач без выделения отдельных источников данных;

– технологичности, предполагающего комплексное использование различных технологий накопления и обработки геомагнитных данных, включая программно-инструментальные средства их комплексного анализа и визуализации.

В основе предлагаемой концепции лежит идея интеграции моделей консолидации и федерализации данных при их объединении в единое пространство с унифицированным для пользователей представлением информации [18]. Обе архитектуры являются однонаправленными и используются для предоставления наборов распределенных данных потребителю. Отличаются они друг от друга тем, что в первом случае происходит копирование данных в отдельное информационное хранилище, а во втором наборы данных формируются динамически без перемещения их из исходных источников.

С учетом перечисленных особенностей архитектур интеграции предлагается использовать консолидацию для сбора и долговременного физического хранения аналитических геомагнитных данных, а фе-

дерализацию — для формирования и виртуального хранения наборов оперативных геомагнитных данных. При этом под аналитическими данными будем понимать результаты всех предшествующих текущим суткам наблюдений параметров магнитного поля Земли, а под оперативными — геомагнитные данные, зарегистрированные магнитными обсерваториями и вариационными станциями в течение последних суток и автоматически преобразуемые в аналитические по их истечении. Результирующий набор может быть получен как интеграцией оперативных и аналитических данных, так и сформирован из выборки одного из них. При этом потребители геомагнитных данных имеют возможность в реальном или близком к реальному времени получить нужную информацию, охватываемую этим пространством [17, 18].

Еще одной отличительной особенностью предлагаемой концепции является развитие архитектуры федерализации данных: результаты наблюдений за параметрами магнитного поля Земли копируются в аналитическое хранилище после прохождения процедуры предварительной обработки. Необходимость такой доработки вызвана тем, что временные ряды геомагнитных данных являются, как правило, неполными и содержат различного рода аномалии, что может, в свою очередь, сказаться на достоверности результатов, их обработки и анализа. В рамках концепции восстановление временных рядов геомагнитных данных реализуется предложенным авторами индуктивным методом [19], основанным на статистических методах обработки временных рядов и принципах машинного обучения с использованием размеченных данных и отличающегося тем, что признаковым описанием фрагмента временного ряда выступает пара предшествующего и следующего за ним фрагментов того же ряда, в совокупности образующих обучающую выборку для поиска недостающего фрагмента по набору его признаков с последующим линейным масштабированием для восстановления исходного тренда информационного сигнала.

Таким образом, физическая независимость и скрытая от потребителя базовая неоднородность геомагнитных данных обеспечивают оперативность и высокую эффективность выполнения множественных обращений к данным под управлением единого метода доступа. При этом нивелируется проблема асинхронного обмена данными между несколькими источниками, свойственная консолидации данных и приводящая к тому, что запрашиваемая информация теряет свою актуальность на момент обращения к ней потребителя.

С учетом вышесказанного и согласно известной классификации [17], единое пространство геомагнитных данных является смешанным, поскольку сочетает в себе элементы централизован-

ной (с общим хранилищем данных) и децентрализованной (с распределенным хранилищем данных) архитектур. Присутствие в его структуре эвристически сложных вычислительных процедур позволяет полностью автоматизировать процедуру формирования хранилища геомагнитных данных, оставляя человеку-оператору только роль в подготовке данных на уровне их источников (станций и обсерваторий) за пределами единого пространства геомагнитных данных. При этом указанные вычислительные процедуры предназначены не только для сбора данных из распределенных источников, но и для выполнения аналитических операций и визуализации различных срезов данных.

5. Формализация структуры единого пространства геомагнитных данных. Структура единого пространства геомагнитных данных представлена тремя разделами, обеспечивающими сбор, хранение и обработку гетерогенных данных (рисунок 1). Основным его компонентом является распределенный набор источников геомагнитных данных, представленный как магнитными сетями, так и отдельными станциями и обсерваториями, которые осуществляют мониторинг параметров магнитного поля Земли и его вариаций.

Обозначим совокупность источников геомагнитных данных как D. Тогда распределенные средства сбора и хранения разнородных геомагнитных данных можно представить, как:

$$D = \langle M_n, M_s, M_o \rangle : \forall s \in M_s : s \notin \{M_n\};\ \forall o \in M_o : o \notin \{M_n\}; s \neq o,$$

где M_n — множество магнитных сетей, M_s — множество автономных вариационных станций, M_o — множество автономных магнитных обсерваторий.

Отметим, что справедливо соотношение:

$$\exists\, d \in \{M_n\} : d \in M_{n_i}, d \in M_{n_j}, i \neq j,$$

где d — произвольный экземпляр из множества магнитных сетей M_n.

Важно, что поставщики геомагнитных данных предоставляют специализированные веб-сервисы, которые по стандартным инфокоммуникационным протоколам осуществляют передачу данных потребителям, но не обеспечивают их предварительную обработку. В результате конечный информационный продукт представляет собой набор неполных временных рядов с пропусками и аномалиями, содержащих разноформатные результаты неоднородных геомагнитных наблюдений, проведенных с различной частотой (в среднем — от нескольких секунд до минуты).

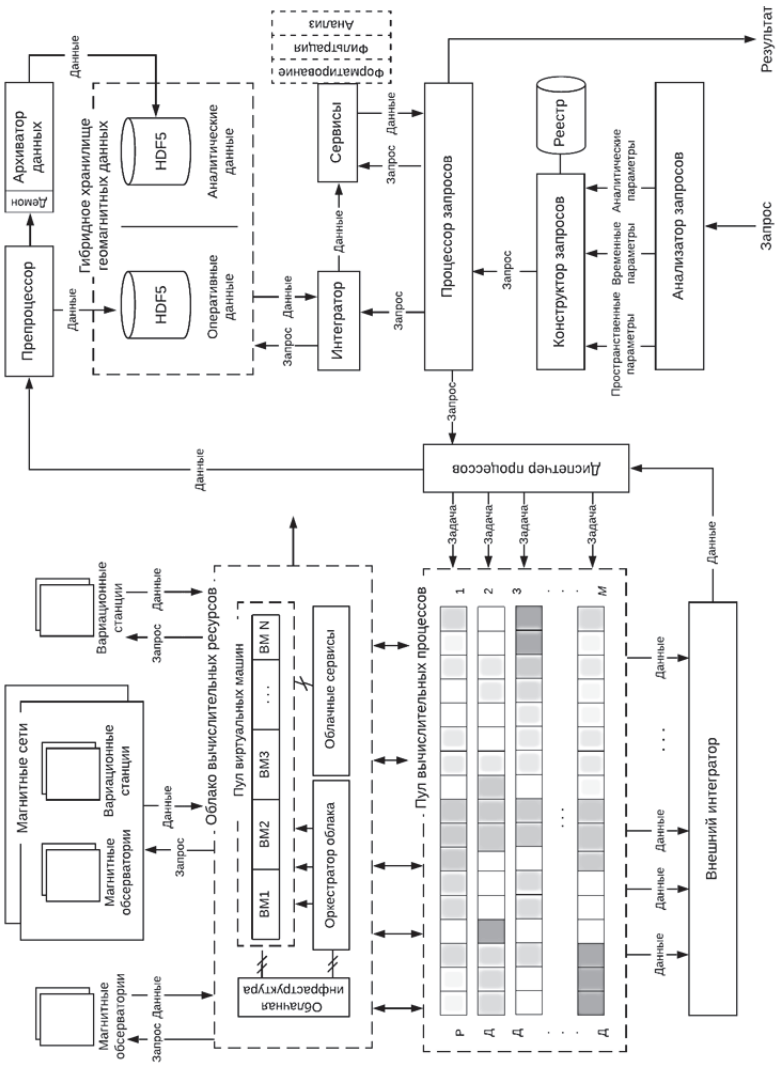


Рис. 1. Концепция единого пространства геомагнитных данных

Все доступные в едином пространстве поставщики геомагнитных данных обозначены в специальном информационном разделе — реестре R, содержащем URL доступа к соответствующим сервисам данных:

$$R = \{R_i\}, i = \overline{1,n}; n = (l_1 - l_2) + l_3 + l_4,$$
$$l_1 = \|M_n\|; l_2 = \|M_n^*\|, M_n^* = \cap M_n; l_3 = \|M_s\|; l_4 = \|M_o\|.$$

Следующий компонент единого пространства – гибридное хранилище геомагнитных данных H, сочетающее в себе физически хранимые аналитические данные A и размещенные в виртуальном кэше оперативные данные B, прошедшие предварительную обработку:

$$H = A \cup B.$$

В зависимости от параметров запроса результат формируется как комбинированием данных обоих хранилищ, так и применением к ним фильтрации по пространственно-временным признакам.

Формирование как содержимого хранилища, так и ответа на запрос к нему осуществляется связанной модульной структурой M единого пространства геомагнитных данных:

$$M = \{M_i\}, i = \overline{1,8}.$$

Всего в архитектуре единого пространства геомагнитных данных предложены восемь модулей, каждый из которых обеспечивает определенные этапы сбора, обработки и анализа гетерогенных данных. Так, анализатор запросов декомпозирует полученное от потребителя сообщение на набор параметров P, выделяя среди них пространственные, временные и аналитические:

$$P = \{S, T, A\},$$

где S — множество пространственных параметров (географические координаты, перечень станций, обсерваторий и пр.), T — множество временных параметров (временной промежуток, за который требуется получить данные), A — множество аналитических параметров (при необходимости — вид применяемого анализа (Фурье-преобразование, частотный, амплитудный и пр.)).

На следующем шаге выделенные параметры передаются конструктору запросов, формирующему тексты команд T для обраще-

ния к разделам гибридного хранилища геомагнитных данных, в том числе предварительно из реестра извлекаются локаторы для требуемых источников:

$$C = U \cup T,$$

где U — множество локаторов источников на основании множества S.

Основным компонентом модульной структуры единого пространства геомагнитных данных является процессор запросов, который на программном уровне оперирует текстами полученных от конструктора команд и передает их выполнение соответствующим программным интерпретаторам. При этом направление передачи команд от процессора к интерпретаторам зависит от типа запроса.

Так, если запрос предполагает обращение только к аналитическим данным, то управление получает интегратор, объединяющий наборы результирующих и разреженных по параметрам данных. Интегратор, в свою очередь, может передать полученный массив аналитическим сервисам, обеспечивающим возможность форматирования данных (приведение к одному из поддерживаемых форматов, в том числе IAGA2002), их фильтрации, а также аналитической обработки. Полученный в результате набор данных возвращает управление процессору запросов для дальнейшей работы с потребителем.

В случае, если выполнение запроса предполагает обращение к исходным распределенным источникам данных (на основании выборки локаторов из реестра), то процессор передает управление диспетчеру процессов, декомпозирующему исполняемую процедуру выборки на множество параллельно исполняемых подпроцессов. Результирующие данные поступают обратно в диспетчер процессов из внешнего интегратора, получающего управление по завершению соответствующих вычислительных подпроцессов и объединяющего полученные в ходе их выполнения результаты.

На следующем шаге полученные геомагнитные данные передаются в препроцессор для прохождения процедуры предварительной обработки — обнаружения выбросов и аномалий, восстановления временного ряда, форматирования и прочее. Если массив данных содержит наблюдения параметров геомагнитного поля только за текущие сутки, то он передается в виртуальный кэш гибридного хранилища для последующего его использования интегратором и аналитическими сервисами. В противном случае данные передаются в специальный модуль — архиватор. Особенностью последнего является то, что он выполняется процессом-демоном, автоматически инициирующим процедуру заполнения хранилища данными по истечении суток. При этом оперативные данные

формируются по запросу, заданному не только потребителем геомагнитных данных, но и самим процессом-демоном (работа процесса ведется по принципу классического планировщика CronTab), который, в свою очередь, эмулирует запрос на получение геомагнитных данных от всех доступных источников за текущие сутки.

Еще один компонент в структуре единого пространства геомагнитных данных играет решающую роль в обеспечении высокой вычислительной скорости обращения к распределенным источникам геомагнитных данных. Таким компонентом является облако вычислительных ресурсов O, предоставляющее среду выполнения программного кода на базе модели PaaS (Platform as a Service, платформа как сервис) [20, 21]. Облачная инфраструктура единого пространства геомагнитных данных обслуживает ряд виртуальных машин, каждой из которых выделяется один или несколько заданных вычислительных процессов по получению и обработке массивов данных. При этом задачи управления виртуальными машинами, их диспетчеризации, распределения аппаратных и программных ресурсов решает специализированный модуль — оркестратор облака.

Одновременное завершение всех параллельных вычислительных процессов не является обязательным, но каждый из них формирует одну часть результирующего набора разнородных геомагнитных данных, который после дополнительной обработки унифицируется и передается во внешний интегратор, формирующий данные для гибридного хранилища (как аналитические, так и оперативные). Таким образом, время ожидания результирующего набора данных соответствует максимальному времени, отводимому на выполнение процессам из вычислительного пула облачной инфраструктуры. При этом отметим, что число виртуальных машин, как правило, меньше выделенных вычислительных процессов, поэтому требуется диспетчеризация и повторное использование ресурсов, что осуществляется оркестратором облачной инфраструктуры.

6. Технологии единого пространства геомагнитных данных. С технической точки зрения единое пространство геомагнитных данных представляет собой распределенную информационную систему, обеспечивающую обработку, анализ и визуализацию больших объемов гетерогенных геомагнитных данных. Однако реализация такой системы имеет ряд нюансов, связанных со спецификой как самих данных, так и их источников.

Требование доступности единого пространства геомагнитных данных широкому кругу потребителей обуславливает необходимость разработки веб-ориентированной информационной среды как промежуточного звена между распределенными источниками данных

и пользователями. Традиционная клиент-серверная архитектура в сочетании с современными технологиями веб-программирования и дизайна (PWA, progressive web applications) позволяет создать независимое от платформы и простое в использовании приложение, доступное пользователям, независимо от типа устройства. Дополнительно к этому большой объем данных, сложность аналитических операций, необходимость доступного представления потребителям результатов обработки геомагнитных данных и обеспечение высокой вычислительной скорости в веб-среде обуславливают выбор в пользу некомпилируемого расширяемого языка программирования, текст на котором можно выполнить на виртуальной машине со специально выделенными областями памяти, потоков и стека.

Далее в рамках предложенной концепции для решения проблемы высоких аппаратных требований к организации гибридного хранилища предлагается хранить геомагнитные данные, которые прошли предварительную обработку, в иерархическом бинарном формате HDF5 (hierarchical data format), специально разработанном для больших объемов цифровой информации. Структура документов HDF5 схожа с иерархической файловой системой, при этом для доступа к данным применяются пути, сформированные на основании POSIX-синтаксиса, а метаданные задаются в виде набора именованных атрибутов соответствующих объектов [22]. Применяемые алгоритмы сжатия сокращают физический объем требуемого для хранения данных дискового пространства, а широкий спектр инструментально-программных средств обработки формата позволяет существенно повысить вычислительную скорость выполнения основных операций над данными. Указанные преимущества являются доводом в пользу выбора формата HDF5 для организации хранилища геомагнитных данных.

Кроме того, с целью повышения вычислительной скорости получения геомагнитных данных из гетерогенных источников предлагается введение виртуального вычислительного кластера на базе модели распределенных вычислений MapReduce и принципов многопроцессности. Однако реализация предлагаемого подхода на веб-платформе технически невозможна из-за использования здесь концепции мьютекса GIL (Global Interpreter Lock). В этой связи предлагается использовать принцип многопоточности, разделив главный узел на множество параллельно выполняемых независимых потоков, число которых кратно вычислительной мощности серверов, обеспечивающих работу единого пространства геомагнитных данных. Тогда вычислительная нагрузка будет распределена между всеми доступными (или их большинством) ядрами процессора, а параллельно выполняемые процессы не замедлят работу друг друга и в совокупности приведут к увеличе-

нию вычислительной скорости работы с большими объемами геомагнитных данных. Предлагаемая схема многопоточной обработки больших объемов геомагнитных данных приведена на рисунке 2.

Показано, что вычислительный процесс декомпозируется на множество групп, состоящих из n потоков. Диспетчер процессов присваивает каждому из них идентификатор, направляя его значение в модуль управления ресурсами потоков — мьютекс. Здесь потоки формируют очередь, при этом каждый из них гарантированно получает управление и свой набор вычислительных ресурсов — состояние потока, которое, в свою очередь, выделяется в результате взаимодействия мьютекса, диспетчера процессов и пула вычислительных ресурсов. Аналогичным образом происходит освобождение ресурсов: завершивший выполнение поток передает идентификатор мьютексу, который направляет его диспетчеру, освобождающему соответствующее состояние потока и формирующему новое для следующего в очереди.

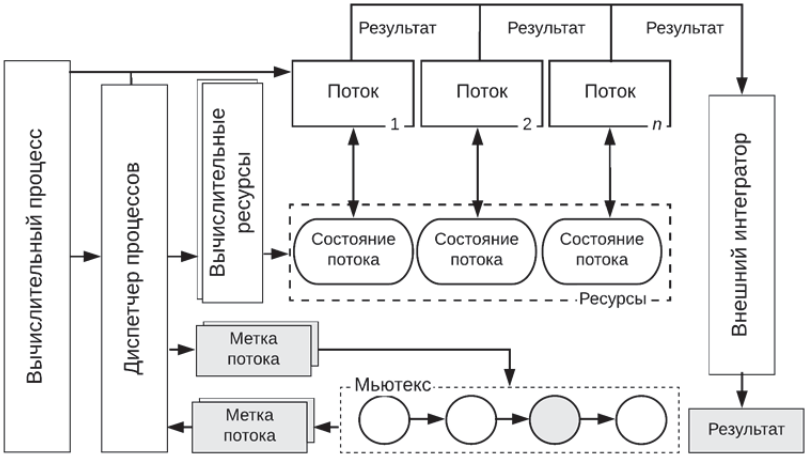


Рис. 2. Схема многопоточной загрузки больших объемов геомагнитных данных

Важно отметить, что параллельно выполняемые вычислительные потоки гораздо менее требовательны к ресурсам, чем параллельно выполняемые вычислительные процессы, и с ними гораздо практичнее выполнять в однопроцессорных системах программы, логика которых требует применения нескольких потоков исполнения. Большинство современных хостингов позволяют разделить таким образом вычислительный процесс в среднем на полтора десятка

потоков и, соответственно, увеличить вычислительную скорость в то же количество раз.

Однако, если выйти за пределы возможностей хостинга и использовать облачную инфраструктуру, то предоставляемые при этом вычислительные скорости и мощности можно дополнительно увеличить. В этом случае вычислительные потоки должны быть равномерно распределены между доступными виртуальными машинами облака, образующими виртуальный вычислительный кластер и также реализующими модель MapReduce [23].

В основе облачной платформы лежит концепция виртуализации, которая обеспечивает предоставление пользователям и системам набора вычислительных ресурсов, абстрагированное от аппаратной реализации. Физические ресурсы и сервисы облака объединяются, и из них оркестратором выделяется пул виртуальных машин, где выполняются гостевые вычислительные процессы. Почасовая тарификация, возможность конфигурирования используемых виртуальных машин (число и степень загрузки ядер процессора) и количество выделяемой памяти позволяют гибко настраивать вычислительные возможности облачной инфраструктуры. В результате число параллельно исполняемых потоков будет кратно вычислительной мощности арендуемой облачной инфраструктуры (количеству виртуальных машин и ядер в них).

7. Пример реализации. На текущем этапе исследований разработан прототип веб-ориентированной информационной системы, реализующей концепцию единого пространства геомагнитных данных на основе инфокоммуникационных технологий и геоинформационного моделирования и включающей ряд сервисов обработки и визуализации геомагнитных данных.

Так, в составе информационной системы реализован сервис «Геомагнитный калькулятор» (рисунок 3), осуществляющий расчет параметров невозмущенного геомагнитного поля в точке с заданными пространственно-временными координатами [24, 25].

Сервис доступен по адресу https://www.geomagnet.ru/geomagnetic_calculator.html и предоставляет пользователям возможность рассчитать следующие параметры для заданной географической точки и временной метки:

– геомагнитные координаты;

– параметры геомагнитного диполя: координаты северного магнитного полюса (в градусах) и магнитный момент (Тл · $м^3$);

– параметры геомагнитного поля: магнитная индукция (нТл · км), компоненты и полный вектор поля (нТл), магнитное склонение и магнитное наклонение (градусы).

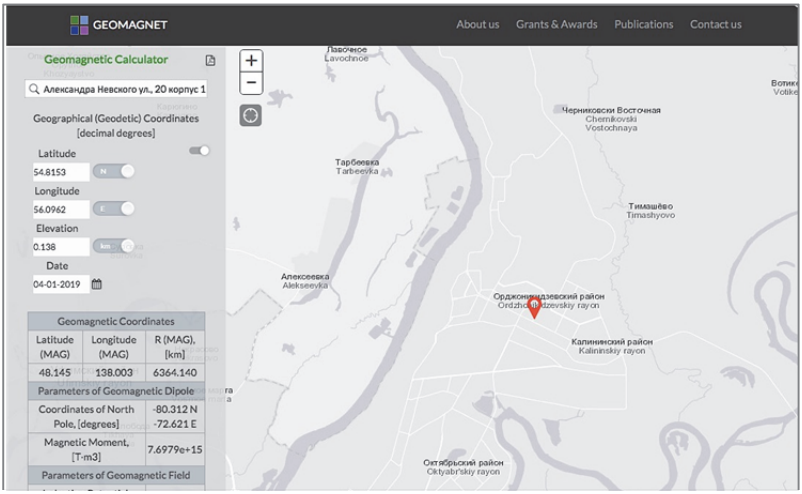


Рис. 3. Экранная форма сервиса «Геомагнитный калькулятор»

Сервис трехмерной визуализации обеспечивает графическую интерпретацию геомагнитных данных и характеризует распределение параметров главного геомагнитного поля по земной поверхности посредством комбинируемых линий уровня (изолиний) и сплайнов, расположенных на интерактивном виртуальном глобусе (рисунок 4).

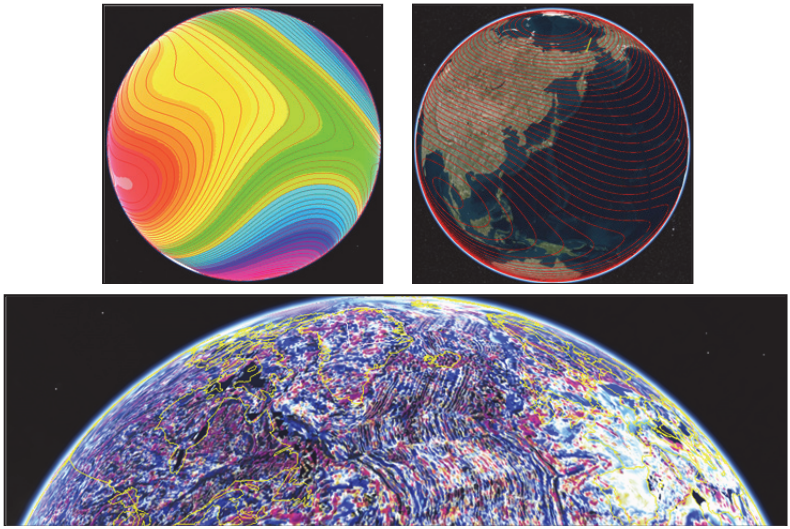


Рис. 4. Экранная форма сервиса трехмерной визуализации геомагнитных данных

Пользователь взаимодействует как непосредственно с трехмерным представлением земной поверхности, используя средства масштабирования и инструментарий геоинформационных технологий (геолокация, прямое и обратное геокодирование и пр.), так и с результатом визуализации параметров главного поля посредством меню различных уровней, ассоциированных с соответствующими графическими элементами (например, каждой линии уровня ставится в соответствие контекстное меню, отражающее значения параметра магнитного поля) [24].

Отличительной особенностью представленной визуализации геомагнитных данных является возможность простого переключения между режимами двух- (2D), трех- (3D) и псевдотрехмерного изображения (2.5D) посредством соответствующего элемента управления, доступного на веб-странице. При этом результаты графической интерпретации геомагнитных данных автоматически адаптируются под выбранный режим представления, что расширяет возможности пользователя при проведении их визуального анализа [24, 25].

Сервис «Магнитные обсерватории» предназначен для предоставления пользователям доступа к результатам наблюдений за параметрами геомагнитного поля и его вариаций, выполняемых одной или несколькими магнитными станциями и обсерваториями.

Особенностью предлагаемого технического решения является информационное обеспечение системы, представленное полученными из распределенных источников геомагнитными данными и регулярно актуализируемое с помощью настраиваемых процессов-демонов. С точки зрения функциональности, сервис обеспечивает возможность поиска магнитных станций и обсерваторий, их фильтрацию, анализ, отображение и вывод доступных данных за указанный период.

На рисунке 5 приведен пример работы информационной системы, демонстрирующий способ выбора магнитной обсерватории с учетом ее пространственной привязки и результаты амплитудно-временного анализа соответствующих геомагнитных данных, полученных по результатам семидневного наблюдения параметров геомагнитного поля и его вариаций.

Применение геоинформационных технологий в рамках данного сервиса позволяет предоставить пользователю несколько альтернативных способов выбора магнитной станции или обсерватории для анализа геомагнитных данных:

– прямое геокодирование: указание значений пространственных координат искомой станции. При этом координаты могут быть указаны в соответствующих полях ввода на экранной форме информационной системы или выбраны путем позиционирования пользователем курсора на виртуальной карте;

– обратное геокодирование: указание наименования, аббревиатуры станции или адреса ее местоположения, которые посредством специализированных веб-ориентированных сценариев преобразуются в наборы пространственных координат. Сопоставление полнотекстовых поисковых запросов осуществляется как с серверной базой пространственных данных, так и с реестром источников геомагнитных данных, предусмотренном предложенной структурой единого пространства геомагнитных данных;

– геолокация: определение реального географического положения электронного устройства, с которого осуществляется выход в Сеть. Для реализации данной функциональности в структуре информационной системы предусмотрен механизм выбора станции или обсерватории, ближайшей к точке, зарегистрированной системой позиционирования.

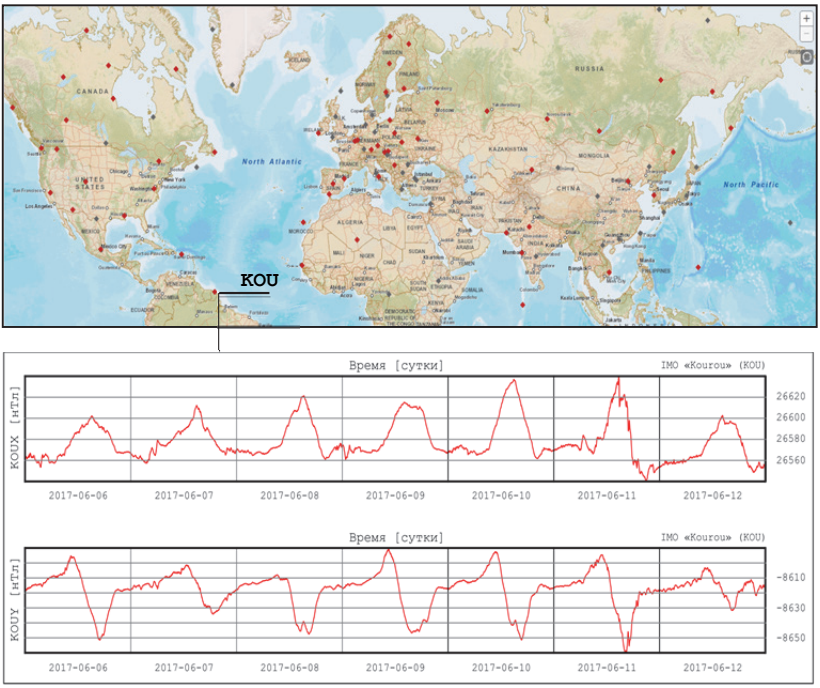


Рис. 5. Экранная форма сервиса «Магнитные обсерватории» по выбору магнитной станции и ее данных

Для выбранного источника геомагнитных данных информационной системой формируется график амплитудно-временного анализа

результатов наблюдений, по умолчанию ориентированный на неделю от заданной временной метки. Временной период можно указать, воспользовавшись соответствующими полями ввода на экранной форме, нагруженными функцией элемента-календаря.

Кроме этого, сервис включает в себя набор инструментов цифровой обработки геомагнитных данных, основанных на классических механизмах цифровой обработки сигналов (ЦОС). Среди них:

– линейная, нелинейная и адаптивная фильтрация информационного сигнала, обеспечивающая как подавление шума и селекцию сигнала в частотной области, так и анализ корреляций параметров смежных информационных сигналов;

– анализ сигнала во временной области, обеспечивающий расчет таких параметров информационного сигнала, как максимальное, минимальное и среднее значения, а также дисперсия и среднеквадратическое отклонение;

– спектральный и частотно-временной анализ, реализуемый как посредством исследования периодограммы, то есть за счет оценки спектральной плотности мощности, основанной на вычислении квадрата модуля преобразования Фурье последовательности данных с использованием статистического усреднения информационного сигнала, так и посредством оценки его вейвлет-скалограммы.

Указанные информационные сервисы реализованы посредством описанного выше набора технологий. Сервис «Магнитный калькулятор» размещен на внешнем сервере и находится на этапе тестирования. Остальные сервисы реализованы в виде исследовательского прототипа и проходят стадию отладки и тестирования на локальном виртуальном сервере.

Проведенные экспериментальные исследования результата реализации предложенной концепции единого пространства геомагнитных данных с применением рассмотренного набора технологий (п. 5) показали, что:

– скорость обработки больших объемов геомагнитных данных за счет применения многопоточного режима увеличивается в число раз, кратное вычислительной мощности сервера;

– аппаратные требования информационной системы в контексте размеров дискового пространства, физически занимаемого гибридным хранилищем, существенно сокращены (более чем в 2,3 раза по сравнению с CSV-представлением) за счет использования иерархического формата больших данных HDF5.

7. Заключение. Всевозрастающая потребность в своевременных достоверных геомагнитных данных предъявляет новые требования

к техническим средствам их сбора, обработки, передачи и анализа. Существующие сервисы лишь частично решают такие задачи, фокусируясь на отдельных магнитных сетях, обсерваториях или станциях. Вместе с тем реалии таковы, что для решения научно-практических задач необходима интеграция геомагнитных данных, получаемых из множества гетерогенных распределенных источников. Кроме того, проблема усугубляется непрерывно возрастающими объемами геомагнитной информации и обнаруживает низкую эффективность традиционных методов и средств их хранения, обработки и анализа.

В статье предложена концепция единого пространства геомагнитных данных, основанная на сочетании моделей консолидации и федерализации при интеграции данных и отличающаяся наличием гибридного информационного хранилища, заполняемого данными по выполнению процесса-демона, предварительной обработкой для повышения качества хранимых временных рядов, иерархическим форматом для снижения вычислительных затрат на размещение данных, модульной структурой и облачными вычислениями. Применение указанной концепции позволит:

1. Обеспечить конечным пользователям единый доступ к геомагнитным данным, предоставляемым гетерогенными источниками, без необходимости обращения к несвязанным информационным ресурсам, а также без избыточной предварительной загрузки разнородных данных с их последующей интеграцией для выполнения аналитических операций. Известные решения в области обработки геомагнитных данных лишь частично решают такую задачу, например, в рамках проекта SuperMag доступны только данные вариационных станций, а информационные ресурсы отдельных станций, обсерваторий и магнитных сетей предоставляют результаты суточных геомагнитных наблюдений, физически представленных в отдельных текстовых файлах.

2. Формировать и хранить полные временные ряды геомагнитных данных, не содержащие аномальных значений. При этом использование ранее предложенного авторами метода обеспечивает ошибку восстановления временного ряда в пределах погрешности геомагнитных измерений, определенной международной ассоциацией IAGA. Существующие решения по восстановлению временных рядов геомагнитных данных не обеспечивают приемлемого значения ошибки и применимы только для устранения единичных пропусков. Кроме того, один из немногих информационных проектов по геомагнитным наблюдениям SuperMag предусматривает удаление обнаруженных во временных рядах аномальных значений, что негативно сказывается на результатах аналитической обработки геомагнитных данных.

3. Сократить вычислительные затраты на хранение и обработку больших объемов геомагнитных данных за счет использования иерархического формата HDF5 и технологий облачных вычислений. В настоящее время массивы геомагнитных данных хранятся в виде набора текстовых файлов и доступны, как правило, по протоколу FTP. Специфика протокола и физическое разделение результатов наблюдений обуславливают низкую вычислительную скорость их загрузки и обработки. Так, к примеру, получение результатов годовых геомагнитных наблюдений с ресурса сети INTERMAGNET и их последующая аналитическая обработка требуют в среднем 4.5 минуты процессорного времени, что с учетом требований к эргономике программного обеспечения (время ожидания отклика не должно превышать 3-5 с) неприемлемо. Предложенная архитектура и стек технологий успешно решают поставленную задачу, сокращая затраты процессорного времени до допустимых пределов.

Перечисленное в совокупности позволит повысить как доступность разнородных результатов наблюдений за параметрами магнитного поля Земли и его вариаций, так и вычислительную скорость сбора и обработки больших объемов геомагнитных данных.

Предложен набор инфокоммуникационных технологий, обеспечивающих эффективную реализацию предложенной концепции. Ожидается, что веб-ориентированная реализация (клиент-серверные приложения по принципам PWA) концепции единого пространства существенно расширит круг потребителей геомагнитных данных, а вычислительные технологии, предназначенные для хранения и обработки больших объемов данных (формат HDF5, облачная инфраструктура и вычисления), позволят обеспечить высокую скорость и сравнительно небольшие вычислительные затраты на выполнение процедур интеграции разнородных геомагнитных данных по запросу потребителя. Кроме того, предлагаемый набор сервисов обеспечит потребителей новыми результатами, позволяющими получить дополнительную информацию из геомагнитных данных без применения дополнительных сторонних средств и систем.

Разработан исследовательский прототип информационной системы, включающий в себя ряд веб-ориентированных сервисов, которые обеспечивают:

– расчет параметров невозмущенного геомагнитного поля в точке с пространственно-временными координатами, заданными посредством геоинформационных технологий геолокации, прямого и обратного геокодирования;

– визуализацию геомагнитных данных с использованием платформонезависимой программируемой веб-ориентированной графики, обеспечивающей возможность переключения между двумерным (2D),

псевдотрехмерным (2.5D) и трехмерным (3D) форматами графического представления распределения параметров геомагнитного поля по земной поверхности;

– получение и амплитудно-временной анализ временных рядов геомагнитных данных, регистрируемых одной или несколькими магнитными обсерваториями и вариационными станциями на протяжении заданного временного интервала.

Ряд сервисов информационной системы размещен на внешнем веб-сервере с открытым физическим доступом, а остальные проходят стадии интеграционного тестирования и отладки на локальном виртуальном сервере. Дальнейшие исследования предполагают полное размещение сервисов на внешнем ресурсе и последующее развитие функциональности, заявленной при описании предложенной структуры единого пространства геомагнитных данных.

Литература

1. *Воробьев А.В., Воробьева Г.Р.* Подход к оценке относительной информационной эффективности магнитных обсерваторий сети INTERMAGNET // Геомагнетизм и аэрономия. 2018. Т. 58. № 5. С. 648–652.
2. *St-Louis B.J. et al.* Intermagnet Technical Reference Manual // The INTERMAGNET office. Retrieved from the World Wide Web. 2011. 100 p.
3. *Love J.J., Chulliat A.* An International Network of Magnetic Observatories // EOS Transactions: American Geophysical Union. 2013. vol. 94. no. 42. pp. 373–374.
4. *Sandhu J.K. et al.* Variations of high latitude geomagnetic pulsation frequencies: A comparison of time of flight estimates and IMAGE magnetometer observations // Journal of Geophysical Research Space Physics. 2018. vol. 123. no. 1. pp. 567–586.
5. *Connors M. et al.* The AUTUMNX magnetometer meridian chain in Québec, Canada // Earth, Planets and Space. 2016. vol. 68. no. 1. pp. 2.
6. *Hughes W.J., Engebretson M.J.* MACCS: Magnetometer array for cusp and cleft studies // Satellite – Ground Based Coordination Sourcebook. 1997. vol. 1198. pp. 119.
7. *Reay S.J. et al.* Magnetic Observatory Data and Metadata: Types and Availability // Geomagnetic Observations and Models. 2011. pp. 149–181.
8. *Хомутов С.Ю.* Обработка магнитных данных на обсерваториях // ИКИР ДВО РАН. 2017. 114 с.
9. *Анисимов С.В. и др.* Информационные технологии в геомагнитных измерениях на геофизической обсерватории Борок // Геофизические исследования. 2008. Т. 9. № 3. С. 62–76.
10. *Золотов С.Ю., Додолин Е.Л.* Методологические основы распределенной информационной системы мониторинга состояния окружающей среды // Доклады ТУСУРа. 2010. № 1(21). С. 203–206.
11. *Кокорев В.А., Шерстюков А.Б.* О метеорологических данных для изучения современных и будущих изменений климата на территории России // Арктика XXI век. Естественные науки. 2015. № 2. С. 5–23.
12. *Соловьев А.А. и др.* Новая геомагнитная обсерватория «Климовская» // Геомагнетизм и аэрономия. 2016. Т. 56. № 3. С. 342–354.
13. *Gvishiani A., Soloviev A.* Geoinformatic advances in geomagnetic data studies and Russian INTERMAGNET segment // Исследования по геоинформатике: труды Геофизического центра РАН. 2016. Т. 4. № 2. С. 8.
14. *Mandea M, Korte M.* Geomagnetic Observations and Models // IAGA Special Sopron Book Series. 2011. vol. 5. pp. 149–181.

15. *Dods J., Chapman S.C., Gjerloev J.W.* Network analysis of geomagnetic substorms using the SuperMAG database of ground-based magnetometer stations // Journal of Geophysical Research-Space Physics. 2015. vol. 120. pp. 7774–7784.
16. Концепция формирования и развития единого информационного пространства России и соответствующих государственных информационных ресурсов. URL: http://www.nsc.ru/win/laws/russ_kon.htm (дата обращения: 04.01.2019).
17. *Куваев В.О. и др.* Варианты построения единого информационного пространства для интеграции разнородных автоматизированных систем // Информация и космос. 2015. № 4. С. 83–87.
18. Методы интеграции данных в информационных системах. URL: http://www.ipr-ras.ru/articles/kogalov10-05.pdf (дата обращения: 04.01.2019).
19. *Воробьев А.В., Воробьева Г.Р.* Индуктивный метод восстановления временных рядов геомагнитных данных // Труды СПИИРАН. 2018. № 57. С. 104–133.
20. *Teixeira C. et al.* The Building Blocks of a PaaS // Journal of Network and Systems Management. 2014. vol. 22. pp. 75.
21. *Van Eyk E. et al.* Serverless is More: From PaaS to Present Cloud Computing // IEEE Internet Computing. 2018. vol. 22. no. 5. pp. 8–17.
22. The HDF Group. Hierarchical Data Format, version 5. URL: http://www.hdfgroup.org/HDF5/ (дата обращения: 04.01.2019).
23. *Miyazaki R., Matsuzaki K., Sato D.* A Generator of Hadoop MapReduce Programs that Manipulate One-dimensional Arrays // Journal of Information Process. 2017. vol. 25. pp. 841–851.
24. *Воробьев А.В., Воробьева Г.Р.* Веб-ориентированная 2D/3D-визуализация параметров геомагнитного поля и его вариаций // Научная визуализация. 2017. Т. 9. № 2. С. 94–101.
25. *Yusupova N. et al.* Web-based solutions in modeling and analysis of geomagnetic field and its variations // Proceedings of REMS 2018 – Russian Federation & Europe Multidisciplinary Symposium on Computer Science and ICT. URL: http://ceur-ws.org/Vol-2254/10000282.pdf. (дата обращения: 04.01.2019).

Воробьев Андрей Владимирович — канд. техн. наук, доцент, кафедра геоинформационных систем факультета информатики и робототехники, Уфимский государственный авиационный технический университет (УГАТУ). Область научных интересов: геоинформационные технологии, цифровая обработка сигналов. Число научных публикаций — 147. geomagnet@list.ru; К.Маркса, 12, 450007, Уфа, Российская Федерация; р.т.: +7(917)345-2299.

Воробьева Гульнара Равилевна — канд. техн. наук, доцент, кафедра вычислительной математики и кибернетики факультета информатики и робототехники, Уфимский государственный авиационный технический университет (УГАТУ). Область научных интересов: геоинформационные и веб-технологии, системы хранения и обработки информации. Число научных публикаций — 121. gulnara.vorobeva@gmail.com; К. Маркса, 12, 450008, Уфа, Российская Федерация; р.т.: +7(917)417-411.

Юсупова Нафиса Исламовна — д-р техн. наук, профессор, заведующий кафедрой, кафедра вычислительной математики и кибернетики факультета информатики и робототехники, Уфимский государственный авиационный технический университет (УГАТУ). Область научных интересов: интеллектуальные методы обработки информации и управления с приложениями в социальных, экономических и технических системах. Число научных публикаций — 566. yussupova@ugatu.ac.ru; К.Маркса, 12, 450007, Уфа, Российская Федерация; р.т.: +7 (347)273-7717.

Поддержка исследований. Работа частично выполнена при финансовой поддержке РФФИ (проект № 18-07-00193-а).

DOI 10.15622/sp.18.2.390-415

A.V. VOROBEV, G.R. VOROBEVA, N.I. YUSUPOVA

CONCEPTION OF GEOMAGNETIC DATA INTEGRATED SPACE

Vorobev A.V., Vorobeva G.R., Yusupova N.I. **Conception of Geomagnetic Data Integrated Space.**

Abstract. As is known, today the problem of geomagnetic field and its variations parameters monitoring is solved mainly by a network of magnetic observatories and variational stations, but a significant obstacle in the processing and analysis of the data thus obtained, along with their spatial anisotropy, are omissions or reliable inconsistency with the established format. Heterogeneity and anomalousness of the data excludes (significantly complicates) the possibility of their automatic integration and the application of frequency analysis tools to them. Known solutions for the integration of heterogeneous geomagnetic data are mainly based on the consolidation model and only partially solve the problem. The resulting data sets, as a rule, do not meet the requirements for real-time information systems, may include outliers, and omissions in the time series of geomagnetic data are eliminated by excluding missing or anomalous values from the final sample, which can obviously lead to both to the loss of relevant information, violation of the discretization step, and to heterogeneity of the time series. The paper proposes an approach to creating an integrated space of geomagnetic data based on a combination of consolidation and federalization models, including preliminary processing of the original time series with an optionally available procedure for their recovery and verification, focused on the use of cloud computing technologies and hierarchical format and processing speed of large amounts of data and, as a result, providing users with better and more homogeneous data.

Keywords: Geomagnetic Data, Magnetic Observatories, Time Series, Big Data, Integrated Information Space, Parallel Computing.

Vorobev Andrei Vladimirovich — Ph.D., Associate Professor, Geoinformation Systems Department of Computer Science and Robotics Faculty, Ufa State Aviation Technical University (USATU). Research interests: geoinformation technologies, digital signal processing. The number of publications — 147. geomagnet@list.ru; 12, K. Marx, 450007, Ufa, Russian Federation; office phone: +7(917)345-2299.

Vorobeva Gulnara Ravilevna — Associate Professor, Computational Mathematics and Cybernetics Department of Computer Science and Robotics Faculty, Ufa State Aviation Technical University (USATU). Research interests: geoinformation and web technologies, systems of information storing and processing. The number of publications — 121. gulnara.vorobeva@gmail.com; 12, K. Marx, 450008, Ufa, Russian Federation; office phone: +7(917)417-411.

Yusupova Nafisa Islamovna — Ph.D., Dr.Sci., Professor, Head of Department, Computational Mathematics and Cybernetics Department of Computer Science and Robotics Faculty, Ufa State Aviation Technical University (USATU). Research interests: intelligent methods of information processing and management with applications in social, economic and technical systems. The number of publications — 566. yussupova@ugatu.ac.ru; 12, K.Marx, 450007, Ufa, Russian Federation; office phone: +7 (347)273-7717.

Acknowledgements. This research is partially supported by RFBR (grant 18-07-00193-a).

References

1. Vorobev A.V., Vorobeva G.R. [Approach to Assessment of the Relative Informational Efficiency of Intermagnet Magnetic Observatories]. *Geomagnetizm i aeronomia – Geomagnetism and Aeronomy.* 2018. Issue 58. vol. 5. pp. 648–652. (In Russ.).

2. St-Louis B.J. et al. Intermagnet Technical Reference Manual. The INTERMAGNET office. Retrieved from the World Wide Web. 2011. 100 p.
3. Love J.J., Chulliat A. An International Network of Magnetic Observatories // *EOS Transactions: American Geophysical Union*. 2013. vol. 94. no. 42. pp. 373–374.
4. Sandhu J.K. et al. Variations of high latitude geomagnetic pulsation frequencies: A comparison of time of flight estimates and IMAGE magnetometer observations. *Journal of Geophysical Research Space Physics.* 2018. vol. 123. no. 1. pp. 567–586.
5. Connors M. et al. The AUTUMNX magnetometer meridian chain in Québec, Canada. *Earth, Planets and Space*. 2016. vol. 68. no. 1. pp. 2.
6. Hughes W.J., Engebretson M.J. MACCS: Magnetometer array for cusp and cleft studies. Satellite – Ground Based Coordination Sourcebook. 1997. vol. 1198. pp. 119.
7. Reay S.J. et al. Magnetic Observatory Data and Metadata: Types and Availability. Geomagnetic Observations and Models. 2011. pp. 149–181.
8. Khomutov S.Yu. *Obrabotka magnitnyh dannyh na observatorijah* [Magnetic data processing at observatories]. IKIR FEB RAS. 2017. 114 p. (In Russ.).
9. Anisimov S.V. et al. [Information technology in geomagnetic measurements at the Borok geophysical observatory]. *Geofizicheskie issledovanija – Geophysical research.* 2008. vol. 9. no. 3. pp. 62–76. (In Russ.).
10. Zolotov S.Yu., Dodolin E.L. [Methodological basis of a distributed information system for monitoring the state of the environment]. *Doklady TUSURa – TUSUR reports.* 2010. vol. 1(21). pp. 203–206. (In Russ.).
11. Kokorev V.A., Sherstukov A.B. [About meteorological data to study current and future climate change in Russia]. *Arktika XXI vek. Yestestvennyye nauki – Arctic XXI century. Natural Sciences.* 2015. vol. 2. pp. 5–23. (In Russ.).
12. Soloviev A.A. et al. [New geomagnetic observatory "Klimovskaya"]. *Geomagnetizm i aeronomia – Geomagnetism and Aeronomy*. 2016. Issue 56. vol 3. pp. 342–354. (In Russ.).
13. Gvishiani A., Soloviev A. [Geoinformatic advances in geomagnetic data studies and Russian INTERMAGNET segment]. *Issledovanija po geoinformatike: trudy Geofizicheskogo centra RAN – Research in geoinformatics: the works of the Geophysical Center RAS 4.* vol. 2. pp. 8.
14. Mandea M, Korte M. Geomagnetic Observations and Models. IAGA Special Sopron Book Series. 2011. vol. 5. pp. 149–181.
15. Dods J., Chapman S.C., Gjerloev J.W. Network analysis of geomagnetic substorms using the SuperMAG database of ground-based magnetometer stations. *Journal of Geophysical Research-Space Physics.* 2015. vol. 120. pp. 7774–7784.
16. Koncepcija formirovanija i razvitija edinogo informacionnogo prostranstva Rossii i sootvetstvujushhih gosudarstvennyh informacionnyh resursov [The concept of the formation and development of a single information space of Russia and the relevant state information resources]. Available at: http://www.nsc.ru/win/laws/russ_kon.htm (accessed: 04.01.2019). (In Russ.).
17. Kuvaev V.O. et al. [Variants of building a single information space for the integration of heterogeneous automated systems]. *Informacija i kosmos – Information and space.* 2015. vol. 4. pp. 83–87. (In Russ.).
18. Metody integracii dannyh v informacionnyh sistemah [Methods of data integration in information systems]. Available at: http://www.ipr-ras.ru/articles/kogalov10-05.pdf (accessed: 04.01.2019). (In Russ.).
19. Vorobev A.V., Vorobeva G.R. [Inductive method of geomagnetic data time series recovery]. Trudy SPIIRAN – SPIIRAS Proceedings. 2018. vol. 57. pp. 104–133. (In Russ.).
20. Teixeira C. et al. The Building Blocks of a PaaS. *Journal of Network and Systems Management.* 2014. vol. 22. pp. 75.

21. Van Eyk E. et al. Serverless is More: From PaaS to Present Cloud Computing. *IEEE Internet Computing*. 2018. vol. 22. no. 5. pp. 8–17.
22. The HDF Group. Hierarchical Data Format, version 5. Available at: http://www.hdfgroup.org/HDF5/ (accessed: 04.01.2019).
23. Miyazaki R., Matsuzaki K., Sato D. A Generator of Hadoop MapReduce Programs that Manipulate One-dimensional Arrays. *Journal of Information Process*. 2017. vol. 25. pp. 841–851.
24. Vorobev A.V., Vorobeva G.R. [Web-based 2D / 3D visualization of geomagnetic field and its variations parameters]. *Nauchnaja vizualizacija – Scientific Visualization*. 2017. Issue 9. vol. 2. pp. 94–101. (In Russ.).
25. Yusupova N. et al. Web-based solutions in modeling and analysis of geomagnetic field and its variations. Proceedings of REMS 2018 – Russian Federation & Europe Multidisciplinary Symposium on Computer Science and ICT. Available at: http://ceur-ws.org/Vol-2254/10000282.pdf (accessed: 04.01.2019).

УДК 630*5:528.85 **DOI** 10.15622/sp.18.2.416-443

Д.М. Черниховский, А.С. Алексеев

ОПРЕДЕЛЕНИЕ СРЕДНИХ ВЫСОТ И ЗАПАСОВ ДРЕВОСТОЕВ НА ОСНОВЕ ОБРАБОТКИ ИНФОРМАЦИИ ТОПОГРАФИЧЕСКОЙ РАДАРНОЙ СЪЁМКИ, ЦИФРОВЫХ МОДЕЛЕЙ РЕЛЬЕФА И ГИС ТЕХНОЛОГИЙ

Черниховский Д.М., Алексеев А.С. **Определение средних высот и запасов древостоев на основе обработки информации топографической радарной съёмки, цифровых моделей рельефа и ГИС технологий.**

Аннотация. Рассмотрены возможности использования глобальных моделей высот рельефа SRTM (Shuttle radar topographic mission — радарная топографическая съемка) для оценки обобщенных характеристик лесных насаждений — средних высот и запасов. Известно, что при выполнении радарных съемок растительный покров препятствует корректному определению высоты земной поверхности. Поверхность, фиксируемая датчиками над покрытой лесом территорией (фаза рассеяния), располагается в верхней части древесного полога. Обзор публикаций подтверждает актуальность данного направления исследований в мире. На основе обзора литературы приведены краткие теоретические основы съемки SRTM, рассмотрены факторы, определяющие значения высот и связанные с ними ошибки, указана возможность определения высоты лесного полога на основе моделей высот.

В качестве модельной территории выбрана часть Учебно-опытного лесничества Ленинградской области. Исходными материалами для выполнения исследования служили геоинформационные базы данных лесоустройства, данные радарной съемки SRTM и топографические карты. Модельная территория разбита регулярной сетью на ячейки с шагом 1 км. Большая часть территории относится к площади, покрытой лесной растительностью. Моделирование рельефа выполнено на основе оцифрованных топографических карт масштаба 1:25000 путем интерполяции методом TIN. Выполнено визуальное и статистическое сравнение двух моделей высот — модели поверхности (на основе данных радарной съемки SRTM) и модели рельефа (на основе топографических карт). С помощью построения профилей выполнена оценка характера изменений высот моделей рельефа и поверхности. Отмечено, что для большей части модельной территории расхождения высот между моделями поверхности и рельефа составляют 15-20 м. Сближение графиков высот отмечается на участках, не покрытых лесной растительностью.

Получен набор линейных регрессионных зависимостей между средними высотами центра фазы рассеяния (независимая переменная) и средними высотами или запасами насаждений (зависимая переменная) в пределах границ лесотаксационных выделов, сгруппированных по преобладающим породам. Выявлено влияние на тесноту связи и значение коэффициента регрессии величины относительной полноты, коэффициента состава, преобладающей древесной породы.

Установленные закономерности могут использоваться в целях совершенствования теории и практики инвентаризации лесов, а также для решения иных задач, связанных с оценкой природных ресурсов на региональном и глобальном уровнях (национальная инвентаризация лесов, определение запасов углерода, оценка биомассы).

Ключевые слова: лесоустройство, цифровая модель рельефа, цифровая модель поверхности, высота центра фазы рассеяния, регрессионный анализ.

1. Введение. Необходимость развития дистанционных методов в лесном хозяйстве Российской Федерации определяется высоким спросом на информацию о лесах, колоссальными размерами покрытой лесом территории, слабо развитой лесной инфраструктурой и значительным

многообразием природных ландшафтов. Совершенствование характеристик материалов дистанционного зондирования и появление новых типов пространственных данных открывает новые возможности для управления природными ресурсами, в том числе для решения задач учета и управления лесами. Можно выделить несколько актуальных направлений развития и применения дистанционных методов в лесном хозяйстве: проведение фундаментальных исследований отображения характеристик лесов на материалах дистанционного зондирования [1-4]; картографирование растительности с изучением структуры и состояния растительных сообществ [5-9]; изучение процессов лесовосстановления [10]; оценка динамики природных и антропогенных ландшафтов [11, 12]; определение отдельных лесотаксационных характеристик [9, 13], оценка лесопатологического состояния [14]. Значительное количество исследований направлено на разработку методик практического применения современных материалов дистанционного зондирования для решения задач лесного хозяйства. Для решения задач инвентаризации лесов рассматриваются материалы аэрофотосъемки [3, 4, 15-17], космической съемки [1, 2, 5, 8, 9, 18], лидарной съемки [19-21]. Определенную нишу для исследований представляет изучение моделей высот лесного полога, получаемых на основе различных видов съемок [22-24].

Для изучения лесных ландшафтов на значительных по площади территориях представляется перспективным использование глобальных цифровых моделей высот (ЦМВ) свободного доступа. Преимуществами данных моделей высот являются их доступность (по сравнению с топографическими картами и высокоточными моделями высот, реализуемыми на коммерческой основе), приемлемая для решения многих задач точность, значительный (глобальный) пространственный охват и цифровая форма представления, удобная для обработки современными программными средствами. Глобальные ЦМВ строятся в основном по данным стереоскопической оптической и интерферометрической радиолокационной космической съёмок. Среди глобальных моделей высот (описывающих всю или почти всю поверхность земного шара) есть как бесплатные, находящиеся в свободном доступе в сети Интернет (GMTED 2010, ASTER GDEM2, SRTM C-band, SRTM X-band), так и распространяемые на коммерческой основе (SPOT DEM, NextMap World 30, NextMap, TanDEM-X Global DEM, World 3D Topographic Data). Глобальные модели высот представляют собой ценные пространственные данные, потенциально пригодные для решения многих задач, в том числе задач, связанных с инвентаризацией и управлением лесами.

Значительное внимание исследователей уделяется определению точностных характеристик модели SRTM, сравнению SRTM с иными

моделями высот или результатами наземных съемок, оценке возможностей использования модели для создания топографических карт [25]. Интерес к использованию глобальных цифровых моделей высот для изучения лесов связан с возможностью определения на их основе многочисленных морфометрических характеристик рельефа (направление и крутизна склонов, показатели кривизны, конвергенции и дивергенции, размеры водосборов, индексы инсоляции и влажности почв). Инструментами для определения и анализа морфометрических характеристик рельефа могут служить специальные геоинформационные системы, например SagaGIS [26, 27]. Теоретические основы математического моделирования и анализа рельефа, а также изучения взаимосвязей между рельефом и другими компонентами геосистем изучаются средствами геоморфометрии (geomorphometry). Методы геоморфометрии широко используются для решения задач геоморфологии, гидрологии, почвоведения, геоботаники, геологии, гляциологии, океанологии, климатологии и других наук о Земле [28]. Представление о предмете изучения, современном состоянии и перспективах развития геоморфометрии можно получить из ряда обзорных публикаций [28-30].

В публикациях, связанных с оценкой взаимосвязей характеристик рельефа с характеристиками лесов, упоминаются следующие морфометрические характеристики рельефа: абсолютная высота [31], среднее квадратическое отклонение и энтропия высот [32], значения уклона, экспозиции склонов и кривизны поверхности [33], абсолютной высоты, кривизны поверхности, формы и экспозиции склонов [34], набор морфометрических характеристик рельефа, определяемых на основе свободных ЦМР [35-38].

Важно отметить, что модели высот, получаемые на основе радарных съемок, относятся не к моделям рельефа (DTM), а к моделям поверхности (DSM). Это характерно для моделей высот, получаемых с помощью оптических, радарных, лидарных и аэрофотосъемок [30]. Модели рельефа (DTM) могут быть получены на основе материалов наземных съемок, GPS-позиционирования, топографических карт. Преимуществами моделей высот, получаемых на основе использования сканирующих устройств воздушного или спутникового базирования (радарные или лидарные съемки), являются очень высокая плотность получаемых данных и регулярность выборки. Поэтому модели поверхности (DSM) более точны в изображении мезо- и микрорельефа по сравнению с иными моделями [30]. Изображения SRTM демонстрируют чувствительность датчика к топографическим

особенностям местности, таким как дренажные сети и холмистый рельеф, а также к особенностям пространственного распределения растительности [39]. С другой стороны, использование материалов подобных съемок может приводить к ошибкам измерений, определяемым физическими ограничениями приборов (радаров и лидаров). С учетом того, что получаемые модели высот отражают не саму земную поверхность, а поверхность объектов, расположенных над ней, для анализа рельефа с их помощью необходимо выполнение предварительной обработки.

Отмечается, что степень покрытия территории древесно-кустарниковой растительностью оказывает негативное влияние на качество определения высот рельефа при радарных съемках [30, 40]. В целях минимизации влияния древесно-кустарниковой растительности на результаты съемок SRTM предлагаются различные алгоритмы по сглаживанию модели, снижению уровня шумов, выявлению и удалению из моделей поверхности участков, покрытых лесной растительностью. Сама растительность (главным образом лесная) при этом рассматривается как негативный фактор, препятствующий прохождению сигналов датчиков и снижающий точностные характеристики модели. Значительный интерес для лесного хозяйства представляет обратная задача — не минимизация влияния растительности на рельеф, а наоборот — выделение из модели высот «слоя растительности» и последующее его изучение [39, 41, 42].

Обзор исследований, посвященных определению высоты лесного полога с использованием данных SRTM, демонстрирует, что указанное направление актуально для разных стран и континентов: США, Китая, России, Австралии, Камбоджи [39, 40, 42-46]. В целом, алгоритмы определения высоты лесного полога с использованием материалов дистанционного зондирования подобны друг другу — высота лесного полога определяется на основе разницы моделей поверхности SRTM и рельефа. В публикациях детально исследуются отдельные вопросы — последствия некорректной регистрации моделей высот [44, 46]; влияние на высоту полога характеристик лесных насаждений — преобладающих древесных пород [41], плотности полога [46]; влияние на высоту полога характеристик рельефа [44, 46]. Во многих исследованиях выполняется сравнение высот, определяемых на основе SRTM, с результатами использования других моделей высот [41, 43, 45]. Публикаций об исследованиях зависимости запаса лесных насаждений от высоты фазы центра рассеяния не обнаружено.

Значительную методическую, познавательную и библиографическую ценность для изучения взаимосвязи высоты

лесных насаждений с результатами съемки SRTM представляет часто цитируемая специалистами статья Келлндорфера и Уокера [39]. В ней изложены теоретические основы съемки SRTM, рассмотрены факторы, определяющие значения высот и связанные с ними ошибки, приведены результаты практических исследований в разных по характеру рельефа и лесной растительности регионах США. Также оценены расхождения между поверхностью SRTM и поверхностью рельефа, связанные с влиянием лесной растительности. Выявлено наличие линейных взаимосвязей между высотами насаждений и разницей высот моделей поверхности и рельефа. Также в данной статье приведены формулы и определения ряда понятий, связанных с изучаемой тематикой.

Высота поверхности SRTM на покрытых лесом участках оказывается больше высоты поверхности открытого рельефа, но ниже средней высоты лесного полога. Разница между поверхностями SRTM и рельефа — высота центра фазы рассеяния (scattering phase center height) h_{spc} зависит от характеристик сенсора и объекта. К характеристикам объекта съемки (лесных насаждений), способных влиять на высоту центра фазы рассеяния, относятся структура и влажность лесной растительности, шероховатость и влажность почвы. К характеристикам сенсора, влияющим на высоту центра фазы рассеяния, относятся длина волны, базовая длина и ориентация, поляризация, угол падения, фазовый шум [39].

Средняя высота центра фазы рассеяния для покрытого лесом участка $\overline{h}_{spc}$ может быть оценена с относительно небольшими погрешностями при условии усреднения достаточного количества значений высот [39]:

$$\overline{h}_{spc} = \frac{1}{N_p}\sum_{i=1}^{N_p}(h_{SRTM} - h_{ED}) + \delta_v, \qquad (1)$$

где h_{SRTM} — высота пикселя модели SRTM, h_{ED} — высота пикселя модели рельефа, N_p — количество пикселей в границах насаждения, δ_v — абсолютная ошибка, представляющая собой вертикальное смещение между поверхностью SRTM и поверхностью рельефа.

Наличие линейных связей между высотами насаждений и высотами центра фазы рассеяния (на основе данных SRTM) подтверждается в ряде исследований [39, 41, 45].

Для выполнения лесоучетных работ на значительных по площади территориях приоритетным направлением считается

использование дистанционных методов. Возможности материалов современных радарных съемок (в частности топографической съемки SRTM) представляются перспективными для определения таких ключевых характеристик лесных насаждений, как средние высоты и запасы. Оценка этих характеристик актуальна не только для выполнения лесоучетных работ (национальной инвентаризации лесов и лесоустройства), но и для решения задач устойчивого управления лесами, экологии и охраны природы, изучения углеродного цикла.

Задачами данной статьи являются:

– оценка расхождений высот между моделями поверхности (данных радарной топографической съемки SRTM) и рельефа (на основе топографических карт) на примере модельной территории — определение высоты центра фазы рассеяния h_{spc};

– оценка взаимосвязей средних высот полога $\overline{h}$ и запасов древостоев $\overline{M}$ со средними высотами центра фазы рассеяния $\overline{h}_{spc}$;

– анализ влияния характеристик насаждений на высоту центра фазы рассеяния h_{spc}.

2. Методика исследования. Методика исследования включала сбор и подготовку исходных пространственных данных с формированием геоинформационного (ГИС) проекта модельной территории, цифровое моделирование рельефа на основе данных SRTM и топографических карт, сравнение расхождений высот между моделями и оценку взаимосвязей высот и запасов лесных насаждений с высотой центра фазы рассеяния. Порядок и краткий состав работ по основным этапам исследования отражены в таблице 1.

Таблица 1. Порядок работ по оценке возможностей использования данных радарной съемки SRTM для определения высот и запасов древостоев

Этап	Содержание	Программное обеспечение
1. Подготовка исходных пространственных данных в цифровых форматах	Формирование геоинформационного проекта модельной территории с наборами векторных и растровых слоев на основе баз данных лесоустройства. Создание векторного слоя регулярной сети с шагом 1 км с расчетом усредненных характеристик лесов внутри ячеек.	Геоинформационные системы (ГИС) QGIS, WinGIS, программа обработки лесоустроительной информации PLP-2015

Продолжение таблицы 1.

Этап	Содержание	Программное обеспечение
2. Создание моделей высот рельефа и поверхности	Поиск, перепроецирование и загрузка данных SRTM. Сканирование и трансформация в геоинформационный проект топографических карт, векторизация горизонталей и высотных отметок. Создание модели рельефа методом TIN. Сравнение моделей высот.	Картографический сервис EarthExplorer (https://earthexplorer.usgs.gov/), ГИС QGIS, Saga GIS, программы MS Excel, Statgraphics
3. Определение высоты центра фазы рассеяния	Определение разницы высот моделей SRTM и рельефа (определение высоты центра фазы рассеяния). Сравнение средних высот полога лесных насаждений со средними высотами центра фазы рассеяния	ГИС QGIS, модуль QProf
4. Моделирование характеристик насаждений	Моделирование высоты и запасов древостоев на основе высоты центра фазы рассеяния (для ячеек регулярной сети и групп лесотаксационных выделов)	Программы PLP-2015, Statgraphics. MS Excel

3. Объекты исследования. В качестве модельной территории выбрана часть Учебно-опытного лесничества Ленинградской области, расположенная в Тосненском районе. Рельеф территории равнинный. Последнее лесоустройство было проведено в 2005 году. Площадь земель лесного фонда представлена на 93,5 % лесными землями, из которых 98,84 % составляют покрытые лесной растительностью земли. Средний состав насаждений — 3С3Е3Б1Ос, средний возраст — 82 года, средний класс бонитета — 2,4, средняя полнота — 0,68, средний запас на 1 га покрытых лесом земель — 214 $м^3$/ га.

Модельная территория ограничена прямоугольником с целым числом квадратных ячеек (рисунок 1а, 1б). Из анализа исключены ячейки, в которых доля земель лесного фонда составила менее 95% от общей площади, а также ячейки, в которых доля земель, покрытых лесной растительностью, составила менее 90% от общей площади (поэтому из 220 исходных ячеек выбрано 150).

Исходными данными для исследования служили материалы лесоустройства (геоинформационные базы данных, лесные карты), выполненного в 2005 году, данные съемки SRTM и топографические карты (рисунок).

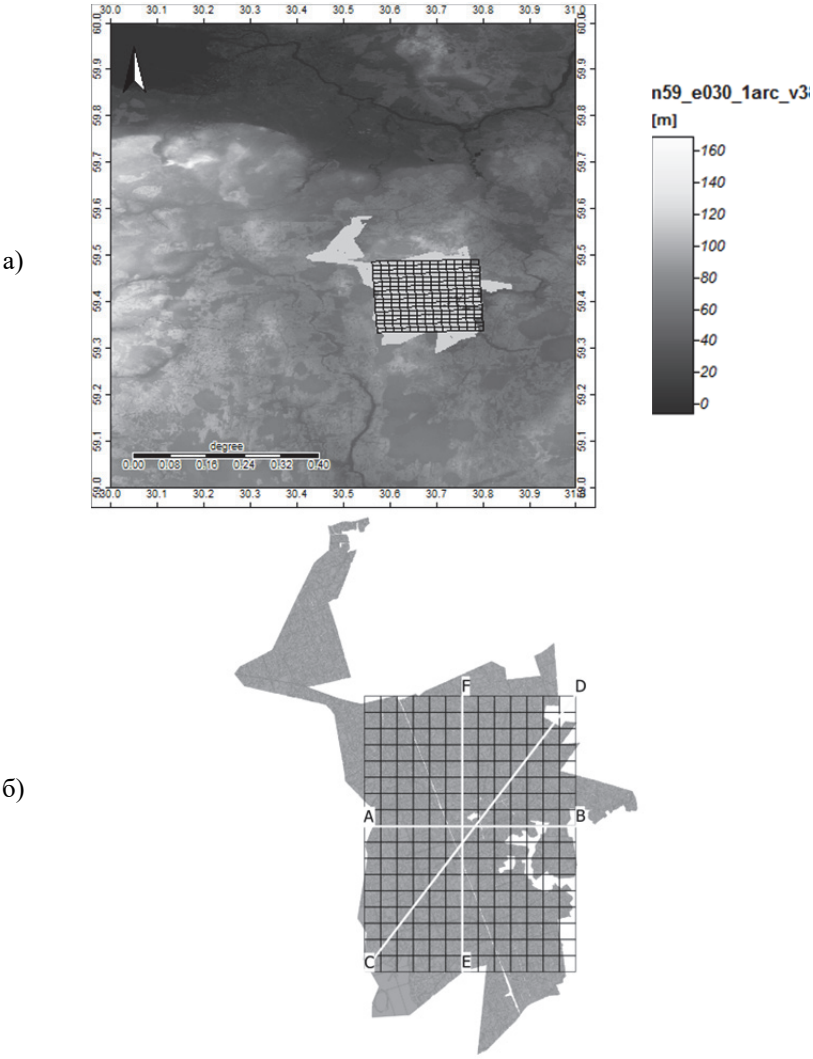


Рис. 1. Исходные пространственные данные: а) расположение модельной территории в виде регулярной сетки прямоугольной формы с шагом 1 км со схемой Учебно-опытного лесничества внутри квадрата матрицы высот SRTM со стороной 1°; б) расположение профилей для анализа высот рельефа, поверхности SRTM и лесного полога

4. Результаты исследования. Средствами ГИС путем вычитания значений высот моделей рельефа из модели SRTM

сформирована новая цифровая модель высот (рисунок 2). Полученная модель высот характеризует положение высоты центра фазы рассеяния. На рисунке 3 показаны профили значений высот обеих моделей, а также приведен график высоты центра фазы рассеяния (h_{SPC}).

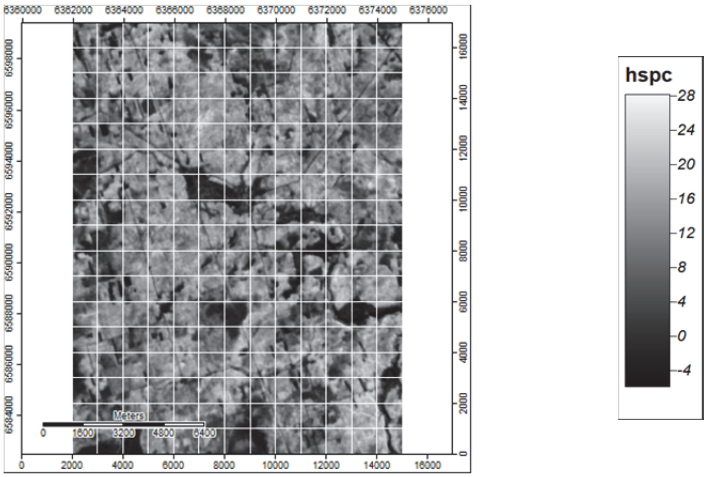
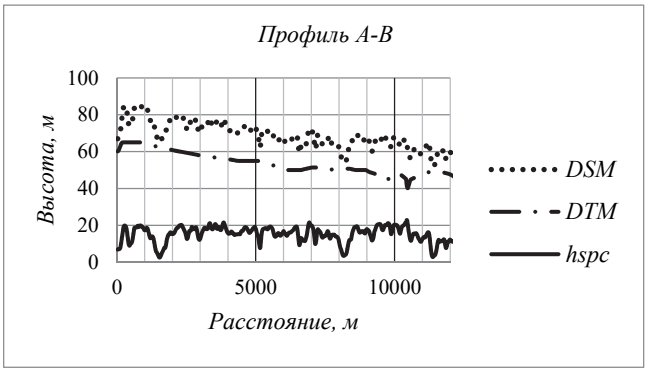
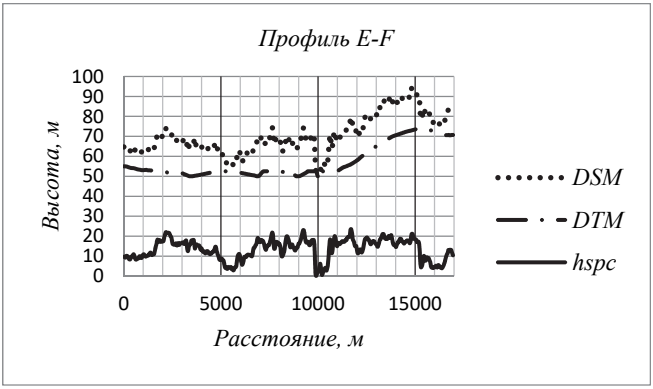


Рис. 2. Изображение модели высот центра фазы рассеяния (разницы моделей поверхности SRTM и рельефа)

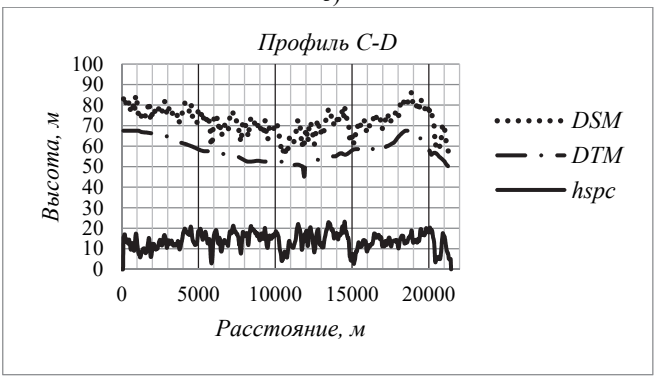


а)

Рис. 3. Профили значений высот, полученных на основе моделей рельефа (DTM) и поверхности (DSM). Разница высот моделей поверхности и рельефа — высота центра фазы рассеяния h_{spc} [38]



б)



в)

Рис. 3. Профили значений высот, полученных на основе моделей рельефа (DTM) и поверхности (DSM). Разница высот моделей поверхности и рельефа — высота центра фазы рассеяния h_{spc} [38]

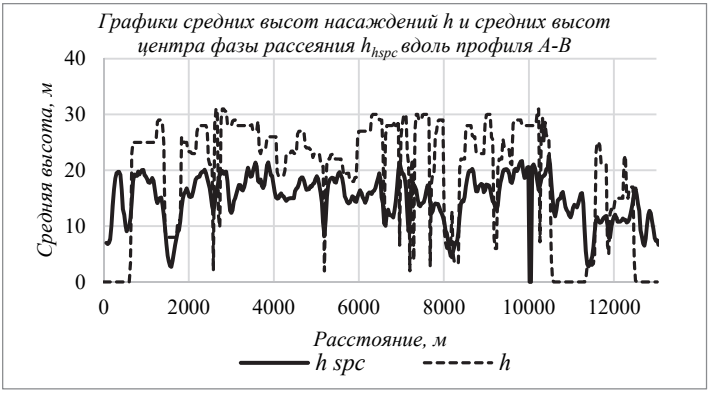
Просмотр изменения высот поверхностей по профилям (рисунок 4) позволяет отметить следующие тенденции:

– характер изменения высот между моделями поверхности и рельефа в целом во всех направлениях однороден, но изменчивость высот модели поверхности заметно выше;

– изменения значений высот поверхности без изменения высот рельефа наблюдаются в местах чередования насаждений с разными средними высотами полога, чередования покрытых и непокрытых лесной растительностью участков;

– изменения значений высот поверхности с одновременным изменением высот рельефа происходят в поймах рек и ручьев.

а)



б)

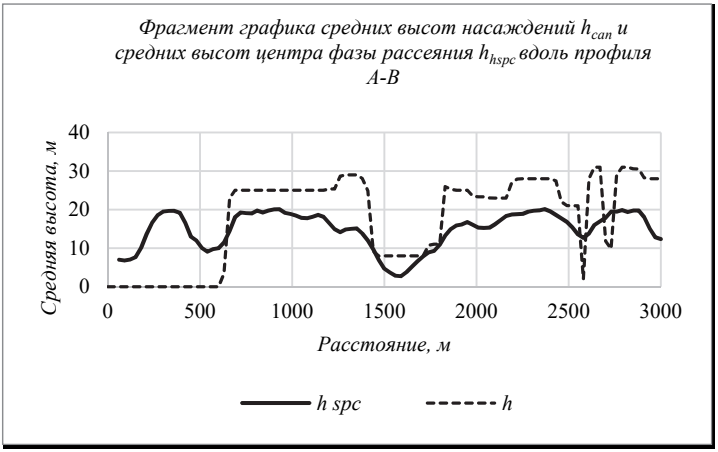


Рис. 4. Сравнение графиков средних высот лесных насаждений со средними высотами центра фазы рассеяния: а) графики средних высот лесных насаждений ($\overline{h}$) и средних высот центра фазы рассеяния ($\overline{h}_{spc}$) профиля A-B; б) увеличенный фрагмент профиля A-B

Графики высоты центра фазы рассеяния на профилях показывают, что для большей части модельной территории (которая относится к лесопокрытой) расхождения высот между моделями поверхности и рельефа составляют 15-20 м. Сближение графиков высот отмечается на участках, не покрытых лесной растительностью [38].

На рисунке 4б показан фрагмент профиля с высотами центра фазы рассеяния ($\overline{h}_{spc}$) и средними высотами насаждений ($\overline{h}$). Ступенчатый характер графика средних высот насаждений с наличием

ровных участков объясняется тем, что для каждого лесотаксационного выдела известно единственное значение высоты (средняя высота первого яруса). При соседстве выделов с разными средними высотами перепад высот получается резким. График поверхности SRTM (высоты фазы рассеяния) более плавный. Возможно, такая плавность при переходе между насаждениями с разными средними высотами в некоторых случаях будет справедливой. Иными причинами плавного характера изменения графика поверхности SRTM могут быть невысокое пространственное разрешение съемки (около 30 м) и влияние смежных объектов.

Графики на рисунке 4 подтверждают, что высота центра фазы рассеяния не соответствует реальной высоте лесных насаждений. Кривая графика средней высоты древостоев в целом проходит выше кривой средней высоты центра фазы рассеяния (рисунок 4а). В публикациях приводятся сведения о закономерностях соотношения высот модели SRTM и лесных насаждений. В частности, отмечается, что поверхность центра фазы рассеяния, фиксируемая сенсором, может составлять от 0,5 до 0,75 от высоты лесного полога [30]. Более глубокому проникновению лучей С-диапазона в лесной полог могут способствовать сухая погода, незначительная плотность полога, небольшая высота насаждений и небольшие размеры листьев и ветвей, коническая форма крон. Такие факторы, как высокая сомкнутость насаждений, преобладание лиственных пород в лесном пологе, наличие ветвей среднего и большого размера препятствуют проникновению лучей радарной съемки в лесной полог [30, 40].

Для оценки высот объектов на основе материалов SRTM (с размером пикселя 30 м) рекомендуется выбирать участки площадью не менее 1,8 га [39]. Также отмечается, что относительная вертикальная ошибка высоты, связанная с фазовым шумом, снижается с увеличением выборки за счет усреднения. В настоящем исследовании использовались следующие градации выделов по площади — все выделы, более 2,5 га, более 5 га, более 7,5 га. В таблице 2 показаны значения высот центра фазы рассеяния для групп категорий земель и отдельных категорий земель. К нелесным землям на модельной территории относятся следующие категории земель: болота, трассы ЛЭП, противопожарные разрывы, карьеры, кладбища, прочие земли и другие. К группе категорий земель «лесные земли, покрытые лесной растительностью» относятся насаждения естественного происхождения, насаждения искусственного происхождения (лесные культуры), насаждения из подроста, насаждения естественные с примесью лесных культур.

Таблица 2. Высоты центра фазы рассеяния по группам категорий земель и отдельным категориям земель. В таблице указаны средние значения высоты центра фазы рассеяния (м), стандартные ошибки (м) и количество выделов в скобках (шт)

Категории и группы категорий земель	Площади выделов				
	без ограничений	не меньше 2,5 га	не меньше 5 га	не меньше 7,5 га	не меньше 10 га
Нелесные земли	7,8±3,8 (219)	7,0±3,5 (81)	5,6±3,4 (34)	4,0±0,9 (22)	4,2±1,0 (11)
Болота	5,6±3,8 14)	3,8±0,8 (10)	3,6±0,7 (9)	3,6±0,7 (9)	3,6±0,5 (6)
Сенокосы	6,1±2,6 (58)	5,1±2,3 (14)	4,1±1,2 (4)	-	-
Пашни, пастбище, выгоны, ландшафтная поляна	5,2±2,9 (17)	3,5±1,4 (9)	3,2±1,0 (4)	3,6±0,7 (3)	-
Непокрытые лесной растительностью лесные земли	10,9±5,2 (122)	10,6±6,0 (26)	11,7±6,5 (12)	12,3±7,9 (5)	-
Несомкнутые лесные культуры	9,3±6,4 (38)	7,2±5,5 (13)	4,0±0,5 (4)	3,9±0,2 (2)	-
Лесные земли покрытые лесной растительностью	14,4±4,0 (7361)	14,2±4,3 (2620)	13,7±4,5 (793)	13,7±4,6 (282)	13,2±4,7 (127)

К лесным землям, не покрытым лесной растительностью — погибшие насаждения, вырубки, прогалины, несомкнутые лесные культуры. Включение в данную таблицу отдельных категорий земель не имеет практического смысла – одни встречаются единично, другие могут включать древесно-кустарниковую растительность (кладбище, плантация, сад, питомник) или постройки (поселок, усадьба, кордон). Также нет смысла анализировать узкие линейные (реки, линии электропередач, противопожарные разрывы) и мелкие по площади (прогалины) объекты, поскольку на высоту модели поверхности будут влиять смежные объекты (как правило, насаждения), а размеры самих объектов могут быть сопоставимы с размером одного пиксела.

Увеличение площади выделов приводит к постепенному снижению высоты центра фазы рассеяния и среднего квадратического отклонения. Показанные значения демонстрируют, что даже для открытых территорий, теоретически лишенных строений и древесной

растительности (болот, пашен, пастбищ), расхождения значений высот поверхности и рельефа существенно отличаются от нуля. Для лесных территорий значения высот центра фазы рассеяния практически не меняются с увеличением площади выдела.

При оценке взаимосвязей между характеристиками лесных насаждений и высотами центра фазы рассеяния использовались ячейки регулярной сети с шагом 1 км (рисунок 5) и лесотаксационные выделы (таблицы 3 и 4). Для выполнения регрессионного анализа использовали линейные уравнения без константы, поскольку высота центра фазы рассеяния (разница высот моделей поверхности и рельефа) для лесопокрытых площадей всегда будет отличаться от нуля.

На рисунке 5 показаны графики двух регрессионных уравнений, отражающих зависимости между средними высотами и средними запасами насаждений от средней высоты центра фазы рассеяния для 150 ячеек регулярной сети с шагом 1 км.

Далее оценивались регрессионные зависимости между средними высотами центра фазы рассеяния (независимая переменная) и средними высотами или запасами насаждений (зависимая переменная) в пределах границ лесотаксационных выделов, сгруппированных по преобладающим породам (таблица 3). Из анализа были исключены насаждения младше 40 лет по той причине, что определение характеристик высот поверхности их полога на основе материалов лесоустройства практически невозможно.

а)

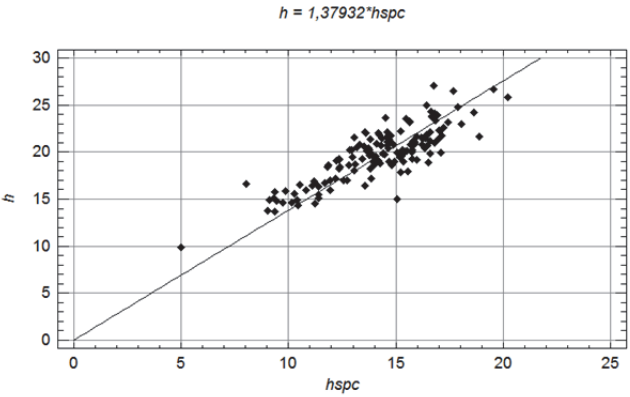


Рис. 5. Графики зависимости средних высот и запасов насаждений от средних высот центра фазы рассеяния SRTM внутри ячеек регулярной сети:
а) зависимость средней высоты насаждений $\overline{h}$ (м), от средней высоты фазы рассеяния $\overline{h}_{spc}$ (м), коэффициент детерминации $R^2 = 99{,}18\%$

б)

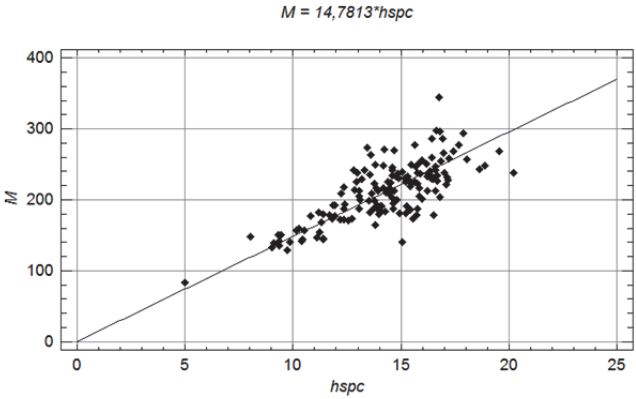


Рис. 5. Графики зависимости средних высот и запасов насаждений от средних высот центра фазы рассеяния SRTM внутри ячеек регулярной сети: б) зависимость среднего запаса насаждений $\overline{M}$ (м3/га) от средней высоты фазы рассеяния $\overline{h}_{spc}$ (м), коэффициент детерминации $R^2 = 98{,}35\%$

Результаты, представленные в таблице 3, демонстрируют наличие тесных регрессионных связей между анализируемыми показателями. Коэффициенты регрессии для всех выделов насаждений четырех групп основных лесообразующих пород в целом близки. Отличия регрессионных коэффициентов между группами насаждений при прочих равных условиях можно трактовать различной пропускной способностью древесного полога. Но структура лесного полога насаждений разных типов достаточно разнообразна.

Рассматриваемые в таблице 3 группы насаждений выделены только по преобладающей в их составе древесной породе в соответствии с действующей Лесоустроительной инструкцией. Выделение отдельных выделов при лесоустройстве допускает определенное варьирование лесотаксационных характеристик насаждений (коэффициентов состава древесных пород, высот, диаметров, возрастов, относительной полноты и бонитета насаждения). Также допустима определенная пространственная неоднородность лесотаксационных выделов (в том числе наличие в выделе единичных деревьев и открытых участков, наличие участков с разной полнотой и сомкнутостью полога, а также неоднородность породного состава на территории выдела). Указанные особенности исходных данных отражаются на неоднородности полога лесных

насаждений (следовательно, и на положении поверхности SRTM — высоты центра фазы рассеяния).

Таблица 3. Регрессионные уравнения зависимостей средней высоты насаждений $\overline{h}$ (м) от средней высоты фазы рассеяния $\overline{h}_{spc}$ (м). В скобках указано количество выделов (шт.)

Критерии оценки и их градации	Группы насаждений			
	сосновые	еловые	березовые	осиновые
Все насаждения	$\overline{h} = 1{,}44\overline{h}_{spc}$ $R^2 = 95{,}83$ (1815)	$\overline{h} = 1{,}46\overline{h}_{spc}$ $R^2 = 96{,}28$ (1851)	$\overline{h} = 1{,}49\overline{h}_{spc}$ $R^2 = 96{,}56$ (1307)	$\overline{h} = 1{,}52\overline{h}_{spc}$ $R^2 = 97{,}16$ (1026)
Площадь выдела, га				
менее 2,5 га	$\overline{h} = 1{,}44\overline{h}_{spc}$ $R^2 = 95{,}48$ (1061)	$\overline{h} = 1{,}45\overline{h}_{spc}$ $R^2 = 96{,}06$ (1264)	$\overline{h} = 1{,}49\overline{h}_{spc}$ $R^2 = 96{,}13$ (854)	$\overline{h} = 1{,}52\overline{h}_{spc}$ $R^2 = 96{,}96$ (673)
2,5-5 га	$\overline{h} = 1{,}43\overline{h}_{spc}$ $R^2 = 96{,}41$ (521)	$\overline{h} = 1{,}48\overline{h}_{spc}$ $R^2 = 96{,}69$ (434)	$\overline{h} = 1{,}48\overline{h}_{spc}$ $R^2 = 97{,}13$ (302)	$\overline{h} = 1{,}5\overline{h}_{spc}$ $R^2 = 97{,}57$ (243)
более 5 га	$\overline{h} = 1{,}46\overline{h}_{spc}$ $R^2 = 96{,}19$ (233)	$\overline{h} = 1{,}51\overline{h}_{spc}$ $R^2 = 97{,}08$ (153)	$\overline{h} = 1{,}47\overline{h}_{spc}$ $R^2 = 97{,}87$ (151)	$\overline{h} = 1{,}53\overline{h}_{spc}$ $R^2 = 97{,}53$ (110)
Относительная полнота, доля единицы				
0,5 и менее	$\overline{h} = 1{,}65\overline{h}_{spc}$ $R^2 = 95{,}03$ (476)	$\overline{h} = 1{,}57\overline{h}_{spc}$ $R^2 = 95{,}82$ (549)	$\overline{h} = 1{,}69\overline{h}_{spc}$ $R^2 = 96{,}05$ (250)	$\overline{h} = 1{,}64\overline{h}_{spc}$ $R^2 = 96{,}89$ (285)
0,6-0,7	$\overline{h} = 1{,}41\overline{h}_{spc}$ $R^2 = 96{,}92$ (1080)	$\overline{h} = 1{,}44\overline{h}_{spc}$ $R^2 = 96{,}74$ (910)	$\overline{h} = 1{,}47\overline{h}_{spc}$ $R^2 = 96{,}86$ (799)	$\overline{h} = 1{,}49\overline{h}_{spc}$ $R^2 = 97{,}44$ (554)
0,8 и выше	$\overline{h} = 1{,}32\overline{h}_{spc}$ $R^2 = 97{,}91$ (259)	$\overline{h} = 1{,}35\overline{h}_{spc}$ $R^2 = 97{,}57$ (392)	$\overline{h} = 1{,}39\overline{h}_{spc}$ $R^2 = 98{,}25$ (258)	$\overline{h} = 1{,}44\overline{h}_{spc}$ $R^2 = 98{,}11$ (187)

Продолжение таблицы 3.

Критерии оценки и их градации	Группы насаждений			
	сосновые	еловые	березовые	осиновые
Коэффициент преобладающей породы в составе, ед.				
5 и менее	$\bar{h} = 1{,}65\bar{h}_{spc}$ $R^2 = 95{,}03$ (476)	$\bar{h} = 1{,}49\bar{h}_{spc}$ $R^2 = 96{,}51$ (1315)	$\bar{h} = 1{,}48\bar{h}_{spc}$ $R^2 = 96{,}77$ (686)	$\bar{h} = 1{,}5\bar{h}_{spc}$ $R^2 = 97{,}22$ (535)
6-7	$\bar{h} = 1{,}41\bar{h}_{spc}$ $R^2 = 96{,}62$ (1080)	$\bar{h} = 1{,}37\bar{h}_{spc}$ $R^2 = 96{,}63$ (412)	$\bar{h} = 1{,}5\bar{h}_{spc}$ $R^2 = 96{,}56$ (510)	$\bar{h} = 1{,}55\bar{h}_{spc}$ $R^2 = 97{,}06$ (364)
8 и более	$\bar{h} = 1{,}32\bar{h}_{spc}$ $R^2 = 97{,}91$ (259)	$\bar{h} = 1{,}39\bar{h}_{spc}$ $R^2 = 94{,}02$ (124)	$\bar{h} = 1{,}47\bar{h}_{spc}$ $R^2 = 95{,}21$ (111)	$\bar{h} = 1{,}47\bar{h}_{spc}$ $R^2 = 97{,}41$ (127)

Для изучения взаимосвязи средней высоты насаждений со средней высотой центра фазы рассеяния помимо преобладающей породы также учитывались следующие критерии: коэффициент состава преобладающей породы, относительная полнота и площадь выдела (таблица 3).

Регрессионный анализ взаимосвязи средней высоты центра фазы рассеяния со средней высотой насаждений позволил установить ряд закономерностей. Величина площади выдела практически не влияет на значения регрессионных коэффициентов и тесноту связи для всех групп насаждений. Увеличение относительной полноты для всех пород однозначно приводит к снижению коэффициентов регрессии. Коэффициент регрессии в данном случае характеризует отличие средней высоты насаждения от средней высоты центра фазы рассеяния (чем больше отличие коэффициента регрессии от единицы, тем более глубоко проникают лучи радарной съемки в лесной полог). Максимальные значения коэффициентов регрессии отмечаются в низкополнотных насаждениях.

Изменения коэффициентов регрессии в зависимости от изменения состава насаждений проявляются неодинаково для насаждений разных пород. Для хвойных насаждений максимальные значения коэффициентов регрессии (следовательно, и максимальная пропускная способность полога для лучей радарной съемки) отмечаются в смешанных насаждениях с долей преобладающей

породы в составе 5 и менее единиц. Увеличение доли преобладающей породы в составе насаждения приводит к снижению коэффициента регрессии. Для лиственных насаждений изменение доли преобладающей породы в составе практически не сказывается на значениях коэффициентов регрессии (они остаются близкими к средним по всем выделам).

Из результатов, представленных в таблице 3, видно, что регрессионные коэффициенты для всех групп насаждений в целом близки и незначительно отличаются от коэффициента 1,38, определенного для всех лесопокрытых участков регулярной сети (рисунок 5а). При этом наиболее низкие значения коэффициентов регрессии отмечаются для сосновых (1,32) и еловых (1,35) насаждений, наиболее высокие (1,44) — для осиновых.

В таблице 4 представлены результаты регрессионного анализа зависимости среднего запаса насаждений от средней высоты фазы рассеяния. Для данного анализа коэффициент регрессии может рассматриваться как «плотность» насаждений (чем выше значение коэффициента регрессии, тем больше запас лесного насаждения при одинаковых значениях высоты фазы рассеяния). Закономерного влияния площади выдела на величину регрессионных коэффициентов и тесноту связей не отмечается. Для всех групп насаждений коэффициенты регрессии увеличиваются с ростом относительной полноты.

Таблица 4. Регрессионные уравнения зависимостей среднего запаса насаждений $\overline{M}$ (м^3/га) от средней высоты центра фазы рассеяния $\overline{h}_{spc}$ (м)

Критерии оценки и их градации	Группы насаждений			
	сосновые	еловые	березовые	осиновые
Все насаждения	$\overline{M} = 15{,}73\overline{h}_{spc}$ $R^2 = 95{,}17$ (1815)	$\overline{M} = 16{,}83\overline{h}_{spc}$ $R^2 = 93{,}35$ (1851)	$\overline{M} = 13{,}96\overline{h}_{spc}$ $R^2 = 95{,}4$ (1307)	$\overline{M} = 16{,}19\overline{h}_{spc}$ $R^2 = 94{,}67$ (1026)
Площадь выдела, га				
менее 2,5 га	$\overline{M} = 15{,}65\overline{h}_{spc}$ $R^2 = 94{,}77$ (1061)	$\overline{M} = 16{,}7\overline{h}_{spc}$ $R^2 = 92{,}77$ (1264)	$\overline{M} = 13{,}7\overline{h}_{spc}$ $R^2 = 94{,}85$ (854)	$\overline{M} = 16{,}31\overline{h}_{spc}$ $R^2 = 94{,}64$ (673)
2,5-5 га	$\overline{M} = 15{,}71\overline{h}_{spc}$ $R^2 = 95{,}64$ (521)	$\overline{M} = 17{,}14\overline{h}_{spc}$ $R^2 = 94{,}6$ (434)	$\overline{M} = 14{,}36\overline{h}_{spc}$ $R^2 = 96{,}21$ (302)	$\overline{M} = 15{,}77\overline{h}_{spc}$ $R^2 = 94{,}54$ (243)

Продолжение Таблицы 4.

Критерии оценки и их градации	Группы насаждений			
	сосновые	еловые	березовые	осиновые
более 5 га	$\overline{M} = 16{,}13\overline{h}_{spc}$ $R^2 = 96{,}08$ (233)	$\overline{M} = 17\overline{h}_{spc}$ $R^2 = 94{,}56$ (153)	$\overline{M} = 14{,}53\overline{h}_{spc}$ $R^2 = 96{,}97$ (151)	$\overline{M} = 16{,}45\overline{h}_{spc}$ $R^2 = 95{,}39$ (110)
Относительная полнота, доля единицы				
0,5 и менее	$\overline{M} = 12{,}7\overline{h}_{spc}$ $R^2 = 93{,}8$ (476)	$\overline{M} = 13{,}04\overline{h}_{spc}$ $R^2 = 93{,}05$ (549)	$\overline{M} = 11{,}46\overline{h}_{spc}$ $R^2 = 94{,}3$ (250)	$\overline{M} = 12{,}51\overline{h}_{spc}$ $R^2 = 95{,}57$ (285)
0,6 – 0,7	$\overline{M} = 15{,}99\overline{h}_{spc}$ $R^2 = 96{,}41$ (1080)	$\overline{M} = 17{,}58\overline{h}_{spc}$ $R^2 = 95{,}98$ (910)	$\overline{M} = 14{,}02\overline{h}_{spc}$ $R^2 = 96{,}05$ (799)	$\overline{M} = 16{,}55\overline{h}_{spc}$ $R^2 = 96{,}94$ (554)
0,8 и выше	$\overline{M} = 18{,}39\overline{h}_{spc}$ $R^2 = 97{,}6$ (259)	$\overline{M} = 20{,}3\overline{h}_{spc}$ $R^2 = 96{,}47$ (392)	$\overline{M} = 15{,}69\overline{h}_{spc}$ $R^2 = 97{,}31$ (258)	$\overline{M} = 20{,}1\overline{h}_{spc}$ $R^2 = 97{,}82$ (187)
Коэффициент преобладающей породы в составе, ед.				
5 и менее	$\overline{M} = 15{,}48\overline{h}_{spc}$ $R^2 = 94{,}79$ (815)	$\overline{M} = 16{,}92\overline{h}_{spc}$ $R^2 = 94{,}09$ (1315)	$\overline{M} = 13{,}84\overline{h}_{spc}$ $R^2 = 95{,}69$ (686)	$\overline{M} = 15{,}26\overline{h}_{spc}$ $R^2 = 95{,}36$ (535)
6-7	$\overline{M} = 15{,}58\overline{h}_{spc}$ $R^2 = 95{,}68$ (454)	$\overline{M} = 16{,}67\overline{h}_{spc}$ $R^2 = 92{,}53$ (412)	$\overline{M} = 14{,}18\overline{h}_{spc}$ $R^2 = 95{,}22$ (510)	$\overline{M} = 16{,}87\overline{h}_{spc}$ $R^2 = 94{,}61$ (364)
8 и более	$\overline{M} = 16{,}33\overline{h}_{spc}$ $R^2 = 95{,}55$ (209)	$\overline{M} = 16{,}24\overline{h}_{spc}$ $R^2 = 87{,}04$ (124)	$\overline{M} = 13{,}73\overline{h}_{spc}$ $R^2 = 94{,}49$ (111)	$\overline{M} = 18{,}1\overline{h}_{spc}$ $R^2 = 95{,}31$ (127)

Для сосновых и осиновых насаждений отмечается также рост коэффициентов регрессии с увеличением в составе доли преобладающей породы (то есть переходом от смешанных к чистым насаждениям).

Для еловых насаждений наоборот — с увеличением доли ели в составе коэффициент регрессии несколько уменьшается. Среди

насаждений основных групп древесных пород наиболее высокие коэффициенты регрессии отмечаются у ельников. Максимальные значения коэффициентов регрессии — у высокополнотных еловых и осиновых насаждений.

Наличие расхождений между результатами регрессионного анализа для разных насаждений в целом закономерно — средняя высота центра фазы рассеяния зависит от плотности и структуры полога, которые, в свою очередь, зависят от составляющих полог древесных пород, сомкнутости полога, его однородности. Объяснение причин установленных в результате регрессионного анализа закономерностей требует проведения более детальных исследований (в частности с учетом всех составляющих первый ярус древесных пород; учетом наличия и характеристик других ярусов). Одной из наиболее очевидных причин различия регрессионных коэффициентов является различие структуры лесного полога преобладающих типов насаждений основных древесных пород.

Сосновые насаждения (особенно чистые, с коэффициентом состава 7-10 единиц) в условиях Учебно-опытного лесничества, как правило, относительно однородны по высоте и сомкнутости. Полог сосновых насаждений относительно ровный и плотный. Еловые насаждения обычно представлены деревьями разного возраста и высоты, с разными размерами крон и промежутков между кронами. Остроконечные кроны ельников разной высоты формируют неровную поверхность полога, что не может не сказываться на положении высоты центра фазы рассеяния. Важно отметить, что съемка SRTM выполнялась зимой (февраль 2000 г.), когда березовые и осиновые насаждения находились в безлистном состоянии, следовательно, глубина проникновения лучей внутрь полога для лиственных насаждений должна быть заведомо больше, чем для хвойных. При выполнении радарной съемки в период вегетации высота центра фазы рассеяния для лиственных насаждений будет отличаться от использованных результатов съемки SRTM.

Для детального исследования поверхности, формируемой при обработке материалов радарных съемок лесных насаждений, целесообразно использовать дополнительные источники информации, позволяющие оценивать варьирование высот лесного полога. Например, стереоизображения, полученные на основе материалов аэрофотосъемки, космической съемки или съемки, выполненной беспилотными летательными аппаратами, а также материалы наземных обследований.

5. Заключение. В результате проведенного исследования предложен новый методический подход к определению важнейших

характеристик лесных насаждений — высот и запасов. Оценка расхождений высот между моделью SRTM и моделью рельефа (на основе топографических карт) на примере Учебно-опытного лесничества Ленинградской области позволила выявить ряд закономерностей.

1. Высота поверхности SRTM отличается от высоты рельефа (определяемой на основе топографических карт). При этом средняя высота центра фазы рассеяния SRTM (разница высот модели SRTM и модели рельефа) для нелесных земель составила 7,0±3,5 м; для лесных земель, покрытых лесной растительностью, — 14,2±4,3 м; для лесных земель, не покрытых лесной растительностью, — 10,6±6,0 м. Увеличение площади выделов приводит к уменьшению значений полученных оценок средней высоты центра фазы рассеяния и среднего квадратического отклонения.

2. Графики высот моделей SRTM и рельефа вдоль профилей отражают синхронный характер изменений. Средняя высота центра фазы рассеяния SRTM располагается ниже средней высоты лесного полога, что подтверждает сведения, приводимые в публикациях [39-41].

3. В результате исследований установлено влияние на зависимости между средней высотой центра фазы рассеяния и средними высотами и запасами лесных насаждений таких показателей, как вид преобладающей древесной породы, относительная полнота насаждения, коэффициент состава преобладающей породы.

С учетом глобального характера данных SRTM выявленные закономерности могут оказаться полезными для проведения лесоучетных работ регионального и глобального уровней, (например, решения задач государственной инвентаризации лесов), а также решения глобальных экологических задач (определения запасов углерода, оценки надземной биомассы).

Целесообразно продолжить исследование в нескольких направлениях: дальнейшее изучение взаимосвязей расхождений моделей высот SRTM и рельефа с характеристиками лесов, оценка возможностей применения для аналогичных целей материалов других радарных съемок, оценка влияния на характеристики лесов моделей высот рельефа и поверхности.

Литература

1. *Жирин В.М., Князева С.В., Эйдлина С.П.* Динамика спектральной яркости породно-возрастной структуры групп типов леса на космических снимках LANDSAT // Лесоведение. 2014. № 5. С. 3–12.
2. *Жирин В.М., Князева С.В., Эйдлина С.П.* Оценка влияния морфологии древесного полога и рельефа на спектральные характеристики лесов по данным Landsat // Исследование земли из космоса. 2016. № 5. С. 10–20.

3. *Толкач И.В., Саевич Ф.К.* Спектральные и яркостные характеристики основных лесообразующих пород на снимках сканера LEICA ADS100 // Труды БГТУ. Серия 1: Лесное хозяйство, природопользование и переработка возобновляемых ресурсов. 2016. № 1(183). С. 24–27.
4. *Толкач И.В. и др.* Закономерности изменчивости спектральных яркостей полога основных лесообразующих пород Беларуси на снимках сканера ADS 100 // Труды БГТУ. Серия 1: Лесное хозяйство, природопользование и переработка возобновляемых ресурсов. 2017. № 2(198). С. 43–49.
5. *Солдатенков А.А.* Дешифрирование состава лесной растительности в условиях среднегорного рельефа // Вестник Адыгейского государственного университета. 2014. Вып. 1(133). С. 127–130.
6. *Сидоренков В.М. и др.* Зонирование территории Удмуртской Республики по категориям среды обитания охотничьих ресурсов на основе данных спутниковой съемки Landsat 8 OLI-TIRS // Лесотехнический журнал. 2015. Т. 5. № 3(19). С. 84–93.
7. *Шарикалов А.Г., Якутин М.В.* Анализ состояния таежных экосистем с использованием методики автоматизированного дешифрирования // Известия Алтайского государственного университета. 2014. Вып. 3-1(83). С. 123–127.
8. *Перепечина Ю.И., Глушенков О.И., Корсиков Р.С.* Учет и оценка лесов, возникших на сельскохозяйственных землях, с использованием данных дистанционного зондирования Земли // Известия высших учебных заведений. Лесной журнал. 2016. № 4(352). С. 71–80.
9. *Перепечина Ю.И., Глушенков О.И., Корсиков Р.С.* Определение лесистости и количественных характеристик лесов по космическим снимкам Sentinel-2 (на примере Шебекинского муниципального района Белгородской обл.) // Лесохозяйственная информация. 2017. № 4(4). С. 85–93.
10. *Белова Е.И., Ершов Д.В.* Опыт оценки естественного лесовосстановления на сплошных вырубках по временным рядам // Лесоведение. 2015. № 5. С. 339–345.
11. *Черных Д.В., Бирюков Р.Ю., Золотов Д.В., Вагнер А.А.* Антропогенные модификации и трансформации ландшафтов в бассейне р. Касмала: классификация и динамика на основе данных дистанционного зондирования // Вестник Алтайской науки. 2014. № 1(19). С. 233–240.
12. *Соромотин А.В., Бродт Л.В.* Мониторинг растительного покрова при освоении нефтегазовых месторождений по данным многозональной съемки LANDSAT // Вестник Тюменского государственного университета. Экология и природопользование. 2018. Т. 4. № 1. С. 37–49.
13. *Терехов А.Г., Макаренко Н.Г., Пак И.Т.* Автоматический алгоритм классификации снимков Quickbird в задаче оценки полноты леса // Компьютерная оптика. 2014. Т. 38. № 3. С. 580–583.
14. *Савченко А.А., Выводцев Н.В.* Оценка возможностей применения данных дистанционного зондирования при мониторинге санитарного и лесопатологического состояния лесов // Ученые заметки ТОГУ. 2015. Т. 6. № 4. С. 658–661.
15. *Balenović I., Seletković A., Pernar R., Jazbec A.* Estimation of the mean tree height of forest stands by photogrammetric measurement using digital aerial images of high spatial resolution // Annals of Forest Research. 2015. vol. 58. no. 1. pp. 125–143.
16. *Архипов В.И., Черниховский Д.М., Березин В.И., Белов В.А.* Современная технология таксации лесов дешифровочным способом «От съемки – к проекту» // Известия Санкт-Петербургской лесотехнической академии. 2014. Вып. 208. С. 22–42.
17. *Алексеев А.С., Михайлова А.А., Черниховский Д.М., Березин В.И.* Метод определения таксационных характеристик насаждений по аэрофотоснимкам сверхвысокого разрешения // Труды Санкт-Петербургского научно-исследовательского института лесного хозяйства. 2017. № 2. С. 67–77.

18. *Faganan M., De Fries R.* Measurement and monitoring of the world's forests. A review and summary of remote sensing technical capability, 2009-2015 // Resources for the Future. 2009. 131 p.
19. *Balenović I., Alberti G., Marjanović H.* Airborne laser scanning – the status and perspectives for the application in the south-east European forestry // South-east European forestry. 2013. vol. 4. no. 2. pp. 59–79.
20. *Kauranne T. et al.* Airborne Laser Scanning Based Forest Inventory: Comparison of Experimental Results for the Perm Region, Russia and Prior Results from Finland // Forests. 2017. vol. 8. no. 3. pp. 72.
21. *Peuhkurinen J. et al.* Predicting Tree Diameter Distributions from Airborne Laser Scanning, SPOT 5 Satellite, and Field Sample Data in the Perm Region, Russia // Forests. 2018. vol. 9. no. 10. pp. 639.
22. *Gašparović M., Milas A., Seletković A., Balenović I.* A novel automated method for the improvement of photogrammetric DTM accuracy in forests // Šumarski list. 2018. vol. 142. no. 11-12. pp. 567–576.
23. *Balenović I., Milas A., Marjanović H.* A comparison of stand-level volume estimates from image-based canopy height models of different spatial resolutions // Remote Sensing. 2017. vol. 9. no. 3. pp. 205.
24. *Balenović I. et al.* Quality assessment of high density digital surface model over different land cover classes // Periodicum biologorum. 2016. vol. 117. no 4. pp. 459–470.
25. *Rabus B., Eineder M., Roth A., Bamler R.* The shuttle radar topography mission – a new class of digital elevation models acquired by spaceborne radar // ISPRS Journal of Photogrammetry and Remote Sensing. 2003. vol. 57. no. 4. pp. 241–262.
26. *Köthe R., Bock M.* Development and use in practice of Saga modules for high quality analysis of geodata // FREE AND OPEN GIS-SAGA-GIS. 2006. vol. 115. pp. 85–96.
27. *Conrad O. et al.* System for Automated Geoscientific Analyses (SAGA) v. 2.1.4 // Geoscientific Model Development. 2015. vol. 8. no. 7. pp. 1991–2007.
28. *Флоринский И.В.* Иллюстрированное введение в геоморфометрию // Электронное научное издание Альманах Пространство и Время. 2016. Т. 11. Вып. 1. URL: 2227-9490e-aprovr_e-ast11-1.2016.71 (дата обращения: 05.12.2018).
29. *Шарый П.А.* Геоморфометрия в науках о Земле и экологии, обзор методов и приложений // Известия Самарского научного центра Российской академии наук. 2006. Т. 8. № 2. С. 458–473.
30. *Hengl T., Reuter H.I.* Geomorphometry: Concepts, Software, Applications // Newnes. 2008. vol. 33. 772 p.
31. *Алексеев А.С., Никифоров А.А.* Влияние рельефа на структуру и продуктивность лесных ландшафтов с применением 3D-моделирования на примере Лисинского учебно-опытного лесхоза // Лесоведение. 2014. № 5. С. 42–53.
32. *Черниховский Д.М., Алексеев А.С.* Влияние формы поверхности рельефа на структуру и продуктивность лесных ландшафтов на примере заповедника «Верхне-Тазовский» Ямало-Ненецкого АО // Лесоведение. 2003. № 5. С. 10–17.
33. *Фарбер С.К.* Структуризация лесных сообществ // Сибирский лесной журнал. 2014. № 1. С. 35–49.
34. *Рахматуллина И.Р., Рахматуллин З.З., Мустафин Р.Ф.* Распространение и продуктивность сосновых насаждений в зависимости от морфометрических показателей рельефа (на примере Бугульминско-Белебеевской возвышенности в пределах Республики Башкортостан) // Вестник Ижевской государственной сельскохозяйственной академии. 2017. № 1(50). С. 42–52.
35. *Черниховский Д.М.* Оценка взаимосвязей морфометрических характеристик рельефа с количественными и качественными характеристиками лесов // Известия Санкт-Петербургской лесотехнической академии. 2016. Вып. 216. С. 69–90.

36. *Черниховский Д.М.* Оценка связей морфометрических характеристик рельефа с количественными и качественными характеристиками лесов на основе цифровых моделей рельефа ASTER и SRTM // Сибирский лесной журнал. 2017. № 3. С. 28–39.
37. *Черниховский Д.М.* Автоматическая классификация поверхности рельефа для изучения количественных и качественных характеристик лесов // Известия Санкт-Петербургской лесотехнической академии. 2017. Вып. 219. С. 74–95.
38. *Черниховский Д.М.* Использование автоматической классификации рельефа Ивахаши и Пайка для оценки количественных и качественных характеристик лесов на основе моделей высот рельефа и поверхности // Известия Санкт-Петербургской лесотехнической академии. 2018. Вып. 223. С. 100–126.
39. *Kellndorfer J. et al.* Vegetation height estimation from Shuttle Radar Topography Mission and National Elevation Datasets // Remote Sensing of Environment. 2004. vol. 93. pp. 339–358.
40. *Gallant J.C., Read A.M., Dowling T.I.* Removal of tree offsets from SRTM and other digital surface models // International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. 2012. vol. 39. no. 14. pp. 275–280.
41. *Sexton J. et al.* A comparison of lidar, radar, and field measurements of canopy height in pine and hardwood forests of southeastern North America // Forest Ecology and Management. 2009. vol. 257. pp. 1136–1147.
42. *Avtar R., Sawada H.* Use of DEM data to monitor height changes due to deforestation // Arabian Journal of Geosciences. 2013. vol. 6. no. 12. pp. 4859–4871.
43. *Sun G., Ranson K.J., Kharuk V.I., Kovacs K.* Validation of surface height from shuttle radar topography mission using shuttle laser altimeter // Remote Sensing of Environment. 2003. vol. 88. pp. 401–411.
44. *Ni W. et al.* Co-Registration of Two DEMs: Impacts on Forest Height Estimation from SRTM and NED at Mountainous Areas // IEEE Geoscience and Remote Sensing Letters. 2014. vol. 11. no. 1. pp. 273–277.
45. *Zhang Z. et al.* Estimation of forest structural parameters from Lidar and SAR data // The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. 2008. vol. 37. pp. 1121–1126.
46. *Miliaresis G., Delikaraoglou D.* Effects of Percent Tree Canopy Density and DEM Misregistration on SRTM/NED Vegetation Height Estimates // Remote Sensing. 2009. vol. 1. pp. 36–49.

Черниховский Дмитрий Михайлович — канд. с.-х. наук, доцент, кафедра лесной таксации, лесоустройства и геоинформационных систем, Санкт-Петербургский государственный лесотехнический университет им. С.М. Кирова (СПбГЛТУ); директор, Центр по развитию дистанционных методов в лесном хозяйстве ООО «Леспроект». Область научных интересов: лесоустройство, дистанционные методы в лесном хозяйстве, геоинформатика, геоморфометрия. Число научных публикаций — 65. cherndm2006@yandex.ru; 5, Институтский пер., 194021, Санкт-Петербург, Российская Федерация; р.т.: +7 (812)550-08-34; факс: +7(812)550-08-15.

Алексеев Александр Сергеевич —д-р геогр. наук, профессор, заведующий кафедрой, кафедра лесной таксации, лесоустройства и геоинформационных систем, Санкт-Петербургский государственный лесотехнический университет им. С.М. Кирова (СПбГЛТУ). Область научных интересов: лесоустройство, лесная экология, мониторинг лесов, ГИС-технологии для учета и управления лесами, математическое моделирование динамики лесных ресурсов. Число научных публикаций — 284. a_s_alekseev@mail.ru; 5, Институтский пер., 194021, Санкт-Петербург, Российская Федерация; р.т.: +7 (812)550-08-34; факс: +7(812)550-08-15.

DOI 10.15622/sp.18.2.416-443

D.M. CHERNIKHOVSKY, A.S. ALEKSEEV

DETERMINATION OF AVERAGE HEIGHTS AND WOOD STOCKS OF FOREST STANDS BASED ON INFORMATION PROCESSING OF TOPOGRAPHIC RADAR SURVEY, DIGITAL ELEVATION MODELS AND GIS TECHNOLOGIES

Chernikhovsky D.M., Alekseev A.S. **Determination of Average Heights and Wood Stocks of Forest Stands Based on Information Processing of Topographic Radar Survey, Digital Elevation Models and GIS Technologies.**

Abstract. The paper studies the possibilities of using global elevation models SRTM (Shuttle radar topographic mission) to assess the characteristics of forest stands – average heights and wood stocks. It is known that in process of radar shooting vegetation is considered as a barrier to correctly determining the height of the earth's surface. The surface, fixed by the sensors above the forest covered territory (scattering phase center height), is located in the upper part of the forest canopy. The review of publications confirms the relevance of this area of investigation in the world. A brief theoretical basis of the SRTM survey, factors determining the values of the heights and the errors associated with them are presented based on literature reviews. The possibility of determining the height of forest canopy based on evaluation models is shown.

The part of Uchebno-Opytnoe Forest District of the Leningrad region was chosen as the model territory. The geographic information databases, data of radar survey SRTM and topographic maps were the origin data for the study. The model territory is divided by a regular network into cells with 1 km step. Most of the territory is covered with forest vegetation. Relief modeling was performed on the basis of digitized topographic maps of 1:25000 scale by interpolation using TIN method. A visual and statistical comparison of both evaluation models – a surface model (based on SRTM radar survey data) and a relief model (based on topographic maps) was done. With help of the profiles construction an assessment of the nature of changes in the heights of the relief and surface models was performed. It is noted that for most of the model territory, the differences in height between the surface and relief models are 15-20 m. The convergence of graphs for heights is observed in areas, which are not covered with forest vegetation.

The set of linear regression dependencies between the scattering phase center heights (independent variable) and average heights or wood stocks (dependent variable) within the borders of forest compartments, grouped by the predominated tree species, was obtained. The influence on the closeness of the relationship and the value of the regression coefficient of such factors as the value of basal area and the share of predominant tree species in composition was found.

The established regularities can be used to improve the theory and practice of forest inventory, as well as to solve other problems related to the assessment of natural resources at the regional and global level (national forest inventory, carbon stock determination, assessment of biomass).

Keywords: Forest Management, Digital Elevation Model, Digital Surface Model, Scattering Phase Center Height, Regression Analysis.

Chernikhovsky Dmitry Mikhailovich — Ph.D., Associate Professor, Department of Forest Inventory, Management and GIS, St. Petersburg State Forest Technical University (SPbFTU); director, Center for the development remote sensing methods in forestry, Lesproekt LLC. Research interests: forest inventory and management, remote sensing in forestry, geoinformatics, geomorphometry. The number of publications — 65. cherndm2006@yandex.ru; 5, Institute per., 194021, St. Petersburg, Russian Federation; office phone: +7 (812)550-08-34; fax: +7(812)550-08-15.

Alekseev Alexander Sergeyevich —Ph.D., Dr.Sci., Professor, Head of Department, Department of Forest Inventory, Management and GIS, St. Petersburg State Forest Technical University (SPbFTU). Research interests: forest inventory and management, forest ecology and monitoring,

GIS technology for forest inventory and management, mathematical modeling of forest resource dynamics. The number of publications — 284. a_s_alekseev@mail.ru; 5, Institute per., 194021, St. Petersburg, Russian Federation; office phone: +7 (812)550-08-34; fax: +7(812)550-08-15.

References

1. Zhirin V.M., Knjazeva S.V., Eidlina S.P. [Dynamics of the spectral brightness of the age-specific structure of forest type groups on LANDSAT satellite images]. *Lesovedenie – Russian Journal of Forest Science*. 2014. vol. 5. pp. 3–12. (In Russ.).
2. Zhirin V.M., Knjazeva S.V., Eidlina S.P. [Estimation of the influence of tree canopy morphology and relief on the spectral characteristics of forests according to Landsat data]. *Issledovanie zemli iz kosmosa – Study of Earth from Space*. 2016. vol. 5. pp. 10–20. (In Russ.).
3. Tolkach I.V., Saevich F.K. [Spectral and brightness characteristics of the main forest-forming species in the images of the LEICA ADS100 scanner]. *Trudy BGTU. Lesnoe hozjajstvo – Proceedings of BSTU. Forestry.* 2016. vol. 1(183). pp. 24–27. (In Russ.).
4. Tolkach I.V. et al. [Patterns of variability of the spectral brightness of the canopy of the main forest-forming species of Belarus in the ADS 100 scanner images]. *Trudy BGTU. Serija 1: Lesnoe hozjajstvo, prirodopol'zovanie i pererabotka vozobnovljaemyh resursov – Proceedings of BSTU. Forestry, nature management and processing of renewable resources*. 2017. vol. 2(198). pp. 43–49. (In Russ.).
5. Soldatenkov A.A. [Interpretation of the composition of forest vegetation in the mid-mountain relief]. *Vestnik Adygejskogo gosudarstvennogo universiteta – The Bulletin of Adyghe State University.* 2014. vol. 1(133). pp. 127–130. (In Russ.).
6. Sidorenkov V.M. et al. [Zoning of the territory of the Udmurt Republic by categories of habitat for hunting resources based on Landsat 8 OLI-TIRS satellite survey data]. *Lesotehnicheskij zhurnal – Forest engineering journal*. 2015. Issue 5. vol. 3(19). pp. 84–93. (In Russ.).
7. Sharikalov A.G., Yakutin M.V. [Analysis of the state of taiga ecosystems using automated interpretation techniques]. *Izvestija Altajskogo gosudarstvennogo universiteta. Nauki o Zemle – Izvestiya of Altai State University Journal.* 2014. vol. 3-1(83). pp. 123–127. (In Russ.).
8. Perepechina Yu.I., Glushenkov O.I., Korsikov R.S. [Accounting and assessment of forests originating on agricultural land using remote sensing data]. *Izvestija vysshih uchebnyh zavedenij. Lesnoj zhurnal – The bulletin of higher educational institutions. Foresty Journal.* 2016. vol. 4(352). pp. 71–80. (In Russ.).
9. Perepechina Yu.I., Glushenkov O.I., Korsikov R.S. [Determination of forest cover and quantitative characteristics of forests using Sentinel-2 satellite images (on the example of the Shebekinsky municipal district of the Belgorod region)]. *Lesohozjajstvennaja informacija – Forestry information.* 2017. vol. 4(4). pp. 85–93. (In Russ.).
10. Belova E.I., Ershov D.V. [Experience in estimating natural reforestation on clear-cuts by time series] *Lesovedenie – Russian Journal of Forest Science*. 2015. vol. 5. pp. 339–345. (In Russ.).
11. Chernyh D.V., Birjukov R.Yu., Zolotov D.V., Vagner A.A. [Anthropogenic modifications and transformations of landscapes in the r. Kasmala: classification and dynamics based on remote sensing data]. *Vestnik Altajskoj nauki – Vestnik Altajskoy nauki.* 2014. vol. 1(19). pp. 233–240. (In Russ.).
12. Soromotin A.V., Brodt L.V. [Vegetation monitoring during the development of oil and gas fields according to LANDSAT multizone survey]. *Vestnik Tyumenskogo gosudarstvennogo universiteta. Ehkologiya i prirodopol'zovanie – UT Research Journal. Natural Resource Use and Ecology.* 2018. Issue 4. vol. 1. pp. 37–49. (In Russ.).
13. Terehov A.G., Makarenko N.G., Pak I.T. [Automatic algorithm for the classification of Quickbird images in the task of assessing the completeness of the forest]. *Komp'juternaja optika – Computer Optics.* 2014. Issue 38. vol. 3. pp. 580–583. (In Russ.).
14. Savchenko A.A., Vyvodcev N.V. [Assessment of the possibilities of using remote sensing data in monitoring the sanitary and forest-pathological state of forests].

Uchenye zametki TOGU – Electronic scientific journal "Scientists notes PNU". 2015. Issue 6. vol. 4. pp. 658–661. (In Russ.).
15. Balenović I., Seletković A., Pernar R., Jazbec A. Estimation of the mean tree height of forest stands by photogrammetric measurement using digital aerial images of high spatial resolution. Annals of Forest Research. 2015. vol. 58. no. 1. pp. 125–143.
16. Arhipov V.I., Chernihovskij D.M., Berezin V.I., Belov V.A. [Modern technology of forest mensuration by interpretation method «From survey – to the project»] *Izvestija Sankt-Peterburgskoj lesotehnicheskoj akademii – News of the Saint Petersburg State Forest Technical Academy.* 2014. vol. 208. pp. 22–42. (In Russ.).
17. Alekseev A.S., Mihajlova A.A., Chernihovskij D.M., Berezin V.I. [Method for determining the taxation characteristics of plantations from ultra-high-resolution aerial photographs]. *Trudy Sankt-Peterburgskogo nauchno-issledovatel'skogo instituta lesnogo hozjajstva – Proceedings of the Saint Petersburg Forestry Research Institute.* 2017. vol. 2. pp. 67–77. (In Russ.).
18. Faganan M., De Fries R. Measurement and monitoring of the world's forests. A review and summary of remote sensing technical capability, 2009-2015. Resources for the Future. 2009. 131 p.
19. Balenović I., Alberti G., Marjanović H. Airborne laser scanning – the status and perspectives for the application in the south-east European forestry. *South-east European forestry*. 2013. vol. 4. no. 2. pp. 59–79.
20. Kauranne T. et al. Airborne Laser Scanning Based Forest Inventory: Comparison of Experimental Results for the Perm Region, Russia and Prior Results from Finland. *Forests*. 2017. vol. 8. no. 3. pp. 72.
21. Peuhkurinen J. et al. Predicting Tree Diameter Distributions from Airborne Laser Scanning, SPOT 5 Satellite, and Field Sample Data in the Perm Region, Russia. *Forests*. 2018. vol. 9. no. 10. pp. 639.
22. Gašparović M., Milas A., Seletković A., Balenović I. A novel automated method for the improvement of photogrammetric DTM accuracy in forests. *Šumarski list*. 2018. vol. 142. no. 11-12. pp. 567–576.
23. Balenović I., Milas A., Marjanović H. A comparison of stand-level volume estimates from image-based canopy height models of different spatial resolutions. *Remote Sensing*. 2017. vol. 9. no. 3. pp. 205.
24. Balenović I. et al. Quality assessment of high density digital surface model over different land cover classes. *Periodicum biologorum*. 2016. vol. 117. no 4. pp. 459–470.
25. Rabus B., Eineder M., Roth A., Bamler R. The shuttle radar topography mission – a new class of digital elevation models acquired by spaceborne radar. *ISPRS Journal of Photogrammetry and Remote Sensing.* 2003. vol. 57. no. 4. pp. 241–262.
26. Köthe R., Bock M. Development and use in practice of Saga modules for high quality analysis of geodata. FREE AND OPEN GIS-SAGA-GIS. 2006. vol. 115. pp. 85–96.
27. Conrad O. et al. System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. *Geoscientific Model Development.* 2015. vol. 8. no. 7. pp. 1991–2007.
28. Florinskij I.V. [An illustrated introduction to geomorphometry]. *Ehlektronnoe nauchnoe izdanie Al'manah Prostranstvo i Vremja – Electronic Scientific Edition Almanac Space and Time.* 2016. Issue 11. vol. 1. Available at: 2227-9490e-aprovr_e-ast11-1.2016.71 (accessed: 05.12.2018). (In Russ.).
29. Sharyi P.A. [Geomorphometry in Earth and Ecology Sciences, a review of methods and applications]. *Izvestiia Samarskogo nauchnogo tsentra Rossiiskoi akademii nauk – Izvestia of RAS Samara Scientific Center of the Russian Academy of Sciences*. 2006. Issue 8. vol. 2. pp. 458–473. (In Russ.).
30. Hengl T., Reuter H.I. Geomorphometry: Concepts, Software, Applications. Newnes. 2008. vol. 33. 772 p.
31. Alekseev A.S., Nikiforov A.A. [The influence of relief on the structure and productivity of forest landscapes using 3D-modeling on the example of Lisinsky training and experimental forestry]. *Lesovedenie – Russian Journal of Forest Science.* 2014. vol. 5. pp. 42–53. (In Russ.).

32. Chernihovskij D.M., Alekseev A.S. [The influence of the relief surface shape on the structure and productivity of forest landscapes on the example of the reserve «Verkhne-Tazovsky» Yamalo-Nenets Autonomous area]. *Lesovedenie – Russian Journal of Forest Science*. 2003. vol. 5. pp. 10–17. (In Russ.).
33. Farber S.K. [Structuring of forest communities]. *Sibirskij Lesnoj Zurnal – Siberian Journal of Forest Science.* 2014. vol. 1. pp. 35–49. (In Russ.).
34. Rahmatullina I.R., Rahmatullin Z.Z., Mustafin R.F. [The distribution and productivity of pine stands depending on the morphometric parameters of the relief (on the example of Bugulminsko-Belebeevsky upland within the Republic of Bashkortostan)]. *Vestnik Izhevskoj gosudarstvennoj sel'skohozjajstvennoj akademii – The Bulletin of Izhevsk State Agricultural Academy.* 2017. vol. 1(50). pp. 42–52. (In Russ.).
35. Chernihovskij D.M. [Assessment of interrelations of morphometric characteristics of relief with quantitative and qualitative characteristics of forests]. *Izvestija Sankt-Peterburgskoj lesotehnicheskoj akademii – News of the Saint Petersburg State Forest Technical Academy.* 2016. vol. 216. pp. 69–90. (In Russ.).
36. Chernihovskij D.M. [Assessment of the relationships between morphometric characteristics of relief with quantitative and qualitative characteristics of forests using ASTER and SRTM digital terrain models]. *Sibirskij Lesnoj Zurnal – Siberian Journal of Forest Science.* 2017. vol. 3. pp. 28–39. (In Russ.).
37. Chernihovskij D.M. [Automatic classification of surface topography to the quantitative and qualitative characteristics of forests]. *Izvestija Sankt-Peterburgskoj lesotehnicheskoj akademii – News of the Saint Petersburg State Forest Technical Academy.* 2017. vol. 219. pp. 74–95. (In Russ.).
38. Chernihovskij D.M. [Using the automatic classification of relief by Ivahashi and Pike to assess the quantitative and qualitative characteristics of forests on the basis of elevation models of terrain and surface]. *Izvestija Sankt-Peterburgskoj lesotehnicheskoj akademii – News of the Saint Petersburg State Forest Technical Academy.* 2018. vol. 223. pp. 100–126. (In Russ.).
39. Kellndorfer J. et al. Vegetation height estimation from Shuttle Radar Topography Mission and National Elevation Datasets. *Remote Sensing of Environment*. 2004. vol. 93. pp. 339–358.
40. Gallant J.C. Read A.M., Dowling T.I. Removal of tree offsets from SRTM and other digital surface models. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. 2012. vol. 39. no. 14. pp. 275–280.
41. Sexton J. et al. A comparison of lidar, radar, and field measurements of canopy height in pine and hardwood forests of southeastern North America. *Forest Ecology and Management*. 2009. vol. 257. pp. 1136–1147.
42. Avtar R., Sawada H. Use of DEM data to monitor height changes due to deforestation. *Arabian Journal of Geosciences.* 2013. vol. 6. no. 12. pp. 4859–4871.
43. Sun G., Ranson K.J., Kharuk V.I., Kovacs K. Validation of surface height from shuttle radar topography mission using shuttle laser altimeter. *Remote Sensing of Environment.* 2003. vol. 88. pp. 401–411.
44. Ni W. et al. Co-Registration of Two DEMs: Impacts on Forest Height Estimation from SRTM and NED at Mountainous Areas. *IEEE Geoscience and Remote Sensing Letters*. 2014. vol. 11. no. 1. pp. 273–277.
45. Zhang Z. et al. Estimation of forest structural parameters from Lidar and SAR data. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. 2008. vol. 37. pp. 1121–1126.
46. Miliaresis G., Delikaraoglou D. Effects of Percent Tree Canopy Density and DEM Misregistration on SRTM/NED Vegetation Height Estimates. *Remote Sensing*. 2009. vol. 1. pp. 36–49.

DOI 10.15622/sp.18.2.444-470

D. PANEVA-MARINOVA, J. STOIKOV, L. PAVLOVA, D. LUCHEV

SYSTEM ARCHITECTURE AND INTELLIGENT DATA CURATION OF VIRTUAL MUSEUM FOR ANCIENT HISTORY

Paneva-Marinova D., Stoikov J., Pavlova L., Luchev D. **System Architecture and Intelligent Data Curation of Virtual Museum for Ancient History**

Abstract. Preserving the cultural and historical heritage of various world nations, and their thorough presentation is a long-term commitment of scholars and researchers working in many areas. From centuries every generation is aimed at keeping record about its labor, so that it could be revised and studied by the next generations. New information and multimedia technologies have been developed during the past couple of years, which introduced new methods of preservation, maintenance and distribution of the huge amounts of collected material. This article aims to present the virtual museum, an advanced system managing diverse collections of digital objects that are organized in various ways by a complex specialized functionality. The management of digital content requires a well-designed architecture that embeds services for content presentation, management, and administration. All elements of the system architecture are interrelated, thus the accuracy of each element is of great importance. These systems suffer from the lack of tools for intelligent data curation with the capacity to validate data from different sources and to add value to data. This paper proposes a solution for intelligent data curation that can be implemented in a virtual museum in order to provide opportunity to observe the valuable historical specimens in a proper way. The solution is focused on the process of validation and verification to prevent the duplication of records for digital objects, in order to guarantee the integrity of data and more accurate retrieval of knowledge.

Keywords: virtual museum, system architecture, functionality, data integrity, knowledge retrieval, data validation, record de-duplication, cultural heritage.

1. Introduction. For a long time, cultural heritage has been maintained in museums, galleries, libraries and research laboratories, where not everyone was able to access this wealth. Digital technologies that have been developed during the past couple of years introduced new solutions of documentation, maintenance and distribution of the huge amounts of collected material. Among these new technologies are virtual museums, which have already proven their worth as a contemporary conceptual solution for access to and attractive presentation of cultural archives. Virtual museums contain diverse collections of digital objects (such as text, images, and media objects) that are organized in various ways and are managed by complex specialized services such as content structuring and grouping, attractive visualization, advanced search (semantic-based search, multilayer and personalized search, context-based search), resources and collection management, indexing, semantic description, knowledge retrieval, metadata management, personalization and content adaptability, content protection and preservation, tracking services, etc. Thus, the valuable cultural heritage wealth is accessible anytime and anywhere, in a friendly, multi-modal, efficient, and affective way.

However, these systems suffer from the lack of tools and services for intelligent data curation with the capacity to add value to data. In this paper, a solution for intelligent data curation in a virtual museum is proposed in order to provide opportunity to observe and analyze valuable ancient history specimens in a proper way and in their historic context, so that some yet undiscovered treasured of the human civilizations be manifested [31]. This solution more specifically is focused on the validation and prevention of duplication of newly added or existing records for digital objects.

Section 2 of this paper discusses some challenges raised during the design and the development of virtual museums. Section 3 and 4 present current concepts for virtual museum system architecture, tracking main functionality and services supporting users' needs. Section 5 includes a discussion on intelligent data curation issues. In section 6, a model of intelligent data curation service is described. The paper ends with some conclusions and further development plans.

2. Virtual museum design issues. The development of the technologies during the last years provides new functionalities and advanced services to contemporary virtual museum (VM) transforming their static complex structures to environment with a dynamic federation of functional units. This change resulted from the needs of the market, the emergence of new technologies, and especially from the request for stricter use of the existing resources and adapting VMs content and services to the needs of different user groups.

Some key research questions, raised during the design and the development of these systems, are:

– How to present the selected resources in a given context and to determine the conditions and use cases – cognitive or educational goals, analysis, creative use, etc.?

– How to help the user not just to view, but to also gain knowledge?

– How to provide knowledge in the most suitable way and form?

– How to adapt the offered information content for each individual user or group in order to achieve their goals and tasks? [29]

– How to choose the most suitable resources for a specific situation and the method of introduction to the domain, which is subject to research, etc.?

The difficulties in solving these research issues are related to the lack of common model and working solutions regarding the basic and the extended functionality, and synchronizing the solutions with the existing standards and regulations in the area; analysis, understanding and better interpretation of digital cultural content; context-dependent use of digital cultural resources; increase and generalisation of visitor experience, contextual techniques for personalising visitor experience, etc. [30].

A considerable interest in this area in recent years is demonstrated by Bulgarian scientists. The main efforts are concentrated in applied aspects, especially for increasing the presence of digital artefacts and collections of the Bulgarian cultural and historical heritage in the global information space. Besides, work is done towards developing ICT tools and systems for digital presentation and preservation of cultural heritage artefacts. There is also intensified interest in fundamental research (priority areas of Informatics, ICT and Cultural Heritage of the Strategy for the Development of Science in Bulgaria till 2020, Innovation Strategy for Intelligent Specialisation, Horizon 2020, etc.) in search of innovations especially in areas/subareas relevant to data processing, access control, intelligent supervision, security, semantics, etc. The current research activities include the study and applications of new methods and tools for the creation, integration and development of innovative systems, managing digital cultural assets [1, 3, 4, 8]. The focus is in researching and exploitation of new or emerging technologies for the development of innovative products, tools, applications and services for the creative digital content production, usage and management. The aim is to transform cultural heritage into digital units, which integration and reuse through research-led methods will have high commercial potential for cultural institutions, tourism, and creative and media industries.

The innovation principles for VM require visual rules, that characterize the different kinds of visual symbols; data rules, that specify the characteristics of the data model, the database schema, and the database instances; -mapping rules, that specify the link between data and visual elements; methodologies and tools for the support of cross-language retrieval from the nodes of the federated architecture must be developed [32].

They include device independent access to a world-wide digital repositories, including interconnected publications and reports; Service's availability 24h/day, 7d/week; Reduced cost and difficulty of content distribution ; Possibility to merge knowledge thematically in one personalized book by making a selection among the large offer of digital content (articles, chapters, extracts, etc.); Optimized management of perishable contents thanks to the maintenance service; Reusability, archiving and availability of the digital content in the "memory of knowledge"; Possibility to create and expand links between small and specialized museum communities; Socio-economic Innovation principles [28].

VM deal with issues of co-operation within the context of an information society comprised of independent organisations with different rules, traditions, organisation structures, motivations for profit, nationality, laws, culture and languages [26, 30] (see Figure 1).

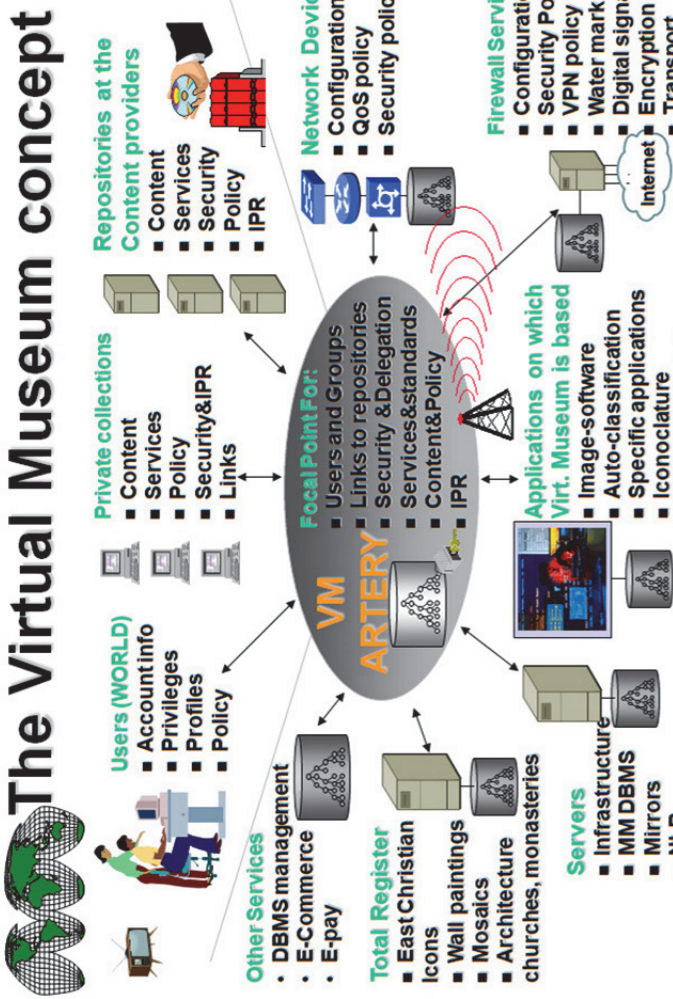


Fig. 1. Virtual Museum example concept

There are many challenges while working on the given task:

– A necessity for clear definition of the user's needs of some specific functionality;

– The presentation of information content in the most suitable way for the chosen user types, the content's ability to be easily found and reached;

– Assuring reusability of the resources in specific context and situation; adapting resources;

– Searching for flexible conceptual solutions, which are easily transferable and implementable via new technological means;

– Synchronization with established standards and specifications, etc.

These questions suggest deep research and analysis of the different components of the system – content, user needs, offered services and its applications. The following are of great importance:

– Building of a straightforward model/specification of the activities that the system will serve;

– Developing and introducing proper functionalities for ensuring flexible access to the resources;

– Analysis of the context in which the resources will be used (including educational one) and searching for methodological approaches and techniques for improving the access to the resources to meet the user needs to the highest extent.

3. Virtual museum system architecture. The virtual museum mainly contains service panels for *Museum content management*, *Museum content presentation*, *Administrative services* (see Figure 2), jointed to *a Media repository* and a *User data repository*.

The *Museum content management* module refers to the activities related to basic content creation: add (annotate and semantic indexing), store, edit, preview, delete, group, and manage multimedia digital objects; manage metadata; search, select (filter), access and browse digital objects.

The Museum content presentation module supports objects and collections display. It also provides collections creation (incl. search, select/browse and group multimedia digital objects according to different criteria and/or context of usage), their metadata/semantic descriptions and attractive visualization, status of collection display.

Content presentation module aims to provide access to all virtual museum services through wide range of contemporary technologies and devices – not only desktop PCs, but mobile phones, tablets, TVs, VR devices, etc. Interactive media technologies are used to provide best user experience within the content of the virtual museum.

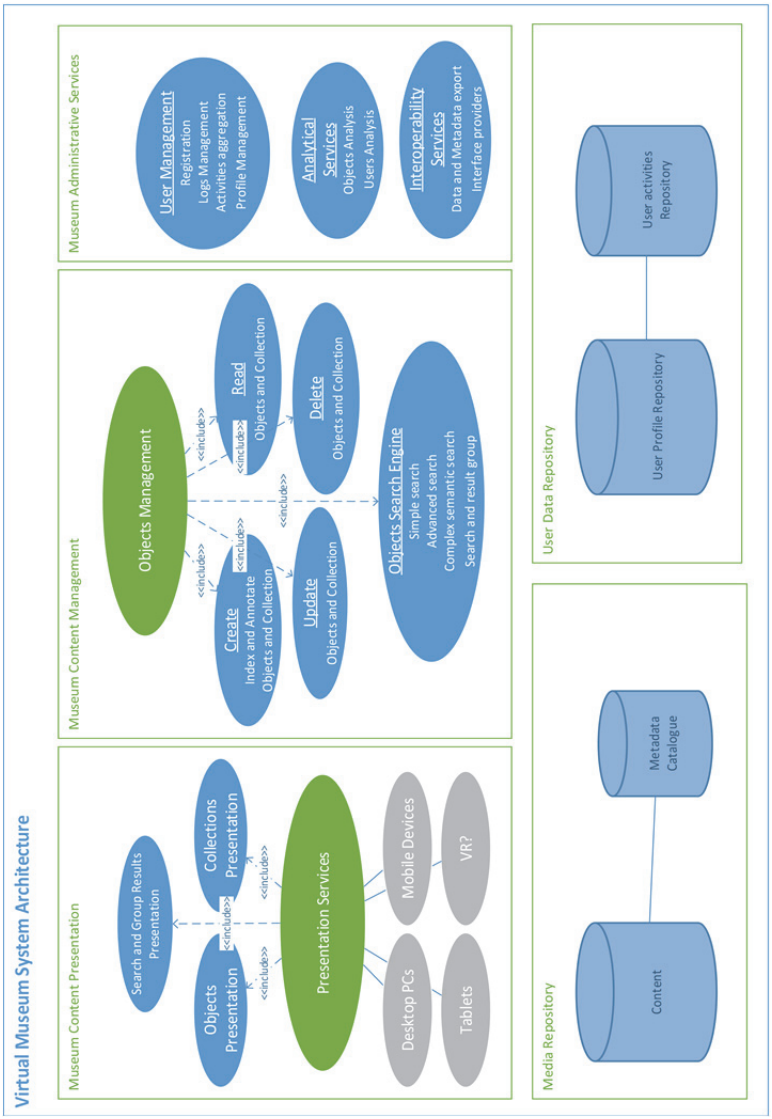


Fig. 2. Virtual Museum System Architecture

The *Administrative services panel* mainly provides user data management, data export, tracking and analysis services.

For every object all semantic and technical metadata are saved in the Media repository. These metadata are represented in catalogue records that point to the original media file/s associated to every object.

The User profile repository manages all user data and their changes.

4. Virtual museum functionalities in details.

4.1. *Museum content management.* The main part of the content creation process is the annotation and semantic indexing of digital objects in order to add them to the museum media repositories. The entering of technical and semantic metadata for the digital objects is implemented through different automated annotation and indexing services.

The technical metadata, that could be expressed in Dublin Core or other standards, are attached to every multimedia object automatically. They cover the general technical information, such as file type and format, identifier, date, provider, publisher, contributor, language, rights, etc.

An annotation template needs to be implemented for the semantic description of digital objects. The template provides several options for easy and fast entering of metadata:

– Autocomplete services (All used (already entered) field values are available in a panel for reuse.);

– Automated appearance of dependencies coming from the relations of the defined classes' (concepts) in an ontological descriptive structure valid for the museum object domain. All main relations and rules expressed in the ontological structure are incorporated during the development of the annotation template;

– Bilingual data entering with automated relation between the relevant values in different languages (if it is applicable or necessary);

– Automated appearance of the number of the used field value, providing regular data tracking (if it is applicable or necessary);

– A tree-based structure of the annotation template. Only checked fields are displayed for entering metadata;

– Possibility for adding more than one media for one metadata description in order to create rich heterogeneous multimedia digital objects tracking (if it is applicable or necessary);

– Reuse of an already created annotation for new objects: the new media object has to replace the older one, the annotation is kept and the new object appears after saving;

– Automated watermarking of the image and video objects;

– Automated resizing and compression of the media objects (image or video);

– Automated identification of file formats;

– Automated conversion of the media object (audio, video, text) in a format suitable for Web-preview;

– Automated terms explanation (if it is applicable or necessary): After saving a new digital object in the museum media repository, a special machine traces for the appearance of dictionary terms in the object metadata description. If some terms are available, the machine adds links to their explanations. In the case of entering a new dictionary term, its presence in the available objects is discovered automatically and a link is added.

– Digital object duplication checks (similarities calculation): In order to avoid duplicate image objects a service that checks the similarity between images is provided. It uses an algorithm that caching images for optimizing their compare (see [1]).

The virtual museum provides a wide range of search services, such as keyword search, extended keyword search, semantic-based search, complex search, search with grouping results, etc. Their realization is based on querying action to the metadata knowledge base. Moreover, five types of conditions for the results set are meant:

– "objects having or NOT characteristic c";

– "objects having value $=, \neq, \leq, \geq, <$ or $>$ for characteristic c". In the search templates, the user could search digital objects with précised criteria.

The search services support content request and delivery via index-based search and browse of managed content and its description.

4.2. *Museum content presentation*. Content presentation is a key activity in the virtual museum. The proper design and intelligent implementation of this service provide a stable base for overall VM functionality.

During the design of the content presentation services a profound analysis was made of content selection and preview possibilities in order to satisfy the user's needs. First, it is necessary to determine the preview possibilities of a separate digital object and its components and after that the preview of grouped objects (collection preview).

The visualization of the rich semantic description of the separate digital object is determined through hidden parts appearing in a new window after link selection. This possibility is used mainly for the long descriptions and for the dictionary terms. Parts of the descriptive data field are hidden, but their values are available for searching in special forms.

During the design of object grouping services, the main ontology classes of the object descriptive structure (*viz.* museum domain ontology) could be selected as object grouping criteria.

Every user can create his private collection of selected objects after search activity. Rich search possibilities (mentioned above) are avail-

able in order to assist collection creation. The user can write the collection's title and short description. He can also select its status: private or shared with other users. New objects for a collection appear automatically after their entering.

Custom collections in virtual museum are dynamic objects. They are criteria based, not list based. Every new object in the virtual museum will be automatically added to a collection, if it meets the criteria defined for the collection. The owner/creator/follower of a collection can be notified when a new object is added and becomes part of the collection. This service uses the following rule:

Let $P = \{p_1, p_2, ..., p_n\}$ be the set of all iconographical objects.

Let $A_m = \{a_{m1}, a_{m2}, \ldots, a_{mk}\}$ be the set (also called a collection) with k iconographical objects with a selected characteristic m, $A_m \subseteq P$.

Let p_{ti} be a new iconographical object, added to the library, $P = P \cup \{p_{ti}\}$.

IF $t \equiv m$ THEN $A_m = A_m \cup \{p_{ti}\}$.

Let $M = \{m_1, m_2, ..., m_r\}$ be a set of characteristics for a collection $A_M = \{a_{M1}, a_{M2}, ..., a_{Mk}\}$ with k iconographical objects.

Let p_{ti} be a new iconographical object, added to the library, $P = P \cup \{p_{ti}\}$, and $M' = \{m_1', m_2', ..., m_r'\}$ be its set of characteristics.

IF $M \subseteq M'$ THEN $A_M = A_M \cup \{p_{it}\}$.

The home page of the library contains a panel with last visited objects, aiding the user's observation of the content. This service uses the following algorithm:

Let p_i be an iconographical object, and $P = \{p_1, p_2, ..., p_n\}$ be the set of all iconographical objects.

Let t_j be the time an object was visited $T = \{t_1, t_2, ..., t_m\}$

$Q = P \times T$

(p_i, t_j) means that the object p_i was visited at the time t_j.

Steps of the algorithm:

1. Create series $Q' = \{(p_{i_1}, t_{j_1}), (p_{i_2}, t_{j_2}), ..., (p_{i_d}, t_{j_d})\}$, where $t_{j1} > t_{j2} > .... > t_{jd}$.

2. Remove all (p_{i_k}, t_{j_k}), where $\exists t_{j_l} : l < k \;\&\; p_{i_k} \equiv p_{j_l}$.

3. Select first $\{q_1, ..., q_v\} \in Q'$.

Every object and collection can be presented in interactive way using any browser compatible device – PC, phone, tablet, or TV. Content presentation services are based on responsive web technologies in order to satisfy the majority of the modern devices diversity and provide great user experience no matter if user consumes the virtual museum services through their phone, tablet, PC, or other smart device. Moreover, the virtual museum should provide technological solutions for people with disabilities [22, 23, 24].

4.3. *Administrative Services*. The Administrative services panel mainly provides user data management, data export, tracking services, and analysis services. The user data management covers the activities related to registration, data changes, level set, and tracking activities of the user. The tracking services have two main branches: tracking of objects, tracking of user' activities. The tracking of objects spies on the activities of add, edit, preview, search, delete, selection, export to XML, and group of objects/collections in order to provide a wide range of statistic data (for frequency of service use, failed requests, etc.) for internal use and generation of inferences about the stable work (stability) and the flexibility of the work and the reliability of the environment. The tracking of users' activities monitors user logs, personal data changes, access level changes and user behaviour in the museum environment. The QlickTech® QlinView® Business Intelligence software could be used as an analysis provider. Therefore, it needs to be connected to the museum tracking services and objects data base by preliminary created data warehouse. It will provide fast, powerful and visual in-memory analysis based on online analytical processing and quickly answered multi-dimensional analytical queries [2]. This information can be used for making conclusions about people's interest in objects, collections and the museum content, in order to further fill the repository of the museum.

The export data services provide the transfer of information packages (for example, packages with digital objects/collections, user profiles, etc.) compatible with other data base systems. For example, with these services a package with objects could be transported in an XML-based structure for new external use in e-learning or e-commerce applications. Moreover, VMs power will increase significantly if they use mechanisms for ubiquitous sharing of their e-artefacts and they distribute attractive content in the social networks, reflecting community demands and needs [27].

During the VM system design the interoperability services and protocols need to be have in mind, *viz.*:

– Services concerning to interoperability and integration - describe the ways in which repositories work with other systems using common standards and protocols. Sometimes these interfaces are used directly by people (e.g. web user interfaces or RSS feeds) and sometimes they are used

by machines (e.g. OAI-PMH and SWORD). Interfaces used by machines are sometimes referred to as m2m (machine-to-machine) interfaces;

– Services supporting linking mechanism - for effective use of distributed electronic resources in libraries. Some examples are Open URL linking, Link resolvers, Electronic resource integration, DOI, CrossRef, Handle.Net. Linking mechanism makes possible to build global museum services and portals, because it provides unique item identifiers, persistent identifiers are used for citation management, etc.;

4.4. *Storage and long term preservation of digital information.* Storage Management gives many benefits:

Digital content relies on common standards for metadata, storage data formats, indexing, etc. with necessity for provision of support for the whole live cycle. The term storage management encompasses the technologies and processes organizations use to maximize or improve the performance of their data storage resources. It's a broad category that includes virtualization, replication, mirroring, security, compression, traffic analysis, process automation, storage provisioning and related techniques. Storage management techniques can be applied to primary, backup or archived storage. Deployment and implementation procedures will vary widely depending on the type of storage management selected and the vendor. In addition, the skills and training of storage administrators and other personnel add another level to an organization's storage management capabilities. There are numerous storage technologies that impact the storage and long term preservation of digital information design, following principles of consolidation of storage into a central location, removal of the storage burden from host OSes, and so on.

Technologies, like storage virtualization, deduplication and compression, allow better utilization of existing storage, resulting in lower costs for operating and maintain storage devices. They simplify the management of storage networks and devices, reducing overall storage operating costs.
The appropriate storage management improves digital data reliability, performance, availability, agility and resilience.

Technologies like replication, mirroring and security are often particularly important for backup and archive digital information. Capacity optimization technologies like parity RAID, delta snapshots, thin provisioning etc. are applicable to storage of both structured and unstructured data.

Over them polices according to systems and software management, physical security, data security, data backups, disaster recovery, redundancy of data (multiple data duplication, digital archives, global web portals, providing content aggregation from various sources distributed over the Internet), are applied.

The emerging new generation of information technologies is a synergy of business intelligence analytics, from personalization towards making data-driven decisions and forecasts providing an integrated business solution gradually alienated from the software toward services and functionalities offered to the users.

The important services in the contemporary virtual museum are: content creation and presentation, crawling, storage, browse, measurement, retrieval, classification/ categorization, filtering, clustering, summarization, mining, preservation, decision support, user modelling/ personalization, etc. A main task for the developers is the proper design and intelligent implementation of these services. The design and the implementation of the described services result from a long-term observation of the users' preferences, cognitive goals, needs, object observation style, and interests, made during the testing processes in several digital content management systems. The main goal was the satisfaction of users' preferences and needs with appropriate navigation, visualization and content presentation techniques.

The authors actively try to develop workable solutions in the field of cultural heritage database management and presentation. The above described solution for virtual museum architecture follows previous developments in the digital content management systems (*viz.* digital libraries, digital repositories, galleries, etc.) for Bulgarian artworks and treasuries (see [3-9]). These systems are successfully implemented to presents the valuable Bulgarian cultural heritage: Bulgarian iconographic art, Bulgarian ethnographic and folklore artefacts, medieval and early modern Bulgarian texts for saints in combination with ethnological data and visual sources, church bells and plates, etc.

4.5. *Security of the Virtual Museum.* The Virtual Museum hosts digital data containing core assets, including user's information, intellectual property, and other critical content. With emerging trends such as Big Data, bring-your-own-device (BYOD) mobility, and global online collaboration sparking an explosion of data, the VM will only become more important and will extending be the target of advanced malware and other cyber attacks [25]. What is needed to be done to secure the Virtual Museum is:

– To shield the VM from advanced persistent threats (APTs) and sophisticated malware found in content stores, web and application servers, and common file shares.

– To stop attacks entering organizations via mobile devices and portable storage.

– To receive on-target analysis to pinpoint possible gaps that need addressing.

– To protect key assets and prevent attacks with products and services that work together and share and threat intelligence.

– To prevent attacks with a nimble, adaptive cyber security strategy.

– To safeguard VMs from attacks that use web servers and other data center infrastructure to host malware.

– To detect threats quickly to reduce lag time before resolution.

– To get reliable, fast malware analysis with agentless network-based threat detection and protection engine.

– To provide continuous, dynamic, non-disruptive resolution to incidents.

Recommended for Virtual Museum architecture is the Adaptive Defense approach to cyber security, which delivers technology, expertise, and intelligence in a unified, nimble framework, which demands the adaptation of the security architecture to prevent today's cyber attacks and avert their worst effects.

5. Intelligent data curation. In the modern era of big data, the curation of data has become more prominent, particularly for software processing high volume and complex data systems [10]. The term is also used in historical uses and cultural heritage digital assets and content management solutions, where increasing cultural and scholarly data from digital projects requires the expertise and analytical practices of data curation. In broad terms, curation means a range of activities and processes done to create, manage, maintain, and validate a component. Specifically, data curation is the attempt to determine what information is worth saving and for how long [11]. The essential elements of a powerful data curation tool are annotations, metadata, standards, models, databases, etc.

Moreover, data curation is intellectually intensive activity that is time consuming and requires a lot of dedicated resources. Taking into account the increasing role and amount of data, curation risks to be a bottleneck for any digital asset management or content management project in the long term. One of the challenges for the automation of data curation is difficulty in completing missing data and the level of granularity. Such a solution, however, looks practical because the data curation process is one of many iterations, consistency and includes complex data evaluation. The human and machine aspect need to be combined in order to solve the two most crucial data-integration problems: linkage of records (which often refers to linking records across disparate sources, referring to the same real-world entity) and schema mapping (mapping columns and attributes of different datasets).

One approach is to use a record to modularize curation processes. Splendiani [12] considers curation activities as functions in a "curation

space" that is exemplified via a "curation record". The curation process is broken down to the following classes of operations:

– *Schema mapping*: Machine-assisted process to identify and map the similar attributes from different data sources together in one, unified data set. The same entities (*e.g.,* events, studies, places) might be described by data sources of different origin in separate ways and in this case the usage of different schemas and vocabularies (a dataset schema is generally an official description of the main attributes and the values that can be taken by them). For instance, one source may refer to a person's credentials by the means of two attributes (Name and Title), another source may use the terms Pers. Name and Royal Title, and a third might use PN and Rank, in order to address the same thing. The major activity in schema mapping is to set a mapping among those attributes. The problem may occur to be more challenging and may involve different conceptualizations especially in the cases when relationships in one source are represented as entities in another. Most often, in the ETL suites are used the most common schema mapping solutions that focus traditionally on the mapping of a small number of such schemas (usually less than ten) that deliver to users a suggested mapping that considers some similarity among column's name and the content of them. With the maturity of the big data stack, however, the enterprises have the power to easily acquire a huge number of different data sources and have at their service applications that can ingest data sources as they are generated. An example from the pharmaceutical industry and the conducted clinical studies can be used, where tens of thousands of studies and assays are conducted by scientists across the world, often using separate technologies and a combination of local schemas and standards. It is essential for the companies' businesses and is often required as mandatory by regulations and laws to use standardized and cross-mapping collected data. This approach has changed the main assumption of most solutions for schema mapping that the suggestions curated by users should be part of a manual process. In such a case the main encountered challenges are: (1) providing of automated solution that requires reasonable interaction with the user, meanwhile being able to map numerous schemas; and (2) designing of matching algorithms that are robust enough to accommodate different languages, formats, reference master data, and data units and granularity [13].

– *Standard setting*: Building a probabilistic machine learning model specific to the organization's domain and stakeholders based on answering a series of yes/no questions to whether two records are the same. Given enough feedback, a pattern is captured that is required to build and maintain logic in order to generate de-duplicated, master data.

– *Validation*: Throughout the human-guided process to build a machine learning model for mastering data, the user is able to see measurable

outputs for each item of yes/no feedback provided. The feedback directly corrects the model. This calculation is culminated in the 'confusion matrix', indicating the precision, recall, accuracy, and F score of the model based on human feedback [14].

Further defining the data curation process, it is based on the organization and integration of data collected from various sources. Data curation includes "all the processes needed for principled and controlled data creation, maintenance, and management, together with the capacity to add value to data" [15]. For example, in science, data curation may indicate the process of extraction of important information from scientific texts, such as research articles by experts, to be converted into an electronic format [16]. Using an efficient master data management solution can greatly facilitate the above-described process. However, a step further in the mastering process is, providing the ability to select a representative record, or golden record, for each set of duplicate, or grouped, data records derived from all data sources. For example, if a cluster of records is identified across systems for the same artist, but each record has a variation on the artist's name the human-guided machine learning approach merges those records and generate a single golden record using the most common values for each attribute - assuming that aligns with how the business perceives their data. The goal of the golden record is to consolidate and generate a single record of truth. This approach is especially applicable when update of existing record is required from several different data sources.

Summing it all up, the curation process can be expressed in terms of rules that embed "atomic operations" like extractors, transformations, etc. The rules can rely on abstraction/inferences for higher genericity and can also be used to produce meta-information [12].

6. Model for intelligent data curation in virtual museum. *Record de-duplication issues.* The linkage of records, the resolution of entity and the deduplication of records are only a few of the terms that describe the need for unification of multiple mentions or database records that describe the same real-world entity. Concerning the example in Table 1 (showing a single schema for simplicity), it is obvious that the records are about Alexander, but they look quite different [17, 18, 19, 20].

Actually, all these records are correct or were correct at some point in time. It is easy for a well-qualified human to determine if such a cluster refers to the same entity, but it is hard for a machine to conduct this judgement. Therefore, more robust algorithms should be utilized to find such matches in the presence of errors, different styles of presentation and mismatches of granularity and references of time.

Table 1. Data unification at scale

Name	Attribute	Title	Year
Alexander			356 – 323 BC
Alexander	the Great		356 – 323 BC
Alexander	of Macedon		356 – 323 BC
Alexander	III		356 – 323 BC
Alexander	III the Great		356 – 323 BC
Alexander	III of Macedon		356 – 323 BC
Alexander	(all atributes)	King of Macedonia	336 – 323 BC
Alexander	(all atributes)	Basileus of Macedonia	336 – 323 BC
Alexander	(all atributes)	Hegemon of Hellenic League	336 - 323 BC
Alexander	(all atributes)	Pharaoh of Egypt	332 – 323 BC
Alexander	(all atributes)	King of Persia	330 – 323 BC
Alexander	(all atributes)	Lord of Asia	331 – 323 BC

The issue is an old one. In the recent decades, the community that conducts researches has come up with many similarity functions, supervised classifiers in order to differentiate matches from non-matches, and clustering algorithms for collecting matching pairs in the same group. Current algorithms can deal with thousands of records (or millions of records partitioned in disjointed groups of thousands of records), similar to schema mapping. Taking into account the massive amount of collected dirty data – and in the context of the abovementioned schema-mapping problem, a number of challenges are faced:

Challenge One: How to scale the quadratic problem (comparing every record to all other records, so computational complexity is quadratic in the number of records).

Challenge Two: How to train and build machine learning classifiers that handle the subtle similarities as in Table 1.

Challenge Three: How to engage humans and domain experts in providing training data, given the nature of the matches, which are rare in most cases.

Challenge Four: How to leverage the knowledge of all domains and previously developed rules and matchers in one integrated tool.

Talking about similarity, both the problems of schema mapping and deduplication occur after finding matching pairs (attributes in the case of schema mapping and records in the deduplication case).

Most of the building blocks can be reused and leveraged for both problems. Regarding correlation, most record matchers depend on some known schema for the two compared records; however, unifying schemas requires some type of schema mapping, even if incomplete. For this reason and many other, the solution at hand is for consolidating these activities and devising core matching and clustering building blocks for the unification of data that could: (1) be leveraged for different activities for unification (in order to avoid piecemeal solutions); (2) scale to a massive number of sources and data; and (3) have human in the loop as a guiding driver of the machine in building classifiers and applying the unification at large scale, in a trusted and explainable way. The idea is to use a human in the loop to resolve ambiguities when the algorithm's confidence on a match falls below a threshold [13].

When extracting data from different sources in cases of initial data upload or record updates, with large masses of data exists the risk of accumulating a lot of duplicate records. In this section will be presented a solution approach for deduplication. The first step is to look for mechanisms to enrich the data. In this way, extra fields can be added to each record which can assist in the deduplication process.

As a result of this step, each record has K attributes of information. The next step depends on the availability of training data. This consists of a collection of pairs of records which a human specifies as matches (*i.e.* duplicates) and a collection of pairs of records that are non-matches.

To this step fits a decision tree model in the following manner. For each attribute is requires a distance function, D (a1, a2), which specifies how far apart are any two values a1 and a2. In general, a distance function can be user-specified. However, for each character string attributes, Jacard and cosine similarity distance are popular metrics, and a human is asked to choose between these two. For numeric data are used arithmetic distance. For each attribute, is chosen a collection of split points based on dividing the training data into L equal sized buckets. Then, for each attribute it tries these L "split points ", and avidly chooses the attribute and the split point that most accurately classifies the training data. In effect each of the L* K cases is a predicate of the form:

Attribute-I < split point => non-match
Attribute-I >= split point => match
And
Attribute-I >= split point => non-match
Attribute-I < split point => match.

After that is selected the predicate that best fits the data at hand. With this "root node" chosen, continues the fit of the two second level nodes. It continues in this fashion until the benefit of additional levels is marginal or

until a user-defined maximum depth, Max, is reached. In effect a decision tree model is fitted to the training data, with parameters D, L and Max.

If there is not enough training data active learning is used to get more. A "cluster review" process can also be employed. This step allows a human to review suggested matches and to correct ones that are in error. Hence, cluster review produces additional training data to refine the model used, and can be thought of as an active learning scheme.

So far are identified collections of records that it thinks represent the same entity, *i.e.* are duplicates. Consider one particular collection and resources that represents Alexander (356 – 323 BC) – a king (basileus) of the ancient kingdom of Macedon, as shown below.

Name:

Alexander the Great (Greek: *Αλέξανδρος ὁ Μέγας*, Bulgarian: *Александър Велики*)

Alexander of Macedon (Greek: *Αλέξανδρος ὁ Μακεδών*, Bulgarian: *Александър Македонски*)

Alexander III (Greek: *Αλέξανδρος Γ΄*, Bulgarian: *Александър III*)

Title with period of reign:

(Alexander) King of Macedonia (336 – 323 BC)

(Alexander) Basileus of Macedonia (336 – 323 BC)

(Alexander) Hegemon of Hellenic League (336 BC)

(Alexander) Pharaoh of Egypt (332 – 323 BC)

(Alexander) King of Persia (330–323 BC), etc.

Apparently, a canonical form for name, a resolution for several values for the title of the ruler that are attached to the name, and the recognition that have several different periods of reign, are requested.

First, user-specified column rules which define how to aggregate the column values in a cluster into a "golden value", are used. Also supported are the options to "choose the most frequent value", "majority consensus", "keep all values" and "choose average value". Based on applying these rules, each cluster of data is reduced to a simpler one with less multi-valued attributes.

Then, is examined each column, looking for patterns of values. For example, in the Alexander cluster, it removes the duplicate value “Alexander” and is left with:

III (The Third)

The Great

Of Macedon.

Then, it assumes that longer strings are better than shorter ones, and forms candidate substitution rules, as follows:

III of Macedon (*Γ΄ὁ Μακεδών*, *III Македонски*)

III The Great (*Γ΄ὁ Μέγας*, *III Велики*)

The above described example is based on content units and their descriptive metadata, for which is in process the development of a virtual museum of ancient history and civilization.

Similar cases can be often observed when documenting historic facts and events in the middle ages. There are situations with substantial number of versions for the name and title of historic figures like the medieval Bulgarian ruler Asparuh, named as Asparuh/Asparukh (Bulg. *Аспарух*), Isperih (Bulg. *Исперих*), Esperih (Bulg. *Есперих*), Ispor (Bulg. *Испор*), Asparhruk, Batiy, etc. with several versions of title "han", "khan", "knyaz", and "tsar".

Then are performed analysis for each multi-valued field in any column that does not have the "keep all" designator. The net result is a collection of possible rules and a count of the number of times each occurs.

Finally, the rules are sorted into frequency order and presents the first one to a human along with a sample of the clusters to which it applies. The human is asked to respond "yes", "no" or "maybe". The rule is automatically applied or discarded in the first two cases. In the third case, it asks a human to start tagging values as "correct" or "not correct". Based on this training data, is formed a decision tree model for the collection of clusters. This process of examining the most frequent possible rules continues until a human decides that the point of diminishing returns has occurred [21].

7. Conclusion. In the last decade the cultural heritage has been radically transformed, with broad acceptance of digital technologies facilitating the exhibition and sharing of its richness, the affirmation of national identity and the development of culture in society. The digital technologies provided the means for active inclusion, engaging and participation of broad user community in the processes of access, exploration and study of this wealth. The investigation of these essential transformations is of crucial value and covers studies of new means and methods for improved attainment, understanding, analysis and interpretation of cultural content, context-dependent use and sharing of the huge databases of cultural and historical objects, new forms of interaction with digital cultural and historical content through the modern digital data transfer channels, modern forms of communication, civic engagement, etc., towards national identity and development of culture in the society.

This paper is focused on the modern digital content management systems that have to cover a complex set of functionalities in order to operate as complete solutions. The management of digital objects in a virtual museum for cultural heritage requires a well-designed architecture that embeds services for content presentation, content management, ad-

ministration of user data and analysis. This set of services is interdependent and demands a high level of data integrity, which is hard to achieve when digital objects originate from disparate data sources with specific and non-standardized data formats and elements. In such a scenario, intelligent data curation that leverages machine learning to clean and unify data, is a sound approach that increases efficiency and eliminates errors of duplicate and inaccurate data. In this process are employed logistic regression, decision trees and ad-hoc models that use training data to fit a model, often assisted by human feedback and active learning. These models undergo continuous evolution and get improved by additional techniques with each new use case.

Further investigations in the above-discussed domain point to a wide variety of directions:

– Creation of workable methods and tools, aiming to increase and generalize the visitors' experience in the virtual museum. Moreover, creative user experiences will support the effective on-line learning through virtual museums.

– Design and development of contextual techniques for personalizing user's work in these platforms.

– Design and development of multimodal interfaces and intelligent visualisation of complex and heterogeneous media objects relying on enhanced usability (incl. user-centric visualisation and analytics, real-time adaptable and interactive visualisation, real-time and collaborative 3D visualisation, dynamic clustering of information, etc.), etc.

The field has great potential for innovations, especially in present world of active imposition of new e-devices. The focus will be also in the research and exploitation of new or emerging technologies (e.g. 3D, augmented and virtual reality, visual computing, smart world, media convergence, social media, etc.) for the development of innovative products, tools, applications, and services for creative digital content production, usage and management. The aim is to transform and customize the valuable parts of mankind's cultural and historical ancestry into digital assets, whose integration and reuse through research-lead methods has high commercial and non-commercial potential for learning and cultural institutions, tourism, creative and media industries.

References

1. Pavlov R., Paneva-Marinova D., Goynov M., Pavlova-Draganova L. Services for content creation and presentation in an iconographical digital library. *International Journal "Serdica Journal of Computing"*. 2010. vol. 4. pp. 279–292.
2. Codd E., Codd S., Salley C. Providing OLAP to user-analysts. 1993. Available at: http://www.minet.uni-jena.de/dbis/lehre/ss2005/sem_dwh/lit/Cod93.pdf (accessed: 10.03.2019.).

3. Paneva-Marinova D., Goynov M., Luchev D. Multimedia digital library: Constructive block in ecosystems for digital cultural assets. Basic functionality and services. LAP LAMBERT Academic Publishing. 2017. 117 p.
4. Luchev D., Paneva-Marinova D., Pavlova-Draganova L., Pavlov R. New digital fashion world. 14th International Conference on Computer Systems and Technologies (CompSysTech'13). 2013. vol. 767. pp. 270–275.
5. Rangochev K., Goinov M., Dimitrova M., Hristova-Shomova I.. Enciclopaedia Slavica Sanctorum: activity, users, statistics. Digital Preservation and Presentation of Cultural and Scientific Heritage. 2013. vol. 3. pp. 81–90.
6. Rangochev K., Dimitrova M.. [Two models for presenting the Balkan folklore heritage in digital libraries]. *Dobre doshli v Kiberiya: zapiski ot digitalniya teren – Welcome to Cyberbia: notes from the digital terrain*. 2014. pp. 397–411. (In Bulg.).
7. Bogdanova G., Todorov T.Y., Noev N. Using graph databases to represent knowledge base in the field of cultural heritage. Digital Preservation and Presentation of Cultural and Scientific Heritage. 2016. vol. 6. pp. 199–206.
8. Pavlova-Draganova L., Paneva-Marinova D., Pavlov R., Goynov G. On the wider accessibility of the valuable phenomena of Orthodox iconography through digital library. Proceedings of the 3rd International Conference dedicated on Digital Heritage (EuroMed 2010). 2010. pp. 173–178.
9. Bogdanova G., Todorov T.Y., Kancheva S. Virtual museum of Russian bells in Bulgaria. Digital Preservation and Presentation of Cultural and Scientific Heritage. 2017. vol. 7. pp. 215–222.
10. Furht B., Escalante A. Handbook of data intensive computing. Springer-Verlag. 2011. 793 p.
11. Borgman C. Big data, little data, no data: scholarship in the networked world. MIT Press. 2015. 416 p.
12. Splendiani A. AI for data curation. Yes, can we? 2017. Available at: https://www.slideshare.net/sergentpepper/artificial-intelligence-in-data-curation (accessed: 10.03.2019.).
13. Ilyas I. Data unification at scale: data tamer. Making Databases Work. Association for Computing Machinery and Morgan & Claypool. 2018. pp. 269–277.
14. Tamr. Agile data mastering raising expectations for master data management (MDM). 2019. Available at: http://www.tamr.com: http://www.tamr.com/wp-content/uploads/2019/01/Tamr_WP_Agile-Data-Mastering-_01-14-19.pdf (accessed: 10.02.2019.).
15. Miller R. Big data curation. Proceedings of the 20th International Conference on Management of Data (COMAD). Computer Society of India. 2014. pp. 4.
16. Blank G. Studyguide for the sage handbook of Internet and online research methods. 2012. Cram101. 80 p.
17. Walbank F. Alexander the Great: King of Macedonia. 2019. Available at: https://www.britannica.com/biography/Alexander-the-Great (accessed: 01.02.2019.).
18. O' Brien J. Alexander the Great: The Invisible Enemy: A Biography. Routledge. 2005. 360 p.
19. Bosworth A.B., Baynham E.J. Alexander the Great in fact and fiction. Oxford University Press. 2000. 384 p.
20. Green P. Alexander of Macedon, 356—323 B.C.: A historical biography. University of California Press. 1991. 617 p.
21. Stonebraker M. Machine learning for data unification practical applications in Tamr's software platform. 2017. Available at: https://www.tamr.com/wp-

content/uploads/2017/07/Machine_Learning_For_Data_Unification_072117_2.pdf (accessed: 10.03.2019.).
22. Georgieva-Tsaneva G., Subev N. Technologies, Standarts and Approaches to Ensure Web Accessibility for Visually Impaired People. Digital Preservation and Presentation of Cultural and Scientific Heritage. 2018. vol. 8. pp. 143–150.
23. Bogdanova G., Noev N. Digitization and preservation of digital resources and their accessibility for blind people. *Cyber-physical systems for social applications*. 2019. pp. 184–206.
24. Karpov A., Ronzhin A. A Universal Assistive Technology with Multimodal Input and Multimedia Output Interfaces. Proceedings of the 8th International Conference on Universal Access in Human-Computer Interaction (UAHCI 2014). 2014. pp. 369–378.
25. Yoshinov R., Kotseva, M, Pavlova D. Specifications for Centralized DataCenter serving the educational cloud for Bulgaria. Proceedings of XII International Conference on Electronics, Telecommunications, Automatics & Informatics (ETAI). 2015. pp. 1–6.
26. Yoshinov R. Bringing up-to-date the principles of the I.DB.I. Artery. Proceedings of IX International Conference on Electronics, Telecommunications, Automatics & Informatics (ETAI). 2009. pp. I3–1.
27. Yoshinov R., Iliev O. "Controlled self-study" in thematic educational community environment. The 47th Spring Conference of the Union of Bulgarian Mathematicians. 2018. pp. 200–213.
28. Yoshinov R., Iliev O. Content reuse – a major problem with modern content storage systems. Eleventh National Conference with International Participation "Education and Research in the Information Society". 2018.
29. Yoshinov R., Kotseva M. The steps for elaboration of the "Rosetta stone" demonstrator. Proceedings of International Conference Inspiring Science Education. 2016. pp. 91–96.
30. Yoshinov R., Arapi P., Kotseva M., Christodoulakis S. Supporting Personalized Learning Experiences on top of Multimedia Digital Libraries. *International journal of education and information technologies*. 2016. vol. 10. pp. 152–158.
31. Trifonov R., Yoshinov R., Jekov B., Pavlova G. Methodology for Assessment of Open Data. *International Journal of Computers*. 2017. vol. 2. pp. 28–37.
32. Yoshinov R.D., Iliev O.P. (2018) The Structural Way for Binding a Learning Material with Personal Preferences of Learners. *Trudy SPIIRAN – SPIIRAS Proceedings*. 2018. vol. 5(60). pp. 189–215.

Paneva-Marinova Desislava Ivanova — Ph.D., associate professor, head of the Mathematical Linguistics Department, Institute of Mathematics and Informatics at the Bulgarian Academy of Sciences (BAS). Research interests: computer science, technologies for knowledge presentation and processing, data management and processing, data analytics, intelligent data curation, Semantic web, digital content management systems, web services. The number of publications — 112. dessi@cc.bas.bg; 8, Akad. G. Bonchev Str., 1113, Sofia, Republic of Bulgaria; office phone: +359888894814.

Stoikov Jordan Stoikov — Ph.D. student at the Mathematical Linguistics Department, Institute of Mathematics and Informatics at the Bulgarian Academy of Sciences (BAS). Research interests: computer science, business management systems and services for intelligent data curation, data integrity, knowledge retrieval, data validation. The number of publications — 2. jstoikov@shieldui.com; 8, Akad. G. Bonchev Str., 1113, Sofia, Republic of Bulgaria; office phone: +35929792874.

Pavlova Lilia Radoslavova — Ph.D., assistant professor at the Laboratory of Telematics, Bulgarian Academy of Sciences (BAS). Research interests: computer science, technologies for knowledge presentation and processing, Semantic web, e-learning. The number of publications — 34. pavlova.lilia@gmail.com; 8, Akad. G. Bonchev Str., 1113, Sofia, Republic of Bulgaria; office phone: +35929793831.

Luchev Detelin Mihailov — Ph.D., associate professor, Mathematical Linguistics Department, Institute of Mathematics and Informatics at the Bulgarian Academy of Sciences (BAS). Research interests: computer science, digital content management systems, technologies for knowledge presentation and processing, data management and processing, web services, mobile applications, history, ethnology. The number of publications — 75. dml@math.bas.bg; 8, Akad. G. Bonchev Str., 1113, Sofia, Republic of Bulgaria; office phone: +359885978788

Acknowledgements. This research is partially supported by the Bulgarian Ministry of Education and Science under the National Research Programme "Cultural heritage, national memory and development of society" approved by DCM №577/17.08.2018 and by the Bulgarian Scientific Fund under the research project № DN02/06/15.12.2016 "Concepts and Models for Innovation Ecosystems of Digital Cultural Assets".

УДК 006.72 **DOI** 10.15622/sp.18.2.444-470

Д.И. Панева-Маринова, Й.С. Стойков, Л.Р. Павлова, Д.М. Лучев

АРХИТЕКТУРА СИСТЕМЫ И ИНТЕЛЛЕКТУАЛЬНАЯ ОБРАБОТКА ДАННЫХ ВИРТУАЛЬНОГО МУЗЕЯ ДРЕВНЕЙ ИСТОРИИ

Йошинов Р.Д., Панева-Маринова Д.И., Стойков Й.С., Павлова Л.Р., Лучев Д.М. **Архитектура системы и интеллектуальная обработка данных виртуального музея древней истории.**

Аннотация. Сохранение культурного и исторического наследия разных народов мира и их тщательное изложение – это долгосрочное обязательство ученых и исследователей, работающих во многих областях. На протяжении веков каждое поколение стремится вести учет своего труда, чтобы его могли пересмотреть и изучить следующие поколения. За последние пару лет были разработаны новые информационные и мультимедийные технологии, которые представили новые методы сохранения, обслуживания и распространения огромного количества собранного материала. Эта статья призвана представить виртуальный музей, передовую систему, управляющую разнообразными коллекциями цифровых объектов, которые по-разному организованы с помощью сложной специализированной функциональности. Управление цифровым содержанием требует хорошо продуманной архитектуры, которая включает в себя сервисы для представления, управления и администрирования содержания. Все элементы архитектуры системы взаимосвязаны, поэтому точность каждого элемента имеет большое значение. Эти системы страдают от недостатка инструментов для интеллектуального курирования данных с возможностью проверки данных из разных источников и повышения ценности данных. В этой статье предлагается решение для интеллектуального курирования данных, которое может быть реализовано в виртуальном музее, чтобы предоставить возможность надлежащим образом наблюдать ценные исторические образцы. Решение сфокусировано на процессах валидации и верификации, чтобы предотвратить дублирование записей цифровых объектов, чтобы гарантировать целость данных и более точный поиск знаний.

Ключевые слова: виртуальный музей, архитектура системы, функциональность, целостность данных, поиск знаний, проверка данных, дедупликация записей, культурное наследие.

Панева-Маринова Десислава Иванова — Ph.D., доцент, заведующая кафедрой математической лингвистики Института математики и информатики Болгарской академии наук (БАН). Научные интересы: информатика, технологии представления и обработки знаний, управление и обработка данных, аналитика данных, интеллектуальное курирование данных, семантическая сеть, системы управления цифровым контентом, веб-сервисы.Число научных публикаций — 112. dessi@cc.bas.bg; ул. Акад. Георги Бончев, бл. 8, 1113, София, Республика Болгария; р.т.: +359888894814.

Стойков Йордан Стойков — аспирант кафедрой математической лингвистики Института математики и информатики Болгарской академии наук (БАН). Сфера научных интересов: компьютерные науки, системы и услуги управления бизнесом для интеллектуального хранения данных, целостности данных, поиска знаний, проверки данных.Число научных публикаций — 2. jstoikov@shieldui.com; ул. Акад. Георги Бончев, бл. 8, 1113, София, Республика Болгария; р.т.: +35929792874.

Павлова Лилия Радославова — Ph.D., научный сотрудник Лаборатории телематики Болгарской академии наук (БАН). Сфера научных интересов: информатика, технологии представления и обработки знаний, семантическая паутина, электронное обучение.Число научных публикаций — 34. pavlova.lilia@gmail.com; ул. Акад. Георги Бончев, бл. 8, 1113, София, Республика Болгария; р.т.: +35929793831.

Лучев Детелин Михайлов — Ph.D., доцент кафедры математической лингвистики Института математики и информатики Болгарской академии наук (БАН). Сфера научных интересов: информатика, системы управления цифровым контентом, технологии представления и обработки знаний, управление и обработка данных, веб-сервисы, мобильные приложения, история, этнология. Число научных публикаций — 75. dml@math.bas.bg; ул. Акад. Георги Бончев, бл. 8, 1113, София, Республика Болгария; р.т.: +359885978788

Поддержка исследований. Исследование выполнено при частичной финансовой поддержке Министерством образования и науки Болгарии в рамках Национальной исследовательской программы «Культурное наследие, национальная память и развитие общества», DCM № 577/17.08.2018, и Болгарским научным фондом в рамках исследовательского проекта № DN02/06/15.12.2016 «Концепции и модели инновационных экосистем цифровых культурных ценностей».

Литература

1. *Pavlov R., Paneva-Marinova D., Goynov M., Pavlova-Draganova L.* Services for content creation and presentation in an iconographical digital library // International Journal "Serdica Journal of Computing". 2010. vol. 4. pp. 279–292.
2. *Codd E., Codd S., Salley C.* Providing OLAP to user-analysts. 1993. URL: http://www.minet.uni-jena.de/dbis/lehre/ss2005/sem_dwh/lit/Cod93.pdf (дата обращения: 10.03.2019.).
3. *Paneva-Marinova D., Goynov M., Luchev D.* Multimedia digital library: Constructive block in ecosystems for digital cultural assets. Basic functionality and services // LAP LAMBERT Academic Publishing. 2017. 117 p.
4. *Luchev D., Paneva-Marinova D., Pavlova-Draganova L., Pavlov R.* New digital fashion world // 14th International Conference on Computer Systems and Technologies (CompSysTech'13). 2013. vol. 767. pp. 270–275.
5. *Rangochev K., Goinov M., Dimitrova M., Hristova-Shomova I.*. Enciclopaedia Slavica Sanctorum: activity, users, statistics // Digital Preservation and Presentation of Cultural and Scientific Heritage. 2013. vol. 3. pp. 81–90.
6. *Рангочев К., Димитрова М.*. Два модела за представяне на българското фолклорно наследство в цифрови библиотеки // Добре дошли в Киберия: записки от дигиталния терен. 2014. С. 397–411.
7. *Bogdanova G., Todorov T. Y., Noev N.* Using graph databases to represent knowledge base in the field of cultural heritage // Digital Preservation and Presentation of Cultural and Scientific Heritage. 2016. vol. 6. pp. 199–206.
8. *Pavlova-Draganova L., Paneva-Marinova D., Pavlov R., Goynov G.* On the wider accessibility of the valuable phenomena of Orthodox iconography through digital library // Proceedings of the 3rd International Conference dedicated on Digital Heritage (EuroMed 2010). 2010. pp. 173–178.
9. *Bogdanova G., Todorov T.Y., Kancheva S.* Virtual museum of Russian bells in Bulgaria // Digital Preservation and Presentation of Cultural and Scientific Heritage. 2017. vol. 7. pp. 215–222.

10. *Furht B., Escalante A.* Handbook of data intensive computing (1 ed.) // Springer-Verlag. 2011. 793 p.
11. *Borgman C.* Big data, little data, no data: scholarship in the networked world // MIT Press. 2015. 416 p.
12. *Splendiani A.* AI for data curation. Yes, can we? 2017. URL: https://www.slideshare.net/sergentpepper/artificial-intelligence-in-data-curation (дата обращения: 10.03.2019.).
13. *Ilyas I.* Data unification at scale: data tamer. Making Databases Work // Association for Computing Machinery and Morgan & Claypool. 2018. pp. 269–277.
14. Tamr. Agile data mastering raising expectations for master data management (MDM). 2019. URL: http://www.tamr.com: http://www.tamr.com/wp-content/uploads/2019/01/Tamr_WP_Agile-Data-Mastering-_01-14-19.pdf (дата обращения: 10.02.2019.).
15. *Miller R.* Big data curation // Proceedings of the 20th International Conference on Management of Data (COMAD). Computer Society of India. 2014. pp. 4.
16. *Blank G.* Studyguide for the sage handbook of Internet and online research methods. 2012. Cram101. 80 p.
17. *Walbank F.* Alexander the Great: King of Macedonia. 2019. URL: https://www.britannica.com/biography/Alexander-the-Great (дата обращения: 01.02.2019.).
18. *O' Brien J.* Alexander the Great: The Invisible Enemy: A Biography // Routledge. 2005. 360 p.
19. *Bosworth A.B., Baynham E.J.* Alexander the Great in fact and fiction // Oxford University Press. 2000. 384 p.
20. *Green P.* Alexander of Macedon, 356–323 B.C.: A historical biography // University of California Press. 1991. 617 p.
21. *Stonebraker M.* Machine learning for data unification practical applications in Tamr's software platform. 2017. URL: https://www.tamr.com/wp-content/uploads/2017/07/Machine_Learning_For_Data_Unification_072117_2.pdf (дата обращения: 10.03.2019.).
22. *Georgieva-Tsaneva G., Subev N.* Technologies, Standarts and Approaches to Ensure Web Accessibility for Visually Impaired People // Digital Preservation and Presentation of Cultural and Scientific Heritage. 2018. vol. 8. pp. 143–150.
23. *Bogdanova G., Noev N.* Digitization and preservation of digital resources and their accessibility for blind people // Cyber-physical systems for social applications. 2019. pp. 184–206.
24. *Karpov A., Ronzhin A.* A Universal Assistive Technology with Multimodal Input and Multimedia Output Interfaces // Proceedings of the 8th International Conference on Universal Access in Human-Computer Interaction (UAHCI 2014). 2014. pp. 369–378.
25. *Yoshinov R., Kotseva, M, Pavlova D.* Specifications for Centralized DataCenter serving the educational cloud for Bulgaria // Proceedings of XII International Conference on Electronics, Telecommunications, Automatics & Informatics (ETAI). 2015. pp. 1–6.
26. *Yoshinov R.* Bringing up-to-date the principles of the I.DB.I. Artery // Proceedings of IX International Conference on Electronics, Telecommunications, Automatics & Informatics (ETAI). 2009. pp. I3–1.
27. *Yoshinov R., Iliev O.* "Controlled self-study" in thematic educational community environment // The 47th Spring Conference of the Union of Bulgarian Mathematicians. 2018. pp. 200–213.
28. *Yoshinov R., Iliev O.* Content reuse – a major problem with modern content storage systems // Eleventh National Conference with International Participation "Education and Research in the Information Society". 2018.

29. *Yoshinov R., Kotseva M.* The steps for elaboration of the “Rosetta stone” demonstrator // Proceedings of International Conference Inspiring Science Education. 2016. pp. 91–96.
30. *Yoshinov R., Arapi P., Kotseva M., Christodoulakis S.* Supporting Personalized Learning Experiences on top of Multimedia Digital Libraries // International journal of education and information technologies. 2016. vol. 10. pp. 152–158.
31. *Trifonov R., Yoshinov R., Jekov B., Pavlova G.* Methodology for Assessment of Open Data // International Journal of Computers. 2017. vol. 2. pp. 28–37.
32. *Yoshinov R.D., Iliev O.P.* (2018) The Structural Way for Binding a Learning Material with Personal Preferences of Learners // Труды СПИИРАН. 2018. Вып. 5(60). pp. 189–215.

УДК 006.72 **DOI** 10.15622/sp.18.2.471-503

А.С. Гуменюк, А.А. Скиба, Н.Н. Поздниченко, С.Н. Шпынов

О МЕРАХ СХОДСТВА РАСПОЛОЖЕНИЯ КОМПОНЕНТОВ В МАССИВАХ ЕСТЕСТВЕННО УПОРЯДОЧЕННЫХ ДАННЫХ

Гуменюк А.С., Скиба А.А., Поздниченко Н.Н., Шпынов С.Н. **О мерах сходства расположения компонентов в массивах естественно упорядоченных данных.**

Аннотация. В настоящее время в публикациях специалистов по анализу массивов естественно упорядоченных данных различной природы (в том числе символьных последовательностей) не имеют широкого распространения математические средства, адекватно учитывающие расположение компонентов. Поэтому затруднены или невозможны измерение и сравнение порядка следования сообщений, выделенных в длинных информационных цепях. Основные подходы при сравнении символьных последовательностей используют вероятностные модели и статистический инструментарий, попарное и множественное выравнивание, позволяющее определить степень сходства цепей с помощью мер редакционного расстояния. Отмеченные подходы почти не уделяют внимания исследованию и обнаружению закономерностей конкретного расположения всех знаков, слов, компонентов массивов данных, составляющих отдельную целостную последовательность. Объектом исследования в наших работах является специальным образом организованный числовой кортеж — расположение компонентов (строй) в символьных или числовых последовательностях. При этом в качестве основы для количественного отображения строя цепи используются интервалы между ближайшими одинаковыми ее компонентами. Перемножение всех интервалов или суммирование их логарифмов позволяет получить числа, которые однозначно отображают расположение компонентов в конкретной последовательности. Эти числа, в свою очередь, позволяют получить целый набор нормированных характеристик строя, среди которых средний геометрический интервал и его логарифм. В данной работе представлен подход для количественного сравнения построений массивов естественно упорядоченных данных (информационных цепей) произвольной природы. Предложены меры сходства-расхождения и процедура сравнения строя цепей, основанные на выделении списка совпадающих и сходных по характеристикам строя подпоследовательностей. При этом для быстрого выделения списка совпадающих компонентов используются ранговые распределения. В работе представлен инструментарий для сравнения построений информационных цепей и продемонстрированы некоторые его возможности при исследовании строя нуклеотидных последовательностей.

Ключевые слова: знаковая последовательность, информационная цепь, строй цепи, глубина строя, средняя удаленность, нуклеотидная последовательность, меры сходства-расхождения, матрица сходства, alignment-free genome comparison, межнуклеотидное расстояние.

1. Введение. Уже более 100 лет используются формальные средства для анализа знаковых последовательностей разной природы. В начале прошлого века при зарождении математической лингвистики появились работы по статистическим исследованиям текстов на естественных языках [1]. В 50-60 годы на фоне широкого использования цифровых вычислительных машин отдельные исследователи применили формальный анализ к музыкальным произведениям, и одновременно

стали использоваться математические модели и средства анализа так называемых «генетических текстов», то есть нуклеотидных, аминокислотных последовательностей и тому подобных. Кроме того, начались интенсивные исследования больших массивов естественно упорядоченных данных измерений (данные мониторинга). В таких массивах обычно требуется учитывать оригинальное расположение выделенных компонентов. В процессе анализа подобных последовательностей исследователи пытаются выявить структуру массива данных. При этом непосредственное исследование структуры никакими средствами не осуществляется, так как не определено и само понятие структуры для цепей данных. Обычно исследователи пытаются опереться на определенную природу компонентов таких цепей (слова, нуклеотиды, триплеты, кодоны, аминокислоты, амплитуды сигналов, высоты звучания нот и тому подобное). Это направление исследований структуры породило большое число «тонких формальных техник», применимых к цепям выделенной природы [2-10].

Общие подходы к исследованию структуры знаковых цепей без опоры на материальную природу компонентов представлены двумя основными направлениями: суждения о структуре цепи на основе статистического распределения ее элементного состава и косвенные суждения о структуре цепи с помощью оценки и исследования локального порядка следования компонентов в цепях.

В первом подходе суждение о структуре полной цепи осуществляется на основе статистического распределения (состава) ее компонентов (двоек, троек и в общем случае n-ок) [11]. Так как компоненты в рамках вероятностной модели цепи представляют собой случайные события (но не величины), в лучшем случае возможно построение статистических распределений, частотно-ранговых распределений или H-статистик. В предельном случае, как это показано в докторской диссертации М.Г. Садовского [12], длина окна для короткого кортежа (n-ки) может быть такой достаточной величины, что при считывании конкретной нуклеотидной цепи L-граммами со сдвигом на один элемент мы получим некоторое конечное множество (алфавит или словарь) L-грамм (n-ок), на основе которого возможно однозначно восстановить расположение всех компонентов исходной цепи. Однако такое численное описание структуры достигается путем введения многократной избыточности за счет $(n-1)$-кратного тиражирования цепи. Суждения же на основе обычного статистического распределения компонентов (или блоков компонентов) данной цепи, полученного экспериментально, не претендуют на возможность восстановления

исходной последовательности. Косвенно такие распределения все же являются количественным описанием взаимного расположения элементов, так как исследователю по умолчанию известно, что он взял не случайную выборку данных, а конкретный текст, нуклеотидную последовательность и тому подобное [13]. Это «проклятие априорного неосознаваемого знания» об очевидной упорядоченности цепи широко распространено в математической лингвистике, биоинформатике (математической биологии) и других аналогичных областях науки.

Другое направление исследований и анализа структуры массивов данных в основном использует мощные известные вероятностно-статистические средства и модели — марковские цепи, потоки заявок и теорию очередей, взвешенные графы, с помощью которых удается хоть и громоздко описать локальную структуру знаковых цепей, но не оригинальное расположение компонентов всей цепи [8, 14].

Особо выделим работу выдающегося математика современности В.И. Арнольда, посвященную разработке теории сложности конечных бинарных последовательностей [15].

Кроме того, косвенный анализ структуры осуществляется путем сравнения пар цепей, одна из которых может быть эталонной. Для этого используются разные меры сходства (различия), среди которых широко используется несимметричная статистическая мера Кульбака — Лейблера, а также мера Левенштейна в форме «редакционного расстояния», которую, по существу, можно считать обобщенной метрикой Хемминга, представляющей расстояние между словами одинаковой длины [16-19].

В настоящее время в противоположность вычислительно сложным методам, основанным на выравнивании последовательностей, в биоинформатике выделяют группы подходов (в основном статистических), называемых alignment-free sequence analysis, которые позволят сравнивать и описывать нуклеотидные последовательности без применения выравнивания и отличаются высоким быстродействием [20-26].

Среди них выделим методы символической динамики, основанные на подходах систем «динамического хаоса», которые обычно используются для визуализации символьных последовательностей в виде траекторий в пространстве некоторой размерности (обычно 2 или 3). В некоторых работах из таких траекторий далее получают интегральные числовые характеристики [27, 28].

В рамках выделенных направлений, по исследованию структуры знаковых последовательностей широко применяется получение вероятностно-статистических и энтропийно-информационных характеристик и оценок. На протяжении десятилетий

глубокие разработки на основе этих средств и моделей, а также с учетом природы объектов со значительным практическим выходом выполняются в Институте математики им. Соболева сибирского отделения РАН в лаборатории под руководством В.Д. Гусева [2-6].

Приведем утверждение из [12], которое определяет методологический подход для анализа и исследования структуры символьных последовательностей: «Как хранение, так и реализация какой-либо информации напрямую обусловлены тем обстоятельством, что в ходе этих процессов актуальную роль играют символьные последовательности. При этом хорошо известно, что в природе фактически нет процессов, связанных с переработкой либо реализацией той или иной информации, которые бы вовлекали всю такого рода символьную последовательность целиком: чтение и обработка файлов вычислительными машинами происходит малыми порциями (байтами) и последовательно, чтение и переработка письменной информации человеком происходит малыми порциями (словами, либо абзацами) и последовательно, чтение и переработка наследственной информации в биологических системах происходит малыми порциями (кодонами) и последовательно. Это простое обстоятельство, тем не менее, имеет важные последствия. Оно требует перехода от рассмотрения всей символьной последовательности в целом к рассмотрению набора ее фрагментов.». Данное утверждение является своего рода постулатом редукционизма или познавательной установкой по Ю. Шрейдеру для данного методологического подхода [29]. По мнению авторов такое допущение существенно ограничивает возможности теории информации, прикладной информатики, в частности математической лингвистики, информатики нуклеотидных последовательностей и средств анализа массивов естественно упорядоченных данных любой природы.

Для исследования знаковых цепей, текстов разной природы и массивов данных измерений разработаны и используются, как отмечалось выше, большое число специальных подходов, процедур и моделей, которые можно дополнить математическим, спектральным, статистическим, корреляционным, фрактальным и другими анализами. Однако почти не уделяется внимания исследованию и обнаружению закономерностей *конкретного расположения всех знаков, слов, компонентов массивов данных, составляющих отдельную целостную последовательность.* Можно констатировать, что до настоящего времени массивы естественно упорядоченных (текстовых) данных обычно рассматриваются как множества (но не кортежи), в которых не принято учитывать и численно представлять расположение их элементов. Можно высказаться

более категорично — до настоящего времени массивы естественно упорядоченных данных обрабатывались, анализировались и исследовались как «вещество», а не как кортежи, информационные цепи или связные «тексты».

По мнению авторов такое положение в некоторой степени объясняется следующими причинами:

1. Отсутствие английского перевода фундаментальной работы М. Мазура «Качественная теория информации», в которой определяются и особо рассматриваются информационные цепи (массивы данных) в отличие от кодовых цепей [30].

2. Отказ от системного подхода, который учитывал бы тексты и массивы естественно упорядоченных данных как абстрактные объекты, каждый из которых представляет собой единое целое.

3. Очевидность определенного расположения компонентов в естественно упорядоченной последовательности, которое нельзя искажать при обработке данных; очевидное не побуждает к формализации.

4. Отсутствие формализма для особого абстрактного объекта, представляющего расположение компонентов и называемого нами строем или построением цепи [31, 32].

Следует отметить, что разные по природе последовательности событий с одинаковыми статистическими распределениями (в дальнейшем — с равномощными составами) могут иметь одинаковый строй. С другой стороны, очевидно, что множество, которое содержит повторяющиеся элементы (мультимножество), может быть основой для построения различных комбинаций типа «перестановки с повторениями». При этом большинство из них будут иметь разное расположение компонентов.

В работах [9, 10, 33-36] использовались интервалы между ближайшими одинаковыми компонентами (межнуклеотидное расстояние) в качестве основы для исследования и сравнения нуклеотидных последовательностей. Однако в них рассматривались только статистические (ранговые) распределения интервалов [37, 38], а также преобразования Фурье и вейвлет-преобразования [39] нуклеотидных последовательностей, что не позволяло описать отдельную последовательность одним числом. Для таких (обычно гиперболических) распределений не определялись статические характеристики (мат ожидание, СКО и тому подобное).

Наконец, особо отметим открытие Ю. Орловым феномена «целостно-завершенного текста», который обнаруживается только при хорошем совпадении статистического рангового распределения

слов данного текста с законом Ципфа — Мандельброта [40, 41]. Ю. Шрейдер формально доказал, что существование такого «идеального» распределения слов для отдельного текста вытекает из фундаментального принципа «минимума симметрии» для целостной системы [29]. При этом для исследования естественных систем им предложена комбинированная методология поочередного использования системного подхода и редукционизма. В России это направление развивается для исследования техноценозов под руководством Б. Кудрина [42]. Применимость закона Ципфа — Мандельброта при исследовании статистических распределений генетических текстов рассматривалось М. Гельфандом [43]. К сожалению, отмеченные разработки советских и российских ученых до настоящего времени не попали в поле зрения англоязычного научного сообщества.

В наших ранних работах [31] предлагается подход, который предназначен для «формального описания и анализа строя» (ФОАС) отдельного текста любой природы (знаковой цепи), в том числе представляющего нуклеотидную последовательность, или массив данных измерений.

2. Формальное описание строя. *Строй цепи* сообщений (событий, знаков и тому подобных) — это кортеж (упорядоченное множество), в котором каждому компоненту цепи поставлено в соответствие натуральное число, причем идентичные по выбранному признаку компоненты отображены одним и тем же числом. Первый компонент кортежа — единица, каждый следующий компонент цепи, отличный от всех предыдущих, обозначается натуральным числом, которое на единицу больше максимального из расположенных ранее в кортеже.

В соответствии с определением для формирования строя необходимо учитывать следующие ограничения:

1. *Алфавит строя* — это множество всех натуральных чисел из диапазона от 1 до m $\{1, 2, 3, 4, 5, \ldots, m\}$.

2. Мощность алфавита m всегда не больше длины строя $m \leqslant n$ (предельный случай, когда длина строя равна размеру алфавита $(m = n)$ и все элементы (числа) встречаются в строе один раз).

3. Первые вхождения элементов алфавита располагаются на позиции строя по возрастанию, начиная с единицы в первой позиции, возможно с пропусками некоторых мест:

`⟨1 2-3-4--5---6 7⟩`.

4. Места на позиции строя, не занятые первыми вхождениями элементов алфавита, заполняются натуральными числами, по значению не превышающими максимального среди всех лежащих слева чисел:

⟨1 2 1 3 2 3 4 4 4 1 5 3 4 5 1 1 1 6 7⟩.

Мощность алфавита строя — это количество различных компонентов в цепи.

Примеры разных последовательностей (кортежей) символов с одинаковым строем приведены на рисунках 1 и 2.

Для сравнения по строю нескольких кортежей реальных сообщений необходимо правильно выполнить однозначное *прямое преобразование* для каждого из них, а затем сравнить полученные строи.

Для кодирования разных знаков при прямом преобразовании цепи сообщений в строй цепи кроме натуральных чисел возможно использовать любой (упорядоченный) алфавит символов достаточно большой мощности. Соответствие между исходной и закодированной таким образом последовательностями называется «совпадением с точностью до переименования». Однако такой алфавит необходимо выбрать или специально построить и самое трудное — сделать его общепринятым. Кроме того, все реальные алфавиты и словари неявно упорядочены натуральными числами для удобства запоминания и использования.

В теоретико-множественном представлении вектором называется кортеж, компонентами которого являются числа. В соответствии с таким определением вектора, назовем специфически сформированный (организованный) кортеж «*вектором строя*».

Таким образом, строй цепи и вектор строя — это синонимы одного и того же абстрактного объекта. Однако на практике следует различать «вектор строя» данной цепи или некоторого их множества и «вектор строя» как элемент множества разных векторов строя.

Заметим, что при несоблюдении ограничений на порядок расположения натуральных чисел мы получим кортеж, точнее вектор, который не представляет собой строй. На рисунке 3 представлен такой вектор.

Рассмотрим отличный от представленного на рисунке 1 строй цепи. Очевидно неоднозначное преобразование данного строя в знаковые последовательности. Для наглядности, пусть они имеют мощность состава элементов такую же, как на рисунке 1. Условимся называть преобразование строя в знаковую цепь «*обратным преобразованием строя*» (рисунок 4).

При одинаковой мощности составов знаковых цепей их частотные распределения одинаковы, то есть инвариантны относительно расположения элементов в цепях, что видно из примеров.

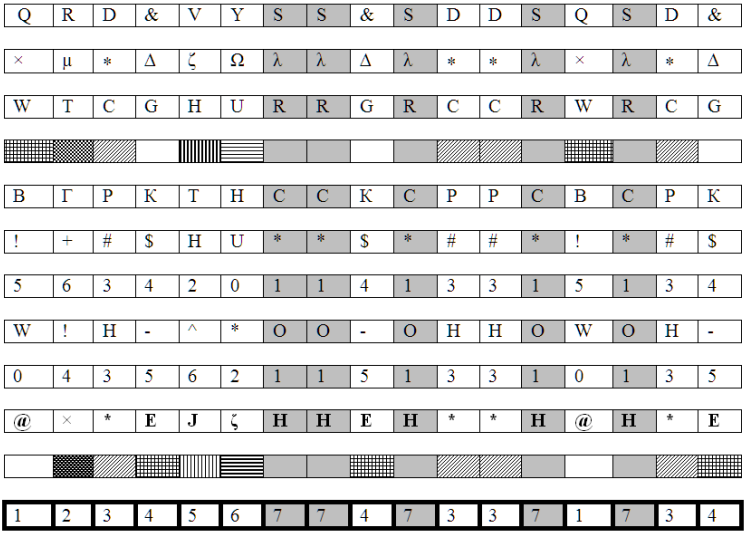


Рис. 1. Пример прямого преобразования 11 разных знаковых цепей, цифровых последовательностей и диаграмм в строй цепи

A	C	C	T	G	A	C	T	G	C	T	A	T	C	G	G	A	T	T	G	A	T	A	C
T	G	G	A	C	T	G	A	C	G	A	T	A	G	C	C	T	A	A	C	T	A	T	G
1	2	2	3	4	1	2	3	4	2	3	1	3	2	4	4	1	3	3	4	1	3	1	2

Рис. 2. Фрагмент двух комплементарных цепочек РНК бактерии Candidatus nitrosopumilus maritimus с одинаковым строем

Строй цепи — это идея или план построения некоторого множества кортежей, цепей реальных сообщений, сигналов или событий. *Строй цепи* в определенном смысле соответствует введенному Гете понятию «*архетип*». Этот термин предложил использовать Юлий Шрейдер для обозначения описания структуры таксона (класса объектов) [29]. Другими словами, *строй цепи — это ее архетип*.

Операция выявления в разных по природе информационных цепях одинаковых построений расширяет возможности междисциплинарных исследований. Однако результат такой операции ограничен описанием строя в форме обычного числового кортежа, хотя и имеющего определение «вектор строя». Рассмотрим более удобное для анализа формальное описание строя, которое позволяет получать компактные числовые характеристики (подобные используемым для описания случайных величин), полезные, в частности, при опознавании строев цепей и

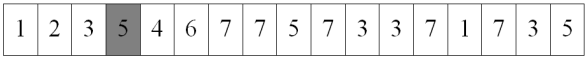


Рис. 3. Вектор натуральных чисел, который не представляет строй цепи

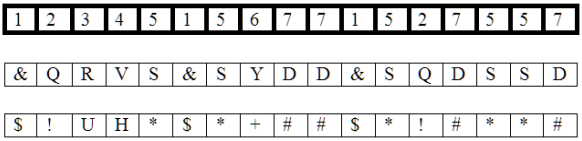


Рис. 4. Обратное преобразование данного строя в две разные знаковые последовательности

определении степени их различия. Для определения такого формализма строя отдельной цепи при обычном (естественном) способе ее чтения «поэлементно» (подряд) введем две нумерации:

– первая нумерует элементы собственного алфавита (словаря) данной знаковой последовательности по порядку их встречи;

– вторая дает сквозную нумерацию всех компонентов кортежа от начала до конца.

Разложим *полную неоднородную* (без пустых мест на позиции) символьную последовательность на m *неполных однородных кортежей*, на позициях которых заняты одинаковыми знаками только некоторые места (рисунок 5). Такое разложение цепи называют *декомпозицией*. Аналогом однородной последовательности является поток однородных событий (заявок) из теории массового обслуживания. Очевидно, что *композиция* всех однородных строев данного полного строя дает полный неоднородный строй, аналогом которого в теории очередей является неоднородный поток событий. Вообще разложение цепи может осуществляться по разным правилам. Декомпозиция строя полной неоднородной знаковой цепи на неполные однородные представлена на рисунке 5.

В приведенных на рисунке 5 примерах используется привязка к концу последовательности, то есть последний интервал считается от последнего вхождения компонента до конца последовательности. Кроме данной привязки могут также использоваться следующие варианты: к началу, к началу и к концу, циклическая, либо отсутствие привязки.

Определим «*интервал*» как расстояние от выделенного в цепи компонента до другого ближайшего, отмеченного в направлении просмотра (рисунок 5); *величина интервала* — это натуральное число, определенное как модуль разности номеров мест двух выделенных

Рис. 5. Декомпозиция строя неоднородной знаковой цепи на неполные однородные цепи и матрица их интервалов

компонентов на позиции кортежа. В дальнейшем для краткости будем называть эти понятия одинаково — интервал.

Назовем направление чтения текста или знаковой цепи «подряд слева направо» «обычным способом считывания». Пусть первое считывание текста осуществляется отличным от обычного способом с самого начала до конца таким образом, что выбираются только элементы строя с номером «1»; при этом последний интервал определяется до знака «финиш» (возможен и другой вариант — определение первого интервала от начала текста — «старта»). Интервалы данной однородной последовательности разместим в соответствии с номерами считываемых элементов в первой строке матрицы. Далее, при втором просмотре строя текста аналогично выберем элементы с номером «2» и разместим вектор интервалов, соответствующий другой однородной последовательности, во второй строке матрицы. В каждой следующей строке помещается вектор интервалов «новой» при очередном просмотре однородной последовательности. Одиночные знаки, слова или сообщения будут представлены всего одним интервалом (до финиша), который размещается в крайнем столбце соответствующей строки матрицы. Число столбцов $n_{j\,max}$ в «*матрице интервалов» однородных цепей* равно числу вхождений самого частого знака (или слова) текста. Незанятые интервалами элементы матрицы заполним нулями. Число строк m равно мощности алфавита или словаря текста. Результаты описанных действий представлены на рисунке 5.

В случае правильного выполнения декомпозиций, полученные множества однородных последовательностей (рисунки 6, 7) будут несовместными (так как не содержат занятых мест с одинаковыми номерами на их позициях). *Композиция* или «совмещение» всех

неполных однородных кортежей дает исходную полную неоднородную последовательность.

T	**T**	**G**	**G**	**G**	**T**	**T**	**C**	**C**	**G**	**G**	**G**	**G**	**G**	**G**	*Cricetulus griseus*
G	**G**	**A**	**A**	**A**	**G**	**G**	**T**	**T**	**A**	**A**	**A**	**A**	**A**	**A**	*Homo sapiens*
1	**1**	**2**	**2**	**2**	**1**	**1**	**3**	**3**	**2**	**2**	**2**	**2**	**2**	**2**	строй, общий для обоих фрагментов

Рис. 6. Фрагменты нуклеотидных цепей *Cricetulus griseus* и *Homo sapiens* с одинаковым строем (длина фрагментов 15). Выделены путем просмотра рибосомальной РНК общей длиной 1871и 1559. Совпадение строя фрагментов начинается с позиций: 1157 для первой цепочки и 778 — для второй цепочки

G	-	-	-	-	G	G	-	-	-	-	-	-	-	G	G	-	-	G
A	-	-	-	-	A	A	-	-	-	-	-	-	-	A	A	-	-	A
1	0	0	0	0	1	1	0	0	0	0	0	0	0	1	1	0	0	1

Рис. 7. Однородные фрагменты нуклеотидных цепей *Cricetulus griseus* и *Homo sapiens* с одинаковым строем (длина фрагментов 19). Совпадение строя начинается с позиций: 108 для первой цепочки и 1847 — для второй цепочки

Следует также отметить описание знаковых цепей и текстов графами. Обычно это взвешенный граф типа «дерево», узлы которого представляют выделенные по определенным признакам символы или слова, а ребра — это интервалы между ними.

Еще раз подчеркнем, что *для исследования построения реальной информационной цепи вводится формальное понятие — строй знаковой последовательности*, который представляет только определенный порядок следования, расположение различимых и одинаковых его компонентов без учета их обозначений и содержимого. Заметим, что цели и методы исследования строя, если его рассматривать как обычный кортеж, не отличаются от исследований реальных текстов и знаковых цепей и тому подобных. Если рассматривать строй как новый абстрактный объект, отображающий информационную цепь, то открывается возможность исследовать и использовать его особые свойства, в том числе применять новые формулы для подсчета информации в массиве данных [32].

3. О мерах сходства-расхождения цепей на основе числовых характеристик строя. Первые разработки по применению развиваемого здесь подхода для сравнения построений цепей представлены в публикациях [44, 45]. Целью же данной статьи является представление разработанного авторами нового подхода и инструментария для сравнения

символьных последовательностей на основе характеристик строя их частей (подпоследовательностей).

Определим *меры сходства и расхождения* оригинальных построений разных знаковых последовательностей A и B на основе числовых характеристик строя отдельной цепи. В данной работе используются характеристики глубины (G) и средней удаленности (g). В частности, глубины расположения всех одинаковых знаков в однородной цепи и любых одинаковых символов в полной неоднородной цепи, соответственно, определяются в виде:

$$G_j = \sum_{i=1}^{n_j} \log \Delta_{ij} = \sum_{i=1}^{n_j} g_{ij}; \tag{1}$$

$$G = \sum_{j=1}^{m} G_j, \tag{2}$$

где Δ_{ij} — интервал от i-го до $(i+1)$ вхождения j-го символа в однородной цепи (например, в литературных текстах это могут быть интервалы между ближайшими одинаковыми словами или буквами, а в нуклеотидных последовательностях — это интервал между двумя ближайшими одинаковыми нуклеотидами — межнуклеотидное расстояние); g_{ij} — удаленность i-го вхождения j-го символа (текущая удаленность); n_j — число вхождений j-го символа (нуклеотида); m — мощность алфавита в последовательности ($m = 4$ — число различных нуклеотидов {A, T, C, G}).

В свою очередь, средние удаленности одинаковых символов в однородной цепи и полной неоднородной знаковой цепи определяются в виде:

$$g_j = G_j / n_j; \tag{3}$$

$$g = G/n, \tag{4}$$

где n — длина последовательности (гена, генома и тому подобного), определяемая числом компонентов (например, нуклеотидов).

С помощью компьютерных экспериментов было установлено, что числовые характеристики строя цепи с высокой точностью и

однозначно отображают оригинальное построение компонентов в данной последовательности [46].

Для приближенного сравнения при условии равенства мощностей алфавитов (словарей), когда $m_A = m_B = m$, сравниваемых цепей A и B, определим меры их расхождения в виде:

$$\Delta G_1 = |G_A - G_B|; \tag{5}$$

$$\Delta g_1 = |g_A - g_B|, \tag{6}$$

где G_A, G_B, g_A, g_B — соответственно глубины и средние удаленности полных цепей A и B. Формула (6) позволяет сравнивать последовательности разной длины.

Назовем отдельные части и так называемые однородные цепи, составляющие полную неоднородную знаковую цепь (кортеж) A, одинаково — частями (целого). Частями целого текста могут быть, например, его предложения, абзацы, параграфы и тому подобное. В качестве частей полного генома принято выделять гены, регуляторные зоны, рибосомальные и транспортные РНК и другие составляющие аннотации генома [47].

Однородная цепь представляет собой кортеж, алфавит которого составляют два компонента: выделенный элемент алфавита полной последовательности и «пустой» элемент. Выделенный элемент алфавита располагается на позиции так же, как и в полной последовательности, а все остальные места заняты «пустым» элементом. В качестве однородной цепи на основе выделенного слова в данном тексте выступает полная многоместная позиция данного текста, в которой «заняты» выделенным словом только отдельные места; остальные места на позиции пусты.

Для точного сравнения построений отдельных частей в разных цепях A и B необходимы следующие действия:

1. Получить два распределения значений интегральной характеристики строя наборов частей (составляющих A и B) — $\{G_{A_j}\}$ и $\{G_{B_k}\}$. В этих распределениях A_j и B_k соответственно j-ая и k-ая части сравниваемых цепей A и B; $j = 1, 2, \ldots, m_A,\ k = 1, 2, \ldots, m_B$; где m_A, m_B — количество частей в цепях A и B, выбранных исследователями для их сравнения.

2. Построить ранговые распределения значений характеристики строя для двух наборов частей A и B, в которых $G_{A_j} \leqslant G_{A_{j+1}}$ и $G_{B_K} \leqslant G_{B_{k+1}}$ (возможно построение убывающих распределений как на рисунке 8).

При совпадении количества частей в разных последовательностях, когда $m_A = m_B = m$, для более точного сравнения можно использовать меры расхождения построений цепей A и B в виде:

$$\Delta G_2 = \sum_{j=1}^{m} |G_{A_j} - G_{B_j}|; \tag{7}$$

$$\Delta g_2 = \sum_{j=1}^{m} |g_{A_j} - g_{B_j}|. \tag{8}$$

И, наконец, для детального сравнения построений цепей A и B, когда совпадают числа вхождения n_j j-ых компонентов у сравниваемых частей данных цепей, можно использовать поинтервальную меру расхождения, в которой сравниваются текущие удаленности в виде:

$$\Delta g_3 = \sum_{j=1}^{m} \sum_{i=1}^{n_j} |g_{A_{ji}} - g_{B_{ji}}|. \tag{9}$$

В случаях, когда размер наборов частей сравниваемых цепей A и B различается и $m_A \neq m_B$, предлагается использовать меры сходства и расхождения, в которых *приоритетно учитываются подмножества сходных по строю (с определенной точностью) отдельных частей, составляющих сравниваемые цепи* A *и* B ($\{A_i\} \subseteq \{A_j\}$ и $\{B_i\} \subseteq \{B_k\}$).

Для предварительного отбора сходных по расположению компонентов в каждой из сравниваемых частей предлагается использовать интегральные характеристики строя: G_{A_i}, G_{B_i}.

При этом сходной парой назовем сравниваемые части цепей A_i и B_i, для которых относительное расхождение значений интегральной характеристики их строя не превышает некоторую величину $\delta \leqslant 1$. Такой критерий сходства пар частей определим в виде:

$$\delta_{AB_i} = \frac{|G_{A_i} - G_{B_i}| \cdot 2}{G_{A_i} + G_{B_i}} \leqslant \delta, \tag{10}$$

где индекс i представляет номер пары совпадающих частей у сравниваемых цепей; при этом количества таких частей могут быть разными, когда $m_{A_i} \neq m_{B_i}$.

Определим среднее относительное отклонение по множеству совпадающих пар $\{\langle A_i, B_i\rangle\}$ в виде:

$$\delta_{AB} = \frac{1}{M_{AB}} \cdot \sum_{i=1}^{M_{AB}} \delta_{AB_i}, \tag{11}$$

где $M_{AB} = |\{\langle A_i, B_i\rangle\}|$ — мощность множества сходных по строю пар частей.

Кроме отмеченного в пунктах 1) и 2) для сравниваемых ранговых распределений значений характеристики строя отдельных частей $\{G_{A_j}\}$ и $\{G_{B_k}\}$ (у которых $m_A \neq m_B$) необходимо определить *сходные по строю части двух цепей* (в количестве $0 \leqslant m_{AB} \leqslant m_A + m_B$). *Для ранговых распределений сходные части можно выделить за небольшое число проходов (без полного перебора).*

Сходные по строю пары частей из состава цепей A и B определяются в два этапа:

– приближенным отбором с помощью критерия (10), в котором используются пары значений характеристики строя G_{A_i}, G_{B_i};

– дальнейшим разделением списка пар, выделенных на первом этапе, при котором пары попадают в одно из трех подмножеств: совпадающие по строю пары, сходные по строю (с малыми различиями строя), «псевдосходные» — сходные только по значениям характеристики G_{A_i}, G_{B_i}, но не являющиеся при этом гомологичными.

Для тонкой селекции сходных пар можно использовать инструменты ФОАС: локальные и однородные характеристики строя, а в случае нуклеотидных последовательностей — общепринятые имена частей (имена компонентов аннотации) [48]. В такой функции «аргументом» является начальная позиция фрагмента (L-граммы) в последовательности. Значения функции вычисляются «скользящим окном» для множества L-грамм — последовательных, равных по длине и смещенных на один элемент относительно друг друга участков цепи (в примере ниже длина L-грамм — $l = 50$ нуклеотидов). Фактически, функция характеристики строя позволяет поэлементно представлять рассматриваемую знаковую цепь. То есть для каждого компонента последовательности (кроме $l - 1$ конечных компонентов) может быть вычислено соответствующее ему значение «локальной» характеристики.

На практике тонкая селекция осуществлялась непосредственным рассмотрением пар графиков функций характеристик. Возможна автоматизация сравнения пар функций.

После выделения сходных по строю частей последовательностей A и B (при условии их наличия) предлагается использовать следующие меры их сходства и расхождения:

1. Относительное число (доля) сходных по строю частей.
2. Среднее относительное расхождение ранговых распределений.
3. Относительная глубина сходных по строю частей.

Первая мера определяется в виде:

$$\delta_1 = \frac{m_{AB}}{m_A + m_B} \leqslant 1, \tag{12}$$

где $m_{AB} = 2 \cdot m_{A_i}$, если $m_{A_i} < m_{B_i}$;
$m_{AB} = 2 \cdot m_{B_i}$, если $m_{A_i} > m_{B_i}$;
$m_{AB} = m_{A_i} + m_{B_i}$, если $m_{A_i} = m_{B_i}$;
$\delta_1 = 1$, если $m_{A_i} = m_A$ и $m_{B_i} = m_B$.

Вторая мера определяется в виде:

$$\delta_2 = \frac{\delta_{AB}}{\delta_1}. \tag{13}$$

С учетом первой меры видно, что уменьшение доли сходных по строю частей искусственно увеличивает расхождение ранговых распределений цепей A и B.

Третья мера представляет отношение суммарных значений характеристики строя сходных частей к сумме этих характеристик для всех составляющих частей сравниваемых цепей A и B, в виде:

$$\delta_3 = \frac{G_{AB}}{G_A + G_B}, \tag{14}$$

где $G_{AB} = 2 \cdot \sum_{i=1}^{m_{A_i}} G_{A_i}$, если $\sum_{i=1}^{m_{A_i}} G_{A_i} < \sum_{i=1}^{m_{B_i}} G_{B_i}$;

$G_{AB} = 2 \cdot \sum_{i=1}^{m_{B_i}} G_{B_i}$, если $\sum_{i=1}^{m_{A_i}} G_{A_i} > \sum_{i=1}^{m_{B_i}} G_{B_i}$;

$G_{AB} = \sum_{i=1}^{m_{A_i}} G_{A_i} + \sum_{i=1}^{m_{B_i}} G_{B_i}$, если $\sum_{i=1}^{m_{A_i}} G_{B_i} = \sum_{i=1}^{m_{B_i}} G_{B_i}$;

$\delta_3 = 1$, если $\sum_{i=1}^{m_{A_i}} G_{A_i} = G_A$ и $\sum_{i=1}^{m_{B_i}} G_{B_i} = G_B$.

4. О процедуре сравнения построений цепей на основе интегральных числовых характеристик строя. Как отмечалось, в качестве информационных цепей выступают массивы естественно упорядоченных данных разной природы: нуклеотидные последовательности (геномы, плазмиды, гены, и др.), лингвистические тексты, нотные записи, массивы данных измерений, в частности временные ряды. Определим процедуру сравнения построений двух наборов частей, представляющих знаковые последовательности A и B. Так как алгоритм (процедура) сравнения двух неубывающих ранговых распределений (за малое число проходов) очевиден для любого специалиста в области информатики, отметим только некоторые его детали:

1. Для пары сравниваемых последовательностей выделяется набор исследуемых частей. Например, для геномов можно использовать не все компоненты из аннотаций.

2. Выбирается характеристика для сравнения частей.

3. Вычисляется отображение всех выбранных частей каждой целостной знаковой цепи (генома, текста и тому подобных) соответствующими значениями характеристики.

4. Части каждой последовательности упорядочиваются по возрастанию (убыванию) вычисленной характеристики. В результате получаются два неубывающих (невозрастающих) ранговых распределения значений этой характеристики (рисунок 8).

5.

В получившихся распределениях производится поиск совпадающих с заданной точностью частей. Особенностью сравнения двух ранговых распределений является необходимость сопоставления очередного элемента данного распределения, в том числе с теми элементами другого распределения, которые выявлены как сходные для предыдущего элемента данного распределения. При этом предполагается, что как в первом, так и во втором распределениях возможны «последовательности» одинаковых по характеристике частей.

6. При нахождении совпадающей части из второй цепи сохраняем данную пару совпадающих (или схожих) частей.

7. Кроме перебора отдельных частей первой цепи на предмет сходства с частями второй цепи, необходимо выполнить данную процедуру для выявления сходства частей второй цепи относительно первой.

8. После выявления всех пар сходных частей первой и второй цепи, вычисляются значения следующих характеристик:

– среднее относительное отклонение для множества сходных пар по формуле (11);

– мера сходства пары сравниваемых цепей по формуле (12), определяемая долей сходных по строю частей в каждой из цепей по отношению к общему числу составляющих их частей;

– мера расхождения данной пары цепей по формуле (13), в которой искусственно завышается среднее относительное отклонение сходных по строю частей за счет доли несовпадающих частей;

– мера сходства по формуле (14), определяемая долей суммы значений характеристики сходных по строю частей.

5. Матрица сходства-расхождения для набора цепей. На практике возникает необходимость попарного сравнения построений некоторого множества цепей. Это могут быть геномы наборов организмов (возможно сходных), оригинальный текст и множество его изложений, множество текстов с одинаковым содержанием, записанных на разных языках.

Результат такого сравнения может быть представлен в виде квадратной матрицы. Ячейки такой матрицы могут содержать как значения всех мер сходства-расхождения, вычисленных по формулам (12), (13), (14), так и только некоторые из них (рисунок 12). Первая строка матрицы формируется как результат сравнения первой цепи со всеми другими цепями. Затем определяется сходство второй цепи со всеми рассматриваемыми цепями и так далее. При этом каждой ячейке матрицы соответствует множество сходных пар частей сравниваемых цепей.

6. Примеры сравнения плазмид разных штаммов бактерии *Coxiella burnetii* на основе числовых характеристик строя их частей (компонентов аннотаций). Для апробации разработанного инструментария использовались знаковые (нуклеотидные) последовательности плазмид (и их аннотации) семи штаммов *Coxiella burnetii*, взятых из GenBank [49]. Их нумерованный список представлен под рисунком 12.

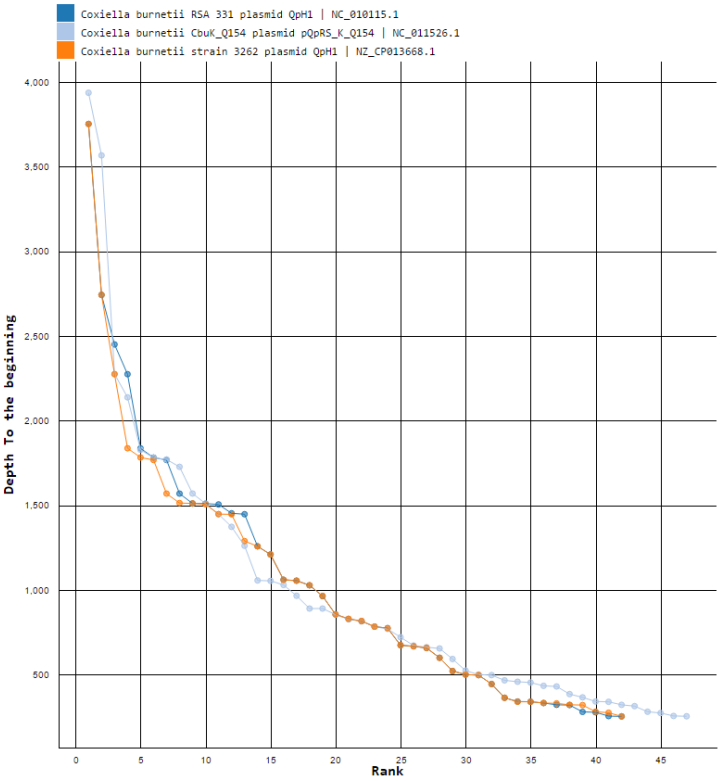


Рис. 8. Ранговые распределения значений характеристики G_j генов и других компонентов аннотаций плазмид штаммов *Coxiella burnetii* с номерами 2, 3, 5

На рисунке 8 для примера представлены ранговые распределения значений характеристики строя G_j сопоставляемых частей (генов и других компонентов аннотаций) плазмид трех микроорганизмов из этого списка. Два распределения представляют 2-ю и 3-ю плазмиды сходные на 76,19% по мере (12). Третье распределение представляет 5-ю плазмиду. По этой мере оно имеет меньшее сходство с ними (2-я с 5-ой на 29,21%, 3-я с 5-ой 24,72%).

Ниже списком представлены некоторые сходные по характеристике G_j пары генов и степень их сходства по мере (10) с порогом $\delta \leqslant 0.1$, вычисленные для пары плазмид штаммов *Coxiella burnetii* с номерами 1 и 3, которые сравнивались по мере (12). На рисунках 9, 10, 11 показаны пары графиков, представляющих функции характеристик строя

сопоставляемых генов.

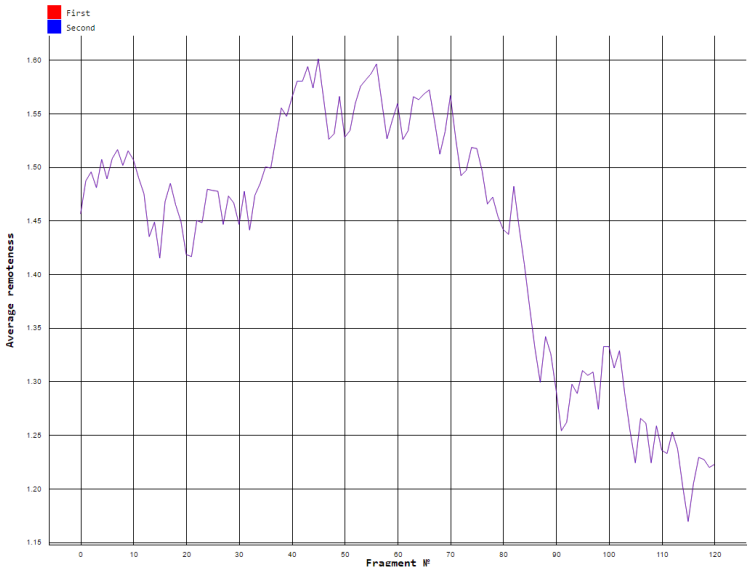


Рис. 9. Графики двух функций характеристики g совпадающих по строю генов 1-ой и 3-ей плазмид (графики полностью совпадают)

Совпадающие по строю гены:
First sequence: Coxiella burnetii RSA 493 plasmid pQpH1 | NC_004704.1
Feature: Coding DNA sequence. Position = 12829. Length = 171.
Product: hypothetical protein. Characteristic value = 252.39884557174787.

Second sequence: Coxiella burnetii RSA 331 plasmid QpH1 | NC_010115.1
Feature: Coding DNA sequence. Position = 1863. Length = 171.
Product: tyrosine recombinase. Characteristic value = 252.39884557174787.

Абсолютное расхождение — 0. Относительное расхождение — 0%.

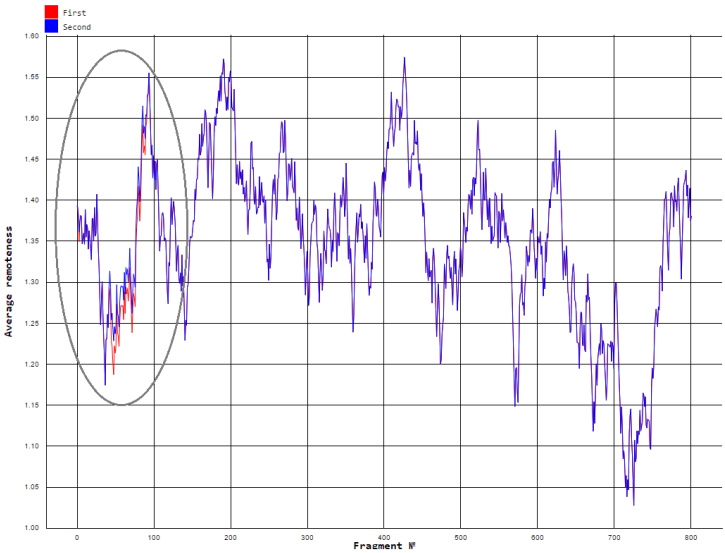


Рис. 10. Графики двух функций характеристики g сходных по строю генов 1-ой и 3-ей плазмид (отмечено место расхождения)

Сходные по строю гены:
First sequence: Coxiella burnetii RSA 493 plasmid pQpH1 | NC_004704.1
Feature: Coding DNA sequence. Position = 19735. Length = 852 Complement.
Product: hypothetical protein. Characteristic value = 1207.4344521538703.

Second sequence: Coxiella burnetii RSA 331 plasmid QpH1 | NC_010115.1
Feature: Coding DNA sequence. Position = 8763. Length = 852. Complement.
Product: hypothetical protein. Characteristic value = 1208.6188767250078.

Абсолютное расхождение — 1.18442457113. Относительное расхождение — 0.098046227455%.

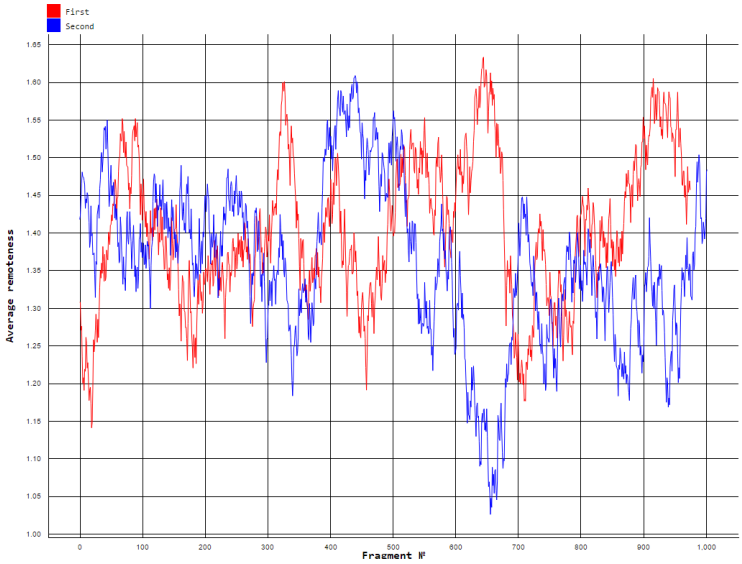


Рис. 11. Графики двух функций характеристики g псевдосходных генов 1-ой и 3-ей плазмид

Псевдосходные гены:
First sequence: Coxiella burnetii RSA 493 plasmid pQpH1 | NC_004704.1
Feature: Coding DNA sequence. Position = 2545. Length = 1026
Product: hypothetical protein. Characteristic value = 1510.194562707094.

Second sequence: Coxiella burnetii RSA 331 plasmid QpH1 | NC_010115.1
Feature: Coding DNA sequence. Position = 9641. Length = 1053. Complement.
Product: hypothetical protein. Characteristic value = 1509.2896095025235.

Абсолютное расхождение — 0.90495320457. Относительное расхождение — 0.059940913941%.

Для демонстрации значений трех мер сходства-расхождения плазмид ниже представлена матрица для всех семи микроорганизмов списка. Рассматриваемые микроорганизмы предварительно упорядочены и пронумерованы на основе одной интегральной характеристики строя — средней удаленности g, вычисленной для каждой плазмиды.

На рисунке 12 по горизонтали и вертикали соответствующими номерами обозначены последовательности:

1 - Coxiella burnetii RSA 493 plasmid pQpH1 | NC_004704.1
2 - Coxiella burnetii strain 3262 plasmid QpH1 | NZ_CP013668.1
3 - Coxiella burnetii RSA 331 plasmid QpH1 | NC_010115.1
4 - Coxiella burnetii str. Namibia plasmid QpRS | NZ_CP007556.1
5 - Coxiella burnetii CbuK_Q154 plasmid pQpRS_K_Q154 | NC_011526.1
6 - Coxiella burnetii MSU Goat Q177 plasmid QpRS | NC_010258.1
7 - Coxiella burnetii Dugway 5J108-111 plasmid pQpDG | NC_009726.1

Рис. 12. Матрица сходства-расхождения плазмид семи штаммов *Coxiella burnetii*

Для проведения исследований сходства цепей разработан комплекс программных средств [50], который позволил построить матрицу сходства для 42 полных геномов организмов семейства Rickettsia.

Кроме отмеченного, компьютерные эксперименты были направлены на исследование однозначности отображения строя конкретной информационной цепи соответствующим значением характеристики строя.

Результаты исследования нуклеотидных последовательностей с использованием характеристики G_j позволили сформулировать следующие предварительные выводы:

– сходные по строю пары частей (компонентов аннотаций) можно предварительно отбирать из состава двух хромосом или плазмид по критерию (10) на основе сравнения значений их интегральных характеристик G_{A_i} и G_{B_i}, различающихся на величину $\delta \leqslant 0,05\%$;

– однако даже при таком малом пороге в списке сходных по строю частей, кроме полностью совпадающих и сходных (с малыми различиями) по строю частей, попадают псевдосходные пары частей, для которых близки только величины G_{A_i} и G_{B_i} (но их построения сильно различаются, как на рисунке 11);

– для формального отсева из списка большинства псевдосходных частей следует подбирать величину порога $\delta < 0,05\%$ для данной пары сравниваемых организмов или некоторого их множества;

– рассмотрение множества пар графиков, представляющих последовательности значений локальной характеристики строя (функции характеристик строя), выявляет, во-первых, некоторое фиксированное подмножество пар частей (в данном исследовании — компонентов аннотаций), строй которых полностью совпадает, если интегральные характеристики равны (G_{A_i} и G_{B_i}); во-вторых, некоторое фиксированное подмножество сходных по строю (с малыми различиями) пар частей; в-третьих, увеличивающееся по мощности (по мере увеличения порога δ) подмножество псевдосходных пар частей, которое частично перемешивается с парами сходных частей по расхождению характеристики;

– в большинстве случаев оригинальное расположение нуклеотидов (компонентов) в данной цепи однозначно отображается уникальной цифровой последовательностью — одним числом интегральной характеристики строя. Однако изредка при равенстве характеристик $G_{A_i} = G_{B_i}$ в принципе могут фиксироваться псевдосовпавшие цепи. Такие совпадения объясняются отличиями в некоторых одинаковых местах цепей, которые не изменяют наборы (межнуклеотидных) интервалов в сравниваемых цепях;

– зачастую несовпадение между частями геномов или плазмид обусловлено низким качеством аннотаций, в частности отсутствием жесткой регламентации (стандартов) при их составлении, не точно или неполно указанной позицией компонента, не полным или некорректным его названием и тому подобным [51].

7. Заключение. В работе отмечено фактическое отсутствие адекватных средств прикладной математики и информатики для исчисления оригинального расположения знаков в символьных последовательностях.

Указан недостаток традиционного методологического подхода — редукционизма, применяемого для описания и анализа массивов естественно упорядоченных данных. Предложено подобные исследования дополнить средствами системного подхода.

Предложены меры сходства и расхождения знаковых цепей на основе числовых характеристик строя.

Представлена процедура сравнения построений цепей на основе ранговых распределений значений числовой характеристики строя, отображающих отдельные части этих цепей. Данная процедура допускает сравнение цепей, мощности алфавитов (словарей) которых могут быть различны.

Сформулированы и рассмотрены возможные исходы при попарном сравнении частей символьных последовательностей по характеристикам строя: полное совпадение; сходство по характеристике с заданной точностью; псевдосходство по заданной характеристике.

Для демонстрации возможностей предложенных мер сходства и процедуры сравнения представлены результаты сравнения плазмид семи штаммов бактерии *Coxiella burnetii* в виде матрицы сходства на основе числовых характеристик строя их частей (генов и других компонентов аннотаций).

Литература

1. *Zipf G., Kingsley G.* Selected Studies of the Principle of Relative Frequency in Language // Harvard University Press. 1932. 128 p.
2. *Гусев В.Д., Косарев Ю.Г., Титкова Т.Н.* Методы поиска и анализ статистических закономерностей в символьных последовательностях // Машинные методы обнаружения закономерностей: материалы всесоюзного симпозиума. 1976. С. 75–84.
3. *Гусев В.Д., Куличков В.А., Никулин А.Е.* Алгоритмы поиска несовершенных повторов в генетических текстах // Анализ символьных последовательностей: вычислительные системы. 1985. Вып. 113. С. 107–122.
4. *Гусев В.Д., Немытикова Л.А.* Векторная мера сложности нуклеотидных последовательностей // Третий сибирский конгресс по прикладной и индустриальной математике (ИНПРИМ-98). 1998. 115 с.
5. *Гусев В.Д., Мирошниченко Л.А., Саломатина Н.В.* Методы выделения структурных единиц в символьных последовательностях. Межъязыковые аналоги // Материалы Всероссийской конференции с международным участием «Знания-Онтологии-Теории». 2009. Т. 2. С. 53–62.
6. *Беликов С.И., Гусев В.Д., Мирошниченко Л.А., Титкова Т.Н.* Сравнительный анализ геномов вирусов клещевого энцефалита: дифференциация по степени вирулентности // Математическая биология и биоинформатика: IV международная конференция. 2012. С. 52–53.
7. *King B.R, Aburdene M., Thompson A., Warres Z.* Application of Discrete Fourier Inter-Coefficient Difference for Assessing Genetic Sequence Similarity // EURASIP Journal on Bioinformatics and Systems Biology. 2014. vol. 2014, no. 1. 8 p.
8. *Srivastava S., Baptista M.S.* Markovian language model of the DNA and its information content // Royal Society open science. 2016. vol. 3. no. 1. pp. 150527.
9. *Nair A.S.S., Mahalakshmi T.* Visualization of genomic data using inter-nucleotide distance signals // Proceedings of IEEE Genomic Signal Processing. 2005. vol. 408.
10. *Afreixo V. et al.* Genome analysis with inter-nucleotide distances // Bioinformatics. 2009. vol. 25(23). pp. 3064-3070.

11. *Jin S. et al.* A Generalized Topological Entropy for Analyzing the Complexity of DNA Sequences // PLoS One. 2014. vol. 9(2). pp. e88519.
12. *Садовский М.Г.* Информационно-статистический анализ нуклеотидных последовательностей: диссертация // Институт биофизики СО РАН. 2004. 394 с.
13. *Amiri S., Dinov I.D.* Comparison of genomic data via statistical distribution // Journal of Theoretical Biology. 2016. vol. 407. pp. 318–327.
14. *Manca V., Bonnici V.* Infogenomics Tools: A Computational Suite for Informational Analyses of Genomes // Journal of Bioinformatics, Proteomics and Imaging Analysis. 2015. vol. 1. no. 1. pp. 7-14.
15. *Арнольд В.И.* Сложность конечных последовательностей нулей и единиц и геометрия конечных функциональных пространств // Публичная лекция. 2006. Т. 13. 14 р.
16. *Kullback S., Leibler R.A.* On information and sufficiency // The Annals of Mathematical Statistics. 1951. vol. 22. no. 1. pp. 79–86.
17. *Левенштейн В.И.* Двоичные коды с исправлением выпадений, вставок и замещений символов // Доклады Академий Наук СССР. 1965. Т. 4. С. 845–848.
18. *Hamming R. W.* Error detecting and error correcting codes // Bell System Technical Journal. 1950. vol. 29(2). pp. 147–160.
19. *Zielezinski A., Vinga S., Almeida J., Karlowski W.M.* Alignment-free sequence comparison: benefits, applications, and tools // Genome Biology. 2017, vol. 18(1):186 p.
20. *Bonham-Carter O, Steele J., Bastola D.* Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis // Briefings in Bioinformatics. 2014. vol. 15(6). pp. 890-905.
21. *Song K. et al.* New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing // Briefings in Bioinformatics. 2014. vol. 15(3). pp. 343-353.
22. *Bernard G. et al.* Alignment-free inference of hierarchical and reticulate phylogenomic relationships // Briefings in Bioinformatics. 2017.
23. *Chan C.X., Ragan M.A.* Next-generation phylogenomics // Biology Direct. 2013. vol. 8(1). pp. 3.
24. *La Rosa M., Fiannaca A., Rizzo R., Urso A.* Alignment-free analysis of barcode sequences by means of compression-based methods // BMC Bioinformatics. 2013. vol. 14(7). pp. S4.
25. *Haubold B.* Alignment-free phylogenetics and population genetics // Briefings in Bioinformatics. 2013. vol. 15(3). pp. 407–418.
26. *Ren J. et al.* Alignment-Free Sequence Analysis and Applications // Annual Review of Biomedical Data Science. 2018. vol. 1. pp. 93–114.
27. *Wang S., Tian F., Feng W., Liu X.* Applications of representation method for DNA sequences based on symbolic dynamics // Journal of Molecular Structure: THEOCHEM. 2009. vol. 909. no. 1-3. pp. 33-42.
28. *Salgado-Garcia R., Ugalde E.* Symbolic Complexity for Nucleotide Sequences: A Sign of the Genome Structure // Journal of Physics A: Mathematical and Theoretical. 2016. vol. 49. no. 44. pp. 445601.
29. *Шрейдер Ю.А., Шаров А.А.* Системы и модели // М.: Радио и связь. 1982. 152 с.
30. *Мазур М.* Качественная теория информации // М.: Мир. 1974. 240 с.
31. *Gumenjuk A., Kostyshin A., Simonova S.* An approach to the research of the structure of linguistic and musical texts // Glottometrics. 2002. vol. 3. pp. 61–89.

32. *Гуменюк А. С., Поздниченко Н. Н., Родионов И. Н., Шпынов С.Н.* О средствах формального анализа строя нуклеотидных цепей // Математическая биология и биоинформатика. 2013. Т. 8. № 1. С. 373-397.
33. *Freitas A., Afreixo V., Cruz S.E.* Mixture models of geometric distributions in genomic analysis of inter-nucleotide distances // Statistics, optimization & information computing Stat. 2013. Vol. 1. no. 1. pp. 8–28.
34. *Wasito I., Veritawati I.* Fractal Dimension Approach for Clustering of DNA Sequences Based on Internucleotide Distance // IEEE 2013 International Conference of Information and Communication Technology (ICoICT). 2013. pp. 82–87.
35. *Tavares A. et al.* Detection of exceptional genomic words: a comparison between species // 22nd International Conference on Computational Statistics (COMPSTAT 2016). 2016.
36. *Zhou L.Q., Li R., Han G.S.* A Method Based on the Improved Inter-Nucleotide Distances of Genomes to Construct Vertebrates Phylogeny Tree // IEEE 2014 7th International Conference on Biomedical Engineering and Informatics. 2014. pp. 776-780.
37. *Kolekar P., Kale M., Kulkarni-Kale U.* Alignment-free distance measure based on return time distribution for sequence analysis: Applications to clustering, molecular phylogeny and subtyping // Molecular Phylogenetics and Evolution. 2012. vol. 65. no. 2. pp. 510–522.
38. *Bonnici V., Manca V.* Recurrence Distance Distributions in Computational Genomics // American Journal of Bioinformatics and Computational Biology. 2015. vol. 3. pp. 5-23.
39. *Messaoudi I., Oueslati A.E., Lachiri Z.* Wavelet analysis of frequency chaos game signal: a time-frequency signature of the C. elegans DNA // EURASIP Journal on Bioinformatics and Systems Biology. 2014. vol. 2014(1). pp. 16.
40. *Орлов Ю.К.* Частотные структуры конечных сообщений в некоторых естественных информационных системах: диссертация // Издательство Тбилисского университета. 1974.
41. *Орлов Ю.К.* Невидимая гармония // Число и мысль. 1980. Вып. 3. С. 70-105.
42. *Кудрин Б.И.* Философия техники: основания постнеклассической философии техники // М.: Техника. 2007. Вып. 36. 196 с.
43. *Попова О.В., Гельфанд М.С.* Существует ли аналог закона Ципфа в генетическом языке? // Научно-техническая информация. Серия 2: Информационные процессы и системы. 2000. № 4. С. 19-24.
44. *Волчкова И.А., Гуменюк А.С.* О мерах сходства разноязычных текстов с одинаковым содержанием. // Материалы Всероссийской конференции с международным участием «Знания – Онтологии – Теории» (ЗОНТ-13). 2013. Т. 1. С. 98-105.
45. *Гуменюк А.С., Волчкова И.А.* Использование средств анализа строя знаковой последовательности для формальной оценки качества перевода. // Омский научный вестник. 2013. Т. 3(123). С. 251-256.
46. *Шпынов С.Н., Гуменюк А.С., Поздниченко Н.Н.* Применение числовой характеристики строя нуклеотидов в геномах прокариот для реклассификации внутри рода Rickettsia // Математическая биология и биоинформатика. 2016. Т. 11. № 2. С. 336-350.
47. The DDBJ/ENA/GenBank Feature Table Definition. URL: http://www.insdc.org/files/feature_table.html (дата обращения: 15.04.2018).
48. *Гуменюк А.С., Поздниченко Н.Н., Шпынов С.Н.* Формальный анализ строя локальной структуры нуклеотидных последовательностей // Вестник Томского государственного университета. 2014. Т. 4(29). С. 23–30.
49. GENBANK DataBase. URL: http://www.ncbi.nlm.nih.gov/nuccore/ (дата обращения: 02.03.2018).

50. *Гуменюк А.С., Поздниченко Н.Н., Скиба А.А., Шпынов С.Н.* Матрица сходства нуклеотидных последовательностей по их компонентам. Свидетельство о государственной регистрации программы для ЭВМ. №2017616679. 09.06.2017.

51. *Поздниченко Н.Н., Гуменюк А.С., Шпынов С.Н.* О картографическом представлении множества геномов прокариот с помощью числовых характеристик строя их компонентов. // Новые информационные технологии в исследовании сложных структур: материалы 11-й международной конференции. 2016. С. 84-85.

Гуменюк Александр Степанович — канд. техн. наук, доцент, доцент кафедры информатики и вычислительной техники факультета информационных технологий и компьютерных систем, Омский государственный технический университет (ОмГТУ). Область научных интересов: формальный анализ строя компонентов в массивах упорядоченных данных, формальный анализ строя литературных текстов, нотных записей и нуклеотидных последовательностей, социальные системы информационных коммуникаций, формализация типов общения, теория и организация обучения студентов на основе делового общения, алгебра ментальных событий, физические основания передачи сообщений. Число научных публикаций — около 200. gumas45@mail.ru; пр. Мира, 11, 644050, Омск, Российская Федерация; р.т.: +7(3812) 65-24-98.

Скиба Артемий Андреевич — инженер-программист, ООО "Компания Элмис". Область научных интересов: формальный анализ строя, формальный анализ строя нотных записей и нуклеотидных последовательностей, биоинформатика. Число научных публикаций — 13. skiba.artem@inbox.ru; Маяковского, 14, 644046, Омск, Российская Федерация;+7(3812) 51-06-30.

Поздниченко Николай Николаевич — старший преподаватель, кафедра информатики и вычислительной техники факультета информационных технологий и компьютерных систем, Омский государственный технический университет (ОмГТУ). Область научных интересов: формальный анализ строя, формальный анализ строя нуклеотидных последовательностей, биоинформатика. Число научных публикаций — 45. nick670@yandex.ru; пр. Мира, 11, 644050, Омск, Российская Федерация; р.т.: +7(3812) 65-24-98.

Шпынов Станислав Николаевич — д-р мед. наук, заведующий лабораторией экологии риккетсий, лаборатория экологии риккетсий, Федеральное государственное бюджетное учреждение «Федеральный научно-исследовательский центр эпидемиологии и микробиологии имени почетного академика Н.Ф. Гамалеи». Область научных интересов: микробиология, экология, эпидемиология, молекулярная биология, биоинформатика. Число научных публикаций — около 200. stan63@inbox.ru; Н.Ф. Гамалеи, 18, 123098, Москва, Российская Федерация; р.т.: +7(499) 193-61-85.

DOI 10.15622/sp.18.2.471-503

A.S. GUMENYUK, A.A. SKIBA, N.N. POZDNICHENKO, S.N. SHPYNOV

ABOUT SIMILARITY MEASURES OF COMPONENTS ARRANGEMENT OF NATURALLY ORDERED DATA ARRAYS

Gumenyuk A.S., Skiba A.A., Pozdnichenko N.N., Shpynov S.N. **About Similarity Measures of Components Arrangement of Naturally Ordered Data Arrays.**

Abstract. At present, mathematical tools that adequately take into account the arrangement of components are not widespread in works of specialists in the fields of research of naturally ordered data of different nature. Therefore, it is difficult or impossible to measure and compare the order of messages allocated in long information chains. The main approaches for comparing symbol sequences are using probabilistic models and statistical tools, pairwise and multiple alignment, which makes it possible to determine the degree of similarity of sequences using edit distance measures. The noted approaches almost do not pay attention to the study and detection of the patterns of the specific arrangement of all symbols, words, and components of data sets that constitute a separate sequence. The object of study in our works is a specifically organized numerical tuple — the arrangement of components (order) in symbolic or numerical sequence. The intervals between the closest identical components of the order are used as the basis for the quantitative representation of the chain arrangement. Multiplying all the intervals or summing their logarithms allows one to get numbers that uniquely reflect the arrangement of components in a particular sequence. These numbers, allow us to obtain a whole set of normalized characteristics of the order, among which the geometric mean interval and its logarithm. In this paper, we present an approach for quantitative comparing the arrangement of arrays of naturally ordered data (information chains) of an arbitrary nature. The measures of similarity/distinction and procedure of comparison of the chain order, based on the selection of a list of equal and similar by the order characteristics of the subsequences, are proposed. Rank distributions are used for faster selection of a list of matching components. The paper presents a toolkit for comparing the order of information chains and demonstrates some of its applications for studying the structure of nucleotide sequences.

Keywords: data array, symbolic sequence, information chain, numeric characteristics of order, depth of order, average remoteness, nucleotide sequence, similarity measures, similarity matrix, alignment-free genome comparison, inter-nucleotide distance.

Gumenyuk Alexander Stepanovich — Ph.D., Associate Professor, Associate Professor of Informatics and computer technology Department of Information Technologies and Computer Systems Faculty, Omsk State Technical University (OmSTU). Research interests: formal order analysis of components in ordered data arrays, formal order analysis of literary works, musical scores and nucleotide sequences, social systems of information communications, formalization of communication types, theory and organization of student learning based on professional communication, mental event algebra, physical foundation of messages transmission. The number of publications — about 200. gumas45@mail.ru; 11, pr. Mira, 644050, Omsk, Russian Federation; office phone: +7(3812) 65-24-98.

Skiba Artemiy Andreevich — Software Developer, Company Elmis. Research interests: formal order analysis, formal order analysis of musical scores and nucleotide sequences, bioinformatics. The number of publications — 13. skiba.artem@inbox.ru; 14, Mayakovskogo, 644046, Omsk, Russian Federation; office phone: +7(3812) 51-06-30.

Pozdnichenko Nikolay Nikolaevich — Senior Lecturer, Informatics and computer technology Department of Information Technologies and Computer Systems Faculty, Omsk State Technical University (OmSTU). Research interests: formal order analysis, formal order analysis of nucleotide sequences, bioinformatics. The number of publications — 45. nick670@yandex.ru; 11, pr. Mira, 644050, Omsk, Russian Federation; office phone: +7(3812) 65-24-98.

Shpynov Stanislav Nikolaevich — Ph.D., Dr.Sci., Head of Laboratory, Laboratory of Rickettsia Ecology, N. F. Gamaleya Federal Research Center for Epidemiology & Microbiology. Research interests: microbiology, ecology, epidemiology, molecular biology, bioinformatics. The number of publications — about 200. stan63@inbox.ru; 18, N.F. Gamaleya, 123098, Moscow, Russian Federation; office phone: +7(499) 193-61-85.

References

1. *Zipf G., Kingsley G.* Selected Studies of the Principle of Relative Frequency in Language. Harvard University Press. 1932. 128 p.
2. Gusev V.D., Kosarev Y.G., Titkova T.N. [Methods of search and analysis of statistical regularities in character sequences]. *Mashinnye metody obnaruzheniya zakonomernostej (Materialy vsesoyuznogo simpoziuma)* [Machine methods for detecting regularities: Materials of the All-Union Symposium]. 1976. pp. 75–84. (In Russ.).
3. Gusev V.D., Kulichkov V.A., Nikulin A.E. [Algorithms for searching for imperfect repetitions in genetic texts]. *Analiz simvol'nyh posledovatel'nostej: Vychislitel'nye sistemy – Analysis of symbol sequences: Computing systems.* 1985. vol. 113. pp. 107–122. (In Russ.).
4. Gusev V.D., Nemytikova L.A. [Vector measure of complexity of nucleotide sequences]. *Tretiy sibirskiy kongress po prikladnoy i industrial'noy matematike (INPRIM-98)* [The Third Siberian Congress on Applied and Industrial Mathematics (APLINM-98)]. 1998. 115 p. (In Russ.).
5. Gusev V.D., Miroshnichenko L.A., Salomatina N.V. [Methods for allocating structural units in character sequences. Interlingual analogues]. *Materialy Vserossiyskoy konferencii s mezhdunarodnym uchastiem «Znaniya-Ontologii-Teorii»* [Materials of the All-Russian Conference with International Participation «Knowledge – Ontologies – Theories»]. 2009. Issue 2. pp. 53–62. (In Russ.).
6. Belikov S.I., Gusev V.D., Miroshnichenko L.A., Titkova T.N. [Comparative analysis of genomes of tick-borne encephalitis viruses: differentiation according to the degree of virulence]. *Matematicheskaya biologiya i bioinformatika: IV mezhdunarodnaya konferenciya* [Mathematical biology and bioinformatics: IV international conference]. 2012. pp. 52–53. (In Russ.).
7. King B.R., Aburdene M., Thompson A., Warres Z. Application of Discrete Fourier Inter-Coefficient Difference for Assessing Genetic Sequence Similarity. *EURASIP Journal on Bioinformatics and Systems Biology.* 2014. vol. 2014. no. 1. 8 p.
8. Srivastava S., Baptista M.S. Markovian language model of the DNA and its information content. *Royal Society open science.* 2016. vol. 3. no. 1. pp. 150527.
9. Nair A.S.S., Mahalakshmi T. Visualization of genomic data using inter-nucleotide distance signals. Processings of IEEE Genomic Signal Processing. 2005. vol. 408.
10. Afreixo V. et al. Genome analysis with inter-nucleotide distances. *Bioinformatics*. 2009. vol. 25(23). pp. 3064–3070.
11. Jin S. et al. A Generalized Topological Entropy for Analyzing the Complexity of DNA Sequences. *PLoS One.* 2014. vol. 9(2). pp. e88519.
12. Sadovskii M.G. *Informacionno-statisticheskiy analiz nukleotidnyh posledovatel'nostey: dissertaciya* [Informational-statistical analysis of nucleotide sequences: Ph.D. Thesis]. Institute of Biophysics SB RAS. 2004. 394 p. (In Russ.).

13. Amiri S., Dinov I.D. Comparison of genomic data via statistical distribution. *Journal of Theoretical Biology.* 2016. vol. 407. pp. 318–327.
14. Manca V., Bonnici V. Infogenomics Tools: A Computational Suite for Informational Analyses of Genomes. *Journal of Bioinformatics, Proteomics and Imaging Analysis.* 2015. vol. 1. no. 1. pp. 7-14.
15. Arnold V.I. [The complexity of finite sequences of zeros and ones and the geometry of finite function spaces]. *Publichnaya lekciya* [Public lecture]. 2006. vol. 13. 14 p.
16. Kullback S., Leibler R.A. On information and sufficiency. *The Annals of Mathematical Statistics*. 1951. vol. 22. no. 1. pp. 79–86.
17. Levenshtein V.I. [Binary codes capable of correcting deletions, insertions, and reversals]. *Doklady Akademiy Nauk SSSR - Soviet Physics Reports*. 1965. vol. 4. pp. 845-848. (In Russ.).
18. Hamming R.W. Error detecting and error correcting codes. *Bell System Technical Journal.* 1950. vol. 29(2). pp. 147–160.
19. Zielezinski A., Vinga S., Almeida J., Karlowski W.M. Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biology.* 2017. vol. 18(1). 186 p.
20. Bonham-Carter O., Steele J., Bastola D. Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis. *Briefings in Bioinformatics.* 2014. vol. 15(6). pp. 890–905.
21. Song K. et al. New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing. *Briefings in Bioinformatics.* 2014. vol. 15(3). pp. 343–353.
22. Bernard G. et al. Alignment-free inference of hierarchical and reticulate phylogenomic relationships. *Briefings in Bioinformatics.* 2017.
23. Chan C.X., Ragan M.A. Next-generation phylogenomics. *Biology Direct.* 2013. vol. 8(1). pp. 3.
24. La Rosa M., Fiannaca A., Rizzo R., Urso A. Alignment-free analysis of barcode sequences by means of compression-based methods. *BMC Bioinformatics.* 2013. vol. 14(7). pp. S4.
25. Haubold B. Alignment-free phylogenetics and population genetics. *Briefings in Bioinformatics.* 2013. vol. 15(3). pp. 407–418.
26. Ren J. et al. Alignment-Free Sequence Analysis and Applications. *Annual Review of Biomedical Data Science.* 2018. vol. 1. pp. 93–114.
27. Wang S., Tian F., Feng W., Liu X. Applications of representation method for DNA sequences based on symbolic dynamics. *Journal of Molecular Structure: THEOCHEM.* 2009. vol. 909. no. 1-3. pp. 33–42.
28. Salgado-Garcia R., Ugalde E. Symbolic Complexity for Nucleotide Sequences: A Sign of the Genome Structure. *Journal of Physics A: Mathematical and Theoretical.* 2016. vol. 49. no. 44. pp. 445601.
29. Shreider Y.A., Sharov A.A. *Sistemy i modeli* [Systems and models]. – M.: Radio i svyaz. 1982. 152 p. (In Russ.).
30. Mazur M. *Kachestvennaya teoriya informacii* [Qualitative information theory]. M.: Mir. 1974. 240 p. (In Russ.).
31. Gumenjuk A., Kostyshin A., Simonova S. An approach to the research of the structure of linguistic and musical texts. *Glottometrics.* 2002. vol. 3. pp. 61–89.
32. Gumenuk A.S., Pozdnichenko N.N., Rodionov I.N., Shpynov S.N. [Formal Analysis of Structures of Nucleotide Chains]. *Matematicheskaya biologiya i bioinformatika - Mathematical biology and bioinformatics.* 2013. Issue 8. vol. 1. pp. 373–397. (In Russ.).
33. Freitas A., Afreixo V., Cruz S.E. Mixture models of geometric distributions in genomic analysis of inter-nucleotide distances. *Statistics, Optimization & Information Computing Stat.* 2013. vol. 1. no. 1. pp. 8–28.

34. Wasito I., Veritawati I. Fractal Dimension Approach for Clustering of DNA Sequences Based on Internucleotide Distance. IEEE 2013 International Conference of Information and Communication Technology (ICoICT). 2013. pp. 82–87.
35. Tavares A. et al. Detection of exceptional genomic words: a comparison between species. 22nd International Conference on Computational Statistics (COMPSTAT 2016). 2016.
36. Zhou L.Q., Li R., Han G.S. A Method Based on the Improved Inter-Nucleotide Distances of Genomes to Construct Vertebrates Phylogeny Tree. IEEE 2014 7th International Conference on Biomedical Engineering and Informatics. 2014. pp. 776–780.
37. Kolekar P., Kale M., Kulkarni-Kale U. Alignment-free distance measure based on return time distribution for sequence analysis: Applications to clustering, molecular phylogeny and subtyping. *Molecular Phylogenetics and Evolution.* 2012. vol. 65. no. 2. pp. 510–522.
38. Bonnici V., Manca V. Recurrence Distance Distributions in Computational Genomics. *American Journal of Bioinformatics and Computational Biology.* 2015. vol. 3. pp. 5–23.
39. Messaoudi I., Oueslati A.E., Lachiri Z. Wavelet analysis of frequency chaos game signal: a time-frequency signature of the C. elegans DNA. *EURASIP Journal on Bioinformatics and Systems Biology.* 2014. vol. 2014(1). pp. 16.
40. Orlov Y.K. *Chastotnye struktury konechnyh soobshcheniy v nekotoryh estestvennyh informacionnyh sistemah: dissertaciya* [Frequency structures of finite messages in some natural information systems. Ph.D. Thesis]. Tbilisi University. 1974. (In Russ.).
41. Orlov Y.K. [Invisible harmony]. Number and thought. 1980. vol. 3. pp. 70-105. (In Russ.).
42. Kudrin B.I. *Filosofiya tekhniki: osnovaniya postneklassicheskoy filosofii tekhniki.* [Philosophy of technology: the foundations of the post-non-classical philosophy of technology]. M.: Tehnika. 2007. vol. 36. 196 p. (In Russ.).
43. Popova O.V., Gelfand M.S. [Is there an analogue of the Zipf's law in genetic language?]. *Nauchno-tekhnicheskaya informaciya. Seriya 2: Informacionnye processy i sistemy - Scientific and technological information. Series 2: Information Processes and Systems.* 2000. vol. 4. pp. 19-24. (In Russ.).
44. Volchkova I.A., Gumenjuk A.S. [On measures of similarity of multilingual texts with the same content]. *Materialy Vserossiyskoy konferencii s mezhdunarodnym uchastiem «Znaniya – Ontologii – Teorii» (ZONT-13)* [Materials of the All-Russian Conference with International Participation «Knowledge – Ontologies – Theories» (KONT-13)] 2013. Issue 1. pp. 98-105. (In Russ.).
45. Gumenjuk A.S., Volchkova I.A. [Application of means of formal order analysis of a sign sequences for a formal assessment of the quality of translation]. *Omskiy nauchnyy vestnik. Seriya «Pribory, mashiny i tekhnologii» - The Journal Omsk Scientific Bulletin. Series «Devices, machines and technologies».* 2013. Issue 3(123). pp. 251-256. (In Russ.).
46. Shpynov S.N., Gumenuk A.S., Pozdnichenko N.N. [Application of the Numerical Characteristic of Formal Order Analysis of the Prokaryotic Genomes for Reclassification within the Genus Rickettsia]. *Matematicheskaya biologiya i bioinformatika - Mathematical biology and bioinformatics.* 2016. Issue 11. vol. 2. pp. 336-350. (In Russ.).
47. The DDBJ/ENA/GenBank Feature Table Definition. Available at: http://www.insdc.org/files/feature_table.html (accessed 15.04.2018).
48. Gumenuk A.S., Pozdnichenko N.N., Shpynov S.N. [Formal analysis of order in the local structure of the nucleotide sequences]. *Vestnik Tomskogo gosudarstvennogo universiteta - Tomsk State University Journal.* 2014. Issue 4(29). pp. 23–30. (In Russ.).
49. GENBANK DataBase. Available at: http://www.ncbi.nlm.nih.gov/nuccore/ (accessed 02.03.2018).
50. Gumenuk A.S., Pozdnichenko N.N., Skiba A.A., Shpynov S.N. [Computer Program «Matrix of similarity of nucleotide sequences by their components»]. *Svidetel'stvo o*

gosudarstvennoy registracii programmy dlya EHVM. №2017616679 [Certificate of State Registration of the Computer Program in the Register of Computer Programs № 2017616679]. 09.06.2017. (In Russ.).

51. Pozdnichenko N.N., Gumenuk A.S., Shpynov S.N. [On the cartographic representation of the set of prokaryotic genomes and their components by means of numerical characteristics of order]. *Novye informacionnye tekhnologii v issledovanii slozhnyh struktur: materialy 11-y mezhdunarodnoy konferencii* [Computer-aided technologies in applied mathematics: 11th international conference]. 2016. pp. 84-85. (In Russ.).

УДК 512.624.5 **DOI** 10.15622/sp.18.2.504-530

А.А. Молдовян, Н.А. Молдовян

НОВЫЕ ФОРМЫ СКРЫТОЙ ЗАДАЧИ ДИСКРЕТНОГО ЛОГАРИФИРОВАНИЯ

Молдовян А.А., Молдовян Н.А. **Новые формы скрытой задачи дискретного логарифмирования.**

Аннотация. Предлагаются новые варианты задачи дискретного логарифмирования в скрытой группе, которая представляет интерес для построения постквантовых криптографических протоколов и алгоритмов. Данная задача формулируется над конечными ассоциативными алгебрами с некоммутативной операцией умножения. В известном варианте указанная задача определяется как суперпозиция операций возведения в степень и автоморфного отображения алгебры, представляющей собой конечное некоммутативное кольцо с глобальной двухсторонней единицей, и называется конгруэнц логарифмированием. Ранее было показано, что последняя задача, заданная в конечной алгебре кватернионов, сводится к задаче дискретного логарифмирования в конечном поле, которое является расширением простого поля, над которым задана конечная алгебра кватернионов, и дальнейшие исследования задачи конгруэнц логарифмирования как примитива постквантовых криптосхем следует проводить в направлении поиска новых ее носителей, для которых такое сведение окажется вычислительно нереализуемым. Представлен ряд новых конечных ассоциативных алгебр, обладающих существенно различающимися свойствами в сравнении с алгеброй кватернионов, в частности в них отсутствует глобальная двухсторонняя единица. Это отличие потребовало новой формулировки задачи дискретного логарифмирования в скрытой группе, отличной от варианта конгруэнц логарифмирования. Предложено несколько вариантов такой формулировки, в которых используются локальные единицы различных типов. Рассматриваются левые, правые и двухсторонние локальные единицы, в качестве которых выступают обратимые и необратимые элементы алгебры. Предложены два общих способа построения конечных ассоциативных алгебр с некоммутативным умножением. Первый способ относится к заданию алгебр, имеющих произвольное натуральное значение размерности $m > 1$, второй — к заданию алгебр произвольных четных размерностей. Впервые разработаны алгоритмы цифровой подписи, основанные на вычислительной трудности задачи дискретного логарифмирования в скрытой группе.

Ключевые слова: криптография, шифры с открытым ключом, постквантовые криптосхемы, задача дискретного логарифмирования, конгруэнц логарифмирование, коммутативные шифры, открытое шифрование, цифровая подпись.

1. Введение. Для обеспечения информационной безопасности современных компьютерных технологий широкое практическое применение нашли криптографические алгоритмы и протоколы [1-2], в том числе двухключевые шифры (криптосхемы с открытым ключом), основанные на вычислительной трудности задачи факторизации чисел специального вида [3] и задачи дискретного логарифмирования (ЗДЛ) [4]. Приемлемый уровень стойкости криптосхем, основанных на этих задачах, определяется тем, что наиболее эффективные алгоритмы их решения, известные в настоящее время

и реализуемые с помощью существующей вычислительной техники, имеют субэкспоненциальную (задача факторизации и ЗДЛ в конечных полях) или экспоненциальную сложность (ЗДЛна эллиптической кривой специального вида).

Значительный прогресс в развитии квантовых вычислений[5,6] обусловил достаточно высокую степень актуальности вопроса оценки вычислительной сложности решения ЗДЛ и задачи факторизации на квантовом компьютере. Исследования, выполненные в данном направлении, показали, что обе рассматриваемые задачи имеют полиномиальную сложность в модели квантовых вычислений [7-9]. Данные результаты и прогнозируемое появления в ближайшее десятилетие практически действующих квантовых вычислителей[10], которые способны эффективно решать задачу взлома существующих криптографических алгоритмов и протоколов, основанных на ЗДЛ и задаче факторизации, обусловливают высокую степень актуальности проблемы создания арсенала протоколов электронной цифровой подписи (ЭЦП), открытого распределения ключей и открытого шифрования, которые были бы удобными для практического применения и стойкими к атакам с использованием квантовых компьютеров.

Алгоритмы симметричной криптографии (шифры с разделяемым секретным ключом), например, блочные и поточные шифры, по мнению специалистов, останутся стойкими к криптоанализу с использованием квантовых вычислителей. Однако для обеспечения достаточно высокой стойкости алгоритмов и протоколов криптографии с открытым ключом, в основу последних требуется положить вычислительно трудные задачи, обладающие сверхполиномиальной вычислительной сложностью при их решении с использованием как обычных, так и квантовых компьютеров. Создание практичных алгоритмов и протоколов постквантовой асимметричной (двухключевой) криптографии связан с поиском новых вычислительно трудных задач, пригодных для использования в качестве примитивов криптосхем с открытым ключом.

Откликом на такой вызов стали объявление Национальным институтом стандартов и технологий (НИСТ; National Institute of Standards and Technology, NIST) конкурса на разработку постквантовых криптосхем с открытым ключом[10] и появление регулярно проводимых тематических конференций по проблематике постквантовой криптографии [11].

Для построения постквантовых двухключевых криптосхем ранее было предложено использовать задачу поиска сопрягающего элемента в некоммутативных группах переплетения (braidgroups)[12,13],

называемых также группами кос. Эта идея привлекла внимание исследователей и была использована для построения алгоритмов открытого шифрования, протоколов открытого согласования секретного ключа и электронной цифровой подписи (ЭЦП). Однако в критических публикациях было показано, что в этом подходе имеются принципиальные трудности, связанные с тем, что задача поиска сопрягающего элемента сводится к решению систем линейных уравнений [14]. Последнее ставит под сомнение безопасность многочисленных двухключевых криптосхем, основанных на вычислительной сложности задачи поиска сопрягающего элемента в группах переплетения [15,16].

Более перспективным представляется подход к построению постквантовых криптосхем с открытым ключом, состоящий в комбинировании ЗДЛ с задачей поиска сопрягающего элемента, и приводящий к так называемой ЗДЛ в циклической группе, скрытой в конечной некоммутативной ассоциативной алгебре (КНАА) [17, 18]. Вычислительная сложность последней задачи (называемой также скрытой ЗДЛ) является сверхполиномиальной при ее решении на обычных вычислительных машинах. Однако в работах [17, 19, 20] были предложены полиномиальные алгоритмы сведения скрытой ЗДЛ, заданной над предложенными в [21-23] конечными алгебрами и представленной в известной на тот момент форме (названной конгруэнц логарифмированием), к ЗДЛ в конечном поле. В связи с этим была поставлена задача поиска новых носителей задачи конгруэнц логарифмирования (ЗКЛ) [17, 19], использование которых позволило бы устранить полиномиальную сводимость к ЗДЛ в конечном поле и обеспечить тем самым сверхполиномиальную сложность ЗКЛ при ее решении на квантовом компьютере, то есть потенциальную возможность разработки постквантовых криптосхем на основе ЗКЛ.

Задача построения алгоритмов и протоколов ЭЦП на основе скрытой ЗДЛ до настоящего момента не была решена. Это связано с тем, что ЗКЛ удобна для построения на ее базе протоколов открытого согласования ключа и алгоритмов открытого и коммутативного шифрования, но неочевидно как ее использовать для построения протокола цифровой подписи.

В настоящей работе решается задача поиска новых КНАА, заданных над простым конечным полем $GF(p)$ и обладающих существенно отличающимися свойствами от известных КНАА, и предлагаются новые варианты задания ЗДЛ в скрытой группе, существенно отличающиеся от ЗКЛ. Также описываются два унифицированных способа задания КНАА больших размерностей, которые позволяют построить два различных подкласса КНАА, включающих алгебры больших размерностей. Первый

способ позволяет построить КНАА произвольных размерностей $m \geq 1$, а второй — КНАА произвольных четных размерностей $m \geq 2$. На основе предложенных новых форм задания ЗДЛ в скрытой группе разработаны постквантовые алгоритмы ЭЦП.

2. Конечные некоммутативные ассоциативные алгебры. Рассмотрим конечное векторное пространство размерности m, заданное над простым полем $GF(p)$. Произвольный элемент этого пространства (вектор) V можно представить в виде упорядоченного набора m элементов конечного поля $GF(p)$: $V = (a, b,\ldots, q)$, а также в виде $V = a\mathbf{e} \oplus b\mathbf{i} \oplus\ldots\oplus q\mathbf{v}$, где $\mathbf{e}$, $\mathbf{i}$ и $\mathbf{v}$ — формальные базисные векторы. В последнем выражении слагаемые $a\mathbf{e}$, $b\mathbf{i}$ и $q\mathbf{v}$ обозначают однокомпонентные векторы $(a, 0, \ldots, 0)$, $(0, b, 0,\ldots, 0)$ и $(0, \ldots, 0, q)$, соответственно, и называются компонентами вектора V. Операция сложения векторов V и $V' = (a', b',\ldots, q')$, для которой примем обозначение $\oplus$, определяется как сложение всех одноименных координат:

$$V \oplus V' = \left(a, b,..., q\right) \oplus \left(a', b',..., q'\right) = \\ = \left(a + a', b + b,... q + q'\right),$$

где знак «+» обозначает операцию сложения в поле $GF(p)$.

Определим операцию умножения двух векторов $V = a\mathbf{e} \oplus b\mathbf{i} \oplus\ldots\oplus q\mathbf{v}$ и $X = x\mathbf{e} \oplus y\mathbf{i} \oplus\ldots\oplus w\mathbf{v}$ (обозначаемую знаком $\circ$) как перемножение каждой компоненты первого операнда с каждой компонентой второго операнда, то есть по следующей формуле:

$$V \circ X = \left(a\mathrm{e} \oplus b\mathrm{i} \oplus ... \oplus q\mathrm{v}\right) \circ \left(x\mathrm{e} \oplus y\mathrm{i} \oplus ... \oplus w\mathrm{v}\right) = \\ = ax\left(\mathrm{e} \circ \mathrm{e}\right) \oplus ay\left(\mathrm{e} \circ \mathrm{i}\right) \oplus ... \oplus aw\left(\mathrm{e} \circ \mathrm{v}\right) \oplus \\ \oplus bx\left(\mathrm{i} \circ \mathrm{e}\right) \oplus by\left(\mathrm{i} \circ \mathrm{i}\right) \oplus ... \oplus bw\left(\mathrm{i} \circ \mathrm{v}\right) \oplus ... \\ ... \oplus qx\left(\mathrm{v} \circ \mathrm{e}\right) \oplus qy\left(\mathrm{v} \circ \mathrm{i}\right) \oplus ... \oplus qw\left(\mathrm{v} \circ \mathrm{v}\right),$$

где координаты рассматриваемых однокомпонентных векторов перемножаются как элементы поля $GF(p)$, а произведение пары формальных базисных векторов в каждом слагаемом заменяется на некоторый однокомпонентный вектор, значение которого выбирается по так называемой таблице умножения формальных базисных векторов (ТУФБВ) [23, 24]. Координаты таких однокомпонентных векторов, отличные от единицы поля $GF(p)$, называются структурными коэффициентами. Если последний равен единице, то однокомпонентный вектор обозначается в виде соответствующего базисного вектора. После выполнения указанной за-

мены в правой части последнего выражения каждое слагаемое представляет собой однокомпонентный вектор. Сумма однокомпонентных векторов с одинаковым формальным базисным вектором равна некоторому другому однокомпонентному вектору с тем же базисным вектором. В общем случае это дает сумму m однокомпонентных векторов вида $a''\mathbf{e} \oplus b''\mathbf{i} \oplus \ldots \oplus q''\mathbf{v}$, то есть вектор $V'' = (a'', b'', \ldots, q'')$.

Определенная таким способом операция умножения пары m-мерных векторов является замкнутой в конечном множестве всевозможных векторов размерности m. Рассмотренное конечное векторное пространство с описанной операцией умножения называется конечной m-мерной алгеброй. Если операция умножения в конечной алгебре является ассоциативной и некоммутативной, то последняя называется КНАА.

Для фиксированных значений размерности конечной алгебры и характеристики поля $GF(p)$ путем разработки соответствующих конкретных ТУФБВ можно задать конечные алгебры различных типов, например, представляющие собой конечные расширенные поля $GF(p^m)$, коммутативные кольца [24, 25] и некоммутативные кольца [21-23] с глобальной двухсторонней единицей, алгебры без глобальной единицы [24]. В качестве носителей ЗДЛ в скрытой группе представляют интерес КНАА, которые в частных случаях могут представлять собой конечные некоммутативные кольца с глобальной двухсторонней единицей. Примером последних являются конечная алгебра кватернионов [24] и 8-мерная КНАА, предложенная в [21]. В литературе описано сравнительно малое число примеров КНАА, в связи с чем представляет интерес разработка унифицированных способов их построения для случаев различных размерностей.

В настоящей работе предлагается следующее обобщение ТУФБВ, использованной в [24] для построения 2-мерных КНАА. Для задания m-мерных алгебр при произвольном натуральном значении $m \geq 2$ может быть использована ТУФБВ общего вида, представленного как таблица 1, в которой формальные базисные векторы обозначены как $\mathbf{e}_i$, $i = 0, 1, \ldots, m-1$, а структурные коэффициенты — как μ_i. В данной таблице в каждой ячейке i-ой строке содержится формальный базисный вектор $\mathbf{e}_i$, а в каждой ячейке j-го столбца структурный коэффициент равен μ_j.Результат умножения $\mathbf{e}_i \circ \mathbf{e}_j$ указан в ячейке, расположенной на пересечении i-ой строки и j-го столбца таблицы. Благодаря такому устройству ТУФБВ произведение двух формальных базисных векторов определяется по следующей простой формуле:

$$\mathrm{e}_i \circ \mathrm{e}_j = \mu_j \mathrm{e}_i. \qquad (1)$$

Используя (1), легко показать, что при умножении произвольных трех формальных базисных векторов выполняется свойство ассоциативности:

$$\left\{\left(\mathrm{e}_i \circ \mathrm{e}_j\right) \circ \mathrm{e}_k = \mu_j \mu_k \circ \mathrm{e}_i;\ \ \mathrm{e}_i \circ \left(\mathrm{e}_j \circ \mathrm{e}_k\right) = \mu_j \mu_k \circ \mathrm{e}_i\right\} \Rightarrow$$
$$\Rightarrow \left(\mathrm{e}_i \circ \mathrm{e}_j\right) \circ \mathrm{e}_k = \mathrm{e}_i \circ \left(\mathrm{e}_j \circ \mathrm{e}_k\right).$$

С учетом определения операции умножения векторов из последней формулы следует, что она обладает свойством ассоциативности. Таким образом, таблица 1 задает общий способ построения КНАА произвольной размерности.

Таблица 1. Таблица умножения формальных базисных векторов, задающая некоммутативную ассоциативную операцию умножения векторов произвольной размерности ($\mu_i \in GF(p)$, $i = 0, 1, \ldots, m-1$)

∘	$\mathbf{e}_0$	$\mathbf{e}_1$	…	$\mathbf{e}_i$	…	$\mathbf{e}_j$	…	$\mathbf{e}_k$	…	$\mathbf{e}_{m-1}$
$\mathbf{e}_0$	$\mu_0\mathbf{e}_0$	$\mu_1\mathbf{e}_0$	…	$\mu_i\mathbf{e}_0$	…	$\mu_j\mathbf{e}_0$	…	$\mu_k\mathbf{e}_0$	…	$\mu_{m-1}\mathbf{e}_0$
$\mathbf{e}_1$	$\mu_0\mathbf{e}_1$	$\mu_1\mathbf{e}_1$	…	$\mu_i\mathbf{e}_1$	…	$\mu_j\mathbf{e}_1$	…	$\mu_k\mathbf{e}_1$	…	$\mu_{m-1}\mathbf{e}_1$
…	…	…	…	…	…	…	…	…	…	…
$\mathbf{e}_i$	$\mu_0\mathbf{e}_i$	$\mu_1\mathbf{e}_i$	…	$\mu_i\mathbf{e}_i$	…	$\mu_j\mathbf{e}_i$	…	$\mu_k\mathbf{e}_i$	…	$\mu_{m-1}\mathbf{e}_i$
…	…	…	…	…	…	…	…	…	…	…
$\mathbf{e}_j$	$\mu_0\mathbf{e}_j$	$\mu_1\mathbf{e}_j$	…	$\mu_i\mathbf{e}_j$	…	$\mu_j\mathbf{e}_j$	…	$\mu_k\mathbf{e}_j$	…	$\mu_{m-1}\mathbf{e}_j$
…	…	…	…	…	…	…	…	…	…	…
$\mathbf{e}_k$	$\mu_0\mathbf{e}_k$	$\mu_1\mathbf{e}_k$	…	$\mu_i\mathbf{e}_k$	…	$\mu_j\mathbf{e}_k$	…	$\mu_k\mathbf{e}_k$	…	$\mu_{m-1}\mathbf{e}_k$
…	…	…	…	…	…	…	…	…	…	…
$\mathbf{e}_{m-1}$	$\mu_0\mathbf{e}_{m-1}$	$\mu_1\mathbf{e}_{m-1}$	…	$\mu_i\mathbf{e}_{m-1}$	…	$\mu_j\mathbf{e}_{m-1}$	…	$\mu_k\mathbf{e}_{m-1}$	…	$\mu_{m-1}\mathbf{e}_{m-1}$

Другой предлагаемый унифицированный способ построения КНАА состоит в задании ТУФБВ, определяющей ассоциативную операцию умножения векторов произвольной четной размерности, по следующей ей формуле:

$$\mathrm{e}_i \circ \mathrm{e}_j = \begin{cases} \mathrm{e}_i, & \text{если } (i+j) \bmod 2 = 0; \\ \mathrm{e}_{m-1-i}, & \text{если } (i+j) \bmod 2 = 1. \end{cases} \qquad (2)$$

Формула (2) задает ассоциативную операцию умножения векторов для произвольного четного значения размерности. Действительно, рассмотрим три произвольных m-мерных вектора:

$$A = \sum_{i=0}^{m-1} a_i \mathrm{e}_i,\ B = \sum_{j=0}^{m-1} b_j \mathrm{e}_j \ \text{ и } \ C = \sum_{k=0}^{m-1} c_k \mathrm{e}_k.$$

В соответствии с определением операции умножения векторов получаем следующие соотношения:

$$\left(A \circ B\right) \circ C = \sum_{i,j,k=0}^{m-1} a_i b_j c_k \left(\mathrm{e}_i \circ \mathrm{e}_j\right) \circ \mathrm{e}_k;$$

$$A \circ \left(B \circ C\right) = \sum_{i,j,k=0}^{m-1} a_i b_j c_k \mathrm{e}_i \circ \left(\mathrm{e}_j \circ \mathrm{e}_k\right).$$

Легко показать, что при четном значении m из выражения (2) следует, что при всех возможных значениях тройки индексов (i, j, k)имеет место равенство $\left(\mathrm{e}_i \circ \mathrm{e}_j\right) \circ \mathrm{e}_k = \mathrm{e}_i \circ \left(\mathrm{e}_j \circ \mathrm{e}_k\right)$. Следовательно, выполняется соотношение $\left(A \circ B\right) \circ C = A \circ \left(B \circ C\right)$. Последнее означает, что операция умножения векторов, определенная по формуле (2), является ассоциативной.

Формула (2) описывает ТУФБВ для произвольного четного значения размерности m.В ячейках таблицы общего вида присутствуют только формальные базисные векторы, то есть однокомпонентные векторы с единичным значением координаты. После составления конкретной ТУФБВ для некоторого фиксированного четного значения размерности, можно перейти к этапу нахождения различных вариантов расстановки (распределения) структурных коэффициентов, при которых свойство ассоциативности операции умножения сохраняется. Таким способом составлена таблица 2 для случая задания 4-мерной алгебры, свойства которой описываются в следующем разделе. При использовании ТУФБВ, в которой распределение формальных базисных векторов по ячейкам таблицы зафиксировано, выбором различных распределений структурных коэффициентов и их значений можно задавать различные варианты операции умножения векторов, придающие существенно различные свойства некоммутативным алгебрам, которые заданы прификсированных значениях размерности и характеристики поля $GF(p)$.

В следующих разделах приводятся результаты изучения конкретных КНАА, причем существенное внимание уделено описанию единичных элементов следующих типов: локальных, глобальных, левосторонних, правосторонних и двухсторонних. Интерес к изучению единиц различного вида определяется тем, что их наличие и возможность выбора единичных элементов алгебры используется для задания новых форм скрытой ЗДЛ и построения постквантовых криптосхем.

Задание ЗДЛ в скрытой группе, содержащейся в некоторой КНАА, в качестве криптографического примитива предполагает нали-

чие большого числа циклических групп, имеющих достаточно большое значение порядка. При этом в криптосхемах, основанных на этой задаче, предполагается выполнение операции возведения векторов в целочисленную степень большой разрядности (от 160 до 512 бит в зависимости от задаваемого уровня стойкости). Для выполнения такой операции для больших значений степени используется алгоритм быстрого возведения в степень, основанный на процедуре последовательного возведения в квадрат.

При построении криптосхем с использованием различных форм скрытой ЗДЛ требуется реализовать модифицированную версию алгоритма быстрого возведения в степень, для которой нет необходимости задавать значение единичного вектора, поскольку его вид изменяется в зависимости от выбора КНАА конкретного вида, от типа используемых единичных элементов и от выбора циклических групп, в которых выполняются вычисления в рамках разрабатываемой криптосхемы.

Данная версия алгоритма быстрого возведения в степень играет значительную роль при выполнении экспериментальных исследований свойств разрабатываемых КНАА. С ее помощью возможно нахождение глобальных и локальных двухсторонних единичных элементов без предварительного вывода математических формул, описывающих координаты единичных элементов.

3. Задание 4-мерной КНАА с параметризуемым умножением. Рассмотрим КНАА размерности 4, которая является кольцом с параметризуемой операцией умножения, различные модификации которой являются взаимно ассоциативными. Умножение векторов задается по ТУФБВ, представленной в виде таблицы 2. Различные модификации операции умножения задаются различными наборами фиксируемых значений структурных коэффициентов μ и λ.

Пусть даны векторы V, W, U и две различные модификации операции умножения $\circ$ и $*$. Под взаимной ассоциативностью операций $\circ$ и $*$ понимается выполнимость следующего соотношения:

$$(V \circ W) * U = V \circ (W * U). \tag{3}$$

Используя определение операции умножения, легко показать, что для КНАА, заданной рассматриваемой ТУФБВ, любые две модификации умножения являются взаимно ассоциативными. Рассмотренные в литературе конечные алгебры не обладают данным свойством, которое представляет самостоятельный интерес для использования в криптографических алгоритмах.

Таблица 2. Строение ТУФБВ для задания КНАА, являющейся кольцом с параметризуемой операцией умножения ($\mu, \lambda \in GF(p)$; $\mu \neq \lambda$)

∘	**e**	**i**	**j**	**k**
e	**e**	μ**k**	μ**e**	**k**
i	**j**	λ**i**	λ**j**	**i**
j	**j**	μ**i**	μ**j**	**i**
k	**e**	λ**k**	λ**e**	**k**

При условии $\lambda \neq \mu$ умножение 4-мерных векторов, выполняемое по таблице 1, задает КНАА, являющуюся конечным кольцом с глобальной единицей, равной вектору:

$$E = \left(\frac{\lambda}{\lambda - \mu}, \frac{1}{\lambda - \mu}, \frac{1}{\mu - \lambda}, \frac{\mu}{\lambda - \mu} \right). \tag{4}$$

В этом кольце обратимы все векторы $V = (a, b, c, d)$, координаты которых удовлетворяют условию $dc - ab \neq 0$. Если имеет место $dc - ab = 0$, то вектор V необратим. Из последнего условия легко найти число необратимых векторов, которое равно $p^3 + p^2 - p$, и значение порядка некоммутативной мультипликативной группы кольца:

$$\Omega = p^4 - \left(p^3 + p^2 - p \right) = p\left(p - 1 \right)\left(p^2 - 1 \right).$$

Для множества необратимых векторов N можно ввести понятие локальных единиц. Если выполняется соотношение $E_{(} \circ N = N$, то вектор $E_{(}$ называется левой локальной единицей, а если $N \circ E_{)} = N$, то вектор $E_{)}$ называется правой локальной единицей для вектора N. В случае выполнимости соотношений $E_{(N)} \circ N = N$ и $N \circ E_{(N)} = N$ вектор $E_{(N)}$ называется двухсторонней локальной единицей для вектора N. Фиксированному необратимому вектору $N = (a, b, c, d)$ соответствует множество различных локальных единиц каждого вида. Множество левых локальных единиц описывается формулой:

$$E_{(} = \left(x, y, z, w \right) = \left(i, \frac{c}{a + \lambda c} - \frac{a + \mu c}{a + \lambda c} j,\ j, \frac{a}{a + \lambda c} - \frac{a + \mu c}{a + \lambda c} i \right), \tag{5}$$

где $i, j = 0, 1, 2, \ldots, p - 1$. В это множество входят необратимые и обратимые элементы рассматриваемого некоммутативного кольца, вклю-

чая глобальную единицу (4). Множество необратимых левых локальных единиц является подмножеством множества (5) и описывается формулой:

$$E'_{(} = (x, y, z, w) = \left(i, \frac{c}{a+\lambda c} - \frac{a+\mu c}{a+\lambda c} \cdot \frac{c}{a} i;\ \frac{c}{a} i;\ \frac{a}{a+\lambda c} - \frac{a+\mu c}{a+\lambda c} i \right), \tag{6}$$

где $i = 0, 1, 2, \ldots, p-1$.

Множество правых локальных единиц вектора N описывается формулой:

$$E_{)} = (x, y, z, w) = \left(\frac{c}{b+c} - \frac{\lambda b+\mu c}{b+c} k, h, k,\ \frac{b}{b+c} - \frac{\lambda b+\mu c}{b+c} h \right), \tag{7}$$

где $h, k = 0, 1, 2, \ldots, p-1$. Множество необратимых правых локальных единиц вектора N описывается формулой:

$$E'_{(} = (x, y, z, w) = \left(\frac{c}{b+c} - \frac{\lambda b+\mu c}{b+c} \cdot \frac{c}{b} h,\ h,\ \frac{c}{b} h,\ \frac{b}{b+c} - \frac{\lambda b+\mu c}{b+c} h \right), \tag{8}$$

где $h = 0, 1, 2, \ldots, p-1$.Множество двухсторонних локальных единиц представляет собой пересечение множеств (5) и (7) и описывается формулой:

$$E_{()} = (x, y, z, w) = \left(\frac{c}{b+c} - \frac{\lambda b+\mu c}{b+c} k,\ \ \frac{c}{a+\lambda c} - \frac{a+\mu c}{a+\lambda c} k, \right.$$
$$\left. k,\ \ \frac{ab+\mu c^2}{(b+c)(a+\lambda c)} - \frac{(a+\mu c)(\lambda b+\mu c)}{(b+c)(a+\lambda c)} k \right), \tag{9}$$

где $k = 0, 1, 2, \ldots, p-1$.В последнем множестве присутствует единственный необратимый вектор, содержащийся одновременно в множествах (6) и (8) и соответствующий значению:

$$k = k_0 = \frac{c^2}{ab + ac + \mu c^2 + \lambda bc} = \frac{c}{d + a + \mu c + \lambda b}. \tag{10}$$

Легко показать, что локальные единицы вектора N являются соответствующими локальными единицами для вектора N^uпри произвольном натуральном значении степени u. Учитывая конечность рассматриваемого кольца, можно показать, что при некотором значении

степени $u = \omega$ имеет место $N^{\omega} = E'_{(N)}$, где $E'_{(N)}$ двухсторонняя локальная единица вектора N, значение которой может быть вычислено по (9) при значении k_0, задаваемом формулой (10). Таким образом, всевозможные степени необратимого вектора N порождают циклическую группу порядка ω с единицей $E'_{(N)}$, что может быть использовано для задания ЗДЛ в скрытой подгруппе, отличной от ЗКЛ.

4. Задание ЗДЛ в скрытой группе необратимых векторов. Пусть в КНАА, рассмотренной в разделе 3, даны обратимый вектор Q и необратимый вектор N, такие, что выполняется неравенство $Q \circ N \neq N \circ Q$. Выберем некоторое произвольное достаточно большое натуральное число u и вычислим вектор F по формуле:

$$F = Q^{q-u} \circ E_{(N)}, \tag{11}$$

где $E_{(N)}$ — некоторый элемент множества (9), q — порядок вектора Q.

Зададим вычисление открытого ключа Y по формуле:

$$Y = Q^{u-t} \circ F \circ N^{x} \circ Q^{t}, \tag{12}$$

где пара случайно выбираемых чисел t и x являются личным секретным ключом владельца открытого ключа Y. Некоторый другой пользователь выбирает секретный ключ в виде пары случайных чисел t' и x' и вычисляет свой открытый ключ:

$$Y = Q^{u-t} \circ F \circ N^{x} \circ Q^{t}.$$

Первый и второй, обменявшись своими открытыми ключами, имеют возможность вычислить общее секретное значение в виде вектора Z по следующим двум формулам:

$$Z = Q^{u-t'} \circ Y^{x'} \circ Q^{t'}; \; Z = Q^{u-t} \circ Y'^{x} \circ Q^{t}. \tag{13}$$

То, что каждая из двух последних формул задает вычисление одного и того же значения, легко доказывается с учетом соотношения (11) и следующего достаточно очевидного равенства:

$$\left(Q^{u-t} \circ F \circ N \circ Q^{t}\right)^{x} = Q^{u-t} \circ F \circ N^{x} \circ Q^{t}.$$

Формулы (12) и (13) задают схему открытого согласования секретного ключа, стойкость которой определяется вычислительной сложность нахождения пары значений t и x по известным параметрам

Y, *Q*, *N*, *F* и *u*. При известном значении *t*нахождение *x* составит ЗДЛ в группе, генерируемой элементом:

$$G = Q^{u-t} \circ F \circ N \circ Q^t.$$

Вычисление двух неизвестных значений *t* и *x*из (12) определяет ЗДЛ в скрытой группе. Последняя задача имеет существенные отличия от ЗКЛ, которая задается формулой:

$$Y = Q^{q-t} \circ G^x \circ Q^t, \tag{14}$$

где *Q* и *G* — пара неперестановочных обратимых элемента конечного некоммутативного кольца; *t*и *x* — неизвестные натуральные значения.

Сравнение показывает, что ЗДЛ в скрытой группе, задаваемая по формуле (12) является более гибкой в плане большего числа задаваемых параметров.

Специфичным для формулы (12) является возможность произвольного выбора степени *u*и локальной двухсторонней единицы, который определяет значение вектора *F*, вычисляемого по формуле (11). Следует заметить, что криптосхема, задаваемая формулами (11), (12) и (13), работает корректно, если в (11) вместо двухсторонней локальной единицы $E_{(N)}$ взять произвольную левую или правую локальную единицу из множеств (5) или (7) соответственно.

Симметричным по отношению к рассмотренному является вариант задания ЗДЛ в скрытой группе по следующим двум формулам:

$$F = E_{(;)} \circ Q^{q-u}, \tag{11′}$$

где $E_{(;)}$ — некоторая локальная единица (левая, правая или двусторонняя); при этом в качестве $E_{(;)}$ можно брать как обратимый, так и необратимый элемент относительно глобальной единицы (4); и

$$Y = Q^{u-t} \circ N^x \circ F \circ Q^t. \tag{12′}$$

Самостоятельное значение имеет способ задания ЗДЛ в скрытой группе по следующим двум формулам:

$$F_1 = Q^{q-u} \circ F_2^{-1} \circ E_{(;)}, \tag{15}$$

где F_2 — произвольно выбираемый обратимый элемент конечного некоммутативного кольца;

$$Y = Q^{u-t} \circ F_1 \circ N^x \circ F_2 \circ Q^t. \tag{16}$$

Используя формулу (15), легко показать, что справедливо следующее соотношение:

$$Q^{u-t} \circ F_1 \circ N^x \circ F_2 \circ Q^t = \left(Q^{u-t} \circ F_1 \circ N \circ F_2 \circ Q^t\right)^x. \quad (17)$$

Пара элементов F_1 и F_2 может быть получена также выбором в качестве F_1 произвольного обратимого элемента и последующим вычислением значения F_2 по формуле:

$$F_2 = E_{(;)} \circ F_1^{-1} \circ Q^{q-u}.$$

В криптосхемах, задаваемых парой формул (11′) и (12′) и парой формул (15) и (16), согласование общего секретного ключа Z по открытому каналу выполняется также по формулам (13).

5. Задание ЗДЛ в скрытой группе алгебры без единицы. В разделе 4 предложены новые варианты задания ЗДЛ в скрытой группе КНАА, являющейся конечным некоммутативным кольцом. В представленных вариантах используется наличие глобальной единицы, относительно которой рассматривается обратимость (и необратимость) элементов КНАА, например, алгебры, представленной в разделе 3.

Для КНАА, не содержащих глобальной двухсторонней единицы, указанные способы задания ЗДЛ в скрытой группе не могут быть применены. Примерами КНАА последнего типа являются ассоциативные алгебры с операцией умножения векторов, задаваемой по ТУФБВ, представленной в таблице 1, для случаев размерности векторного пространства $m = 2$ [24]и$m = 3, 4, 5$.Можно предположить, что для случая произвольной размерности КНАА с операцией умножения определяемой по таблице 1 глобальная двухсторонняя единица отсутствует. Простое строение данной ТУФБВ позволяет предположить, что могут быть получены общие формулы для произвольного значения размерности m, описывающие единичные элементы и делители нуля в КНАА с операцией умножения, заданной по таблице 1, однако это представляется самостоятельной задачей.

Рассмотрение КНАА без глобальной двусторонней единицы в качестве носителя ЗДЛ в скрытой группе предполагает использование другого подхода, в котором не требуется применение обратимых элементов. Потенциальная возможность применения таких КНАА в качестве носителей ЗДЛ в скрытой группе связана с существованием двухсторонних локальных единиц, с которыми связаны подмножества элементов алгебры, которые образуют циклические группы. Например, в

трехмерной алгебре, заданной таблицей 1 для случая значений структурных коэффициентов $\mu_1 = \mu_2 = \mu_3 = 1$ множество правых локальных единиц для элемента $N = (a, b, c)$, координаты которого удовлетворяют условию $a + b + c \neq 0$, описывается формулой:

$$E_{)} = (x, y, z) = (i, j, 1 - i - j),$$

где $i, j = 0, 1, 2, \ldots, p - 1$. Это множество правых единиц является глобальным в том смысле, что входящие в него правые единицы действуют как таковые на всевозможные элементы N, удовлетворяющие указанному условию. Элементу $N = (a, b, c)$ соответствует единственная левая локальная единица, которая совпадает с единственной двухсторонней локальной единицей:

$$E_{(} = E_{()} = (x, y, z) = \left(\frac{a}{a+b+c}, \frac{b}{a+b+c}, \frac{c}{a+b+c} \right). \tag{18}$$

Всевозможные степени элемента N образуют циклическую группу с единицей (18).

Другим примером является 4-мерная КНАА с операцией умножения, определяемой по таблице 3. В данной алгебре существуют локальные единицы только для векторов$N = (a, b, c, d)$, координаты которых удовлетворяют соотношению $ac = bd$. Множество правых локальных единиц вектора N описывается формулой:

$$E_{)} = (x, y, z, w) = \left(i, j, \frac{b}{\mu a + \lambda b} - j, \frac{a}{\mu a + \lambda b} - i \right), \tag{19}$$

где $i, j = 0, 1, 2, \ldots, p - 1$, а множество левых локальных единиц — формулой:

$$E_{(} = (x, y, z, w) = \left(k, \ \frac{a}{\lambda(a+d)} - \frac{\mu}{\lambda} k, \ h, \ \frac{d}{\mu(a+d)} - \frac{\lambda}{\mu} h \right), \tag{20}$$

где $k, h = 0, 1, 2, \ldots, p - 1$.

Множество локальных двухсторонних единиц для Nопределяется пересечением множеств (19) и (20), что дает следующую формулу:

$$E_{(N)} = \left(k, \frac{a}{\lambda(a+d)} - \frac{\mu}{\lambda} k, \frac{b}{\mu a + \lambda b} - \frac{a}{\lambda(a+d)} + \frac{\mu}{\lambda} k, \frac{a}{\mu a + \lambda b} - k \right), \tag{21}$$

где $k = 0, 1, 2, \ldots, p - 1$.

Таблица 3. Правило умножения формальных базисных векторов в 4-мерной КНАА без глобальной единицы (μ, λ ∈ *GF*(*p*); μ ≠ λ)

∘	**:e**	**:i**	**:j**	**:k**
:e	**:μe**	**:μi**	**:μi**	**:μe**
:i	**:λe**	**λi**	**λi**	**λe**
:j	**:λk**	**λj**	**λj**	**λk**
:k	**:μk**	**μj**	**μj**	**μk**

В множестве (21) содержится только один элемент $E_{(N)} = (a', b', c', d')$, координаты которого удовлетворяют условию $a'c' = b'd'$. Этот элемент является единицей циклической группы, которую порождают всевозможные степени *N*, и его координаты могут быть вычислены по формуле (21) при значении:

$$k = k_0 = \frac{a^2}{(a+d)(\mu a + \lambda b)}. \qquad (21')$$

Таким образом, вычисление двухсторонней локальной единицы $E_{(N)}$ может быть выполнено по формуле (21) или путем возведения элемента *N* в степень, кратную порядку циклической группы, генерируемой всевозможными степенями *N*. Это может быть проиллюстрировано следующим вычислительным экспериментом, выполненным для 4-мерного вектора:

$$N = (908829491, 124888084499, 18949746053, 676148046414381)$$

при значениях p = 1108878614179151; μ = 257; λ = 13:

$$N^{p-1} = (1016120000861220,\ 58254033235670,\ 20300751201639,\ 697158116826006).$$

Вычисление по формуле (21) дает значение k_0 = 1016120000861220, подстановка которого в (21) дает следующее:

$$E_{(N)} = N^{p-1}.$$

Элемент $E_{(N)} = (a', b', c', d')$ является правой единицей подмножества КНАА, которое описывается формулой:

$$V_N = V \circ E_{(N)}, \qquad (22)$$

где V пробегает все элементы КНАА. Элемент $E_{(N)}$ является левой единицей подмножества КНАА, которое описывается формулой:

$$V'_N = E_{(N)} \circ V.$$

Ни в одном из двух последних подмножеств КНАА элемент $E_{(N)}$не является двухсторонней единицей для всех элементов подмножества. Каждое из этих подмножеств замкнуто относительно операции умножения. В подмножестве (22) можно задать автоморфизм относительно операции умножения, описываемый формулой:

$$\psi_{N,t}(V_N) = N^{\eta-t} \circ V_N \circ N^t, \tag{23}$$

где η — порядок циклической группы, генерируемой степенями элемента N. Действительно, для произвольных двух элементов V_{N1} и V_{N2} множества (22) выполняются соотношения:

$$\begin{aligned}\psi_{N,t}\left(V_{N1} \circ V_{N2}\right) &= N^{\eta-t} \circ (V_{N1} \circ V_{N2}) \circ N^t = \\ &= N^{\eta-t} \circ \left(V_{N1} \circ E_{(N)} \circ V_{N2}\right) \circ N^t = \\ &= \left(N^{\eta-t} \circ V_{N1} \circ N^t\right) \circ \left(N^{\eta-t} \circ V_{N2} \circ N^t\right) = \\ &= \psi_{N,t}\left(V_{N1}\right) \circ \psi_{N,t}\left(V_{N2}\right).\end{aligned}$$

В силу указанного автоморфизма имеет место следующая формула:

$$\left(N^{\eta-t} \circ V_N \circ N^t\right)^x = N^{\eta-t} \circ V_N^x \circ N^t, \tag{24}$$

которая может быть использована для построения криптосхемы открытого согласования общего секретного ключа, в которой формирование открытого ключа Yпо секретному ключу (x, t) выполняется по формуле:

$$Y = N^{\eta-t} \circ V_N^x \circ N^t, \tag{25}$$

где N, $V_{(N)}$ и η — параметры криптосхемы, а вычисление общего секретного ключа — по формулам:

$$Z = N^{\eta-t} \circ Y'^x \circ N^t, Z = N^{\eta-t'} \circ Y^{x'} \circ N^{t'}. \tag{26}$$

Формула (25) определяет еще один вариант задания ЗДЛ в скрытой группе, который отличается от ЗКЛ тем, что его носителем является КНАА без глобальной единицы. При выборе параметров N и $V_{(N)}$следует выполнить естественное требование, состоящее в выполнении неравенства $N \circ V_{(N)} \neq V_{(N)} \circ N$.

Для рассматриваемой в разделе 5 КНАА был сформулирован ряд положений, касающихся векторов вида $N = (a, b, c, d)$,где $ac = bd$. Число таких векторов равно $p^3 + p^2 - p$. Возникает естественный вопрос об остальных $p^4 - (p^3 + p^2 - p)$ векторов вида $V = (a', b', c', d')$, где $a'c' \neq b'd'$. Данная 4-мерная КНАА является весьма своеобразной и векторы последнего вида при умножении «самоустраняются», то есть умножение двух векторов произвольного вида дает в результате вектор первого типа. Можно сказать, что операция умножения обладает сжимающим свойством.

6. Постквантовые алгоритмы цифровой подписи. Интерес к использованию ЗДЛ в скрытой группе для построения схем электронной цифровой подписи (ЭЦП) связан с актуальностью разработки алгоритмов ЭЦП, стойких к атакам с использованием гипотетического квантового компьютера. Однако для выполнения такой разработки требуется использование новых форм задания указанной вычислительно трудной задачи. В данном разделе предлагаются новые формы задания ЗДЛ в скрытой группе, позволяющие реализовать построение алгоритмов ЭЦП с использованием вычислений в КНАА.

При задании скрытой ЗДЛ над алгеброй с глобальной двухсторонней единицей существенным моментом является использование необратимых элементов в качестве генератора скрытой циклической группы. Для необратимых векторов существует большое число локальных единиц различного типа, например, описываемых формулами (5) и (7) в случае 4-мерной КНАА из раздела 3, задаваемой с помощью таблицы 2. Зададим формирование личного секретного ключа подписанта в соответствии со следующей процедурой:

1. Выбрать случайный необратимый вектор N, локальный порядок которого равен достаточно большому простому числу η.

2. Выбрать случайные векторы E_1, E_2 и E_3из множества локальных единиц, соответствующих вектору N(в качестве E_1 и E_2 выбираются обратимые векторы).

3. Сгенерировать обратимый вектор G, такой, что имеет место неравенство $N \circ G \neq G \circ N$, и вычислить следующие векторы:

$$U = E_1 \circ G^{-1};\ H = U^{-1} \circ E_2; T = E_3 \circ H^{-1}. \qquad (27)$$

4. Сгенерировать случайное число $x < \eta$.

Следует отметить, что вычисляемый на шаге 3 вектор U является обратимым элементом как произведение двух обратимых элементов, поэтому существует обратное к нему значение U^{-1}, используемое при вычислении вектора H. Последний также обратим, и существует обратное к нему значение H^{-1}, которое используется при вычислении вектора T. Нахождение обратных значений выполняется путем решения соответствующих систем из четырех линейных уравнений.

Личный секретный ключ подписанта представляет собой целое число x и тройку векторов N, G и T.Открытый ключ подписанта представляет собой пару векторов Y и Q, вычисляемых по следующим двум формулам:

$$Y = G \circ N^x \circ U; Q = H \circ N \circ T. \tag{28}$$

Вычисляемые векторы Y и Qпринадлежат различным циклическим группам, содержащимся в используемой КНАА, но связаны они с одной и той же замаскированной циклической группой, генерируемой всевозможными степенями необратимого вектора N. Скрытая ЗДЛ состоит в вычислении значения x по известным 4-мерным векторам Y и Q, которые принадлежат разным циклическим группам, содержащимся в используемой КНАА.

Процедура генерации ЭЦП к некоторому заданному электронному документу M выполняется следующим образом:

1. Сгенерировать случайное число $k < \eta$.

2. Вычислить вектор $W = G \circ N^k \circ T$.

3. Вычислить первый элемент ЭЦП в виде двоичного числа $e = F_h(M, W)$, где F_h — некоторая специфицированная хэш-функция.

4. Вычислить второй элемент ЭЦП в виде двоичного числа s:

$$s = k - ex \bmod \eta.$$

Процедура проверки подлинности ЭЦП (e, s) к документу M выполняется по открытому ключу (Y, Q) следующим образом:

1. Вычислить вектор:

$$\tilde{W} = Y^e \circ Q^s.$$

2. Вычислить двоичное число $\tilde{e} = F_h\left(M, \tilde{W}\right)$.

3. Сравнить значения $\tilde{e}$ и e. Если $\tilde{e} = e$, то подпись (e, s) является подлинной. В противном случае подпись отклоняется.

Вычисление значения подписи выполняется с учетом операций возведения в степень в циклической группе, задаваемой вектором N, а проверка подлинности цифровой подписи осуществляется с помощью операций возведения в степень в двух других циклических группах, а именно в конечных группах, порождаемых векторами Y и Q, представляющих собой элементы открытого ключа. В связи с этим корректность предложенной схемы подписи не является очевидной, что делает целесообразным рассмотрение формального доказательства ее корректности. Последнее выполняется путем подстановки значения ЭЦП (e, s) на выходе процедуры генерации подписи на вход процедуры проверки подлинности ЭЦП. С учетом формул (28) и справедливости соотношений $U \circ G = E_1$, $U \circ H = E_2$ и $T \circ H = E_3$, вытекающих из (27), такая подстановка дает следующее:

$$\tilde{W} = Y^e \circ Q^s = \left(G \circ N^x \circ U\right)^e \circ \left(H \circ N \circ T\right)^s =$$

$$= G \circ \left(N^x \circ E_1\right)^{e-1} \circ N^x \circ U \circ H \circ \left(N \circ E_3\right)^{s-1} \circ N \circ T =$$

$$= G \circ \left(N^x\right)^{e-1} \circ N^x \circ E_2 \circ \left(N\right)^{s-1} \circ N \circ T = G \circ N^{ex} \circ E_2 \circ N^s \circ T =$$

$$= G \circ N^{ex+s} \circ T = G \circ N^{ex+(k-ex)} \circ T = G \circ N^k \circ T =$$

$$= W \;\Rightarrow\; \tilde{e} = F_h\left(M, \tilde{W}\right) = F_h\left(M, W\right) = e\,.$$

В представленной схеме ЭЦП процедура проверки подписи состоит в выполнении операций над векторами, принадлежащими двум различным циклическим группам КНАА. При этом при вычислении подписи используется связь элементов Y и Q открытого ключа с одной и той же циклической группой, которая скрыта для всех, кроме владельца открытого ключа. В предположении, что потенциальному атакующему известны значения G, U, H и T (это дает возможность вычислить значения N и N^x), для подделки подписи ему потребуется найти число x, то есть решить задачу дискретного логарифмирования в циклической группе, генерируемой вектором N. При использовании гипотетического квантового вычислительного устройства последняя задача может быть решена за полиномиальное время. Однако, атакующему известны только векторы Y и Q, поэтому возможность применения квантового компьютера для взлома предложенной схемы ЭЦП связана со сведением используемой ЗДЛ в скрытой группе к ЗДЛ в явно заданной циклической группе. В настоящее время этот вопрос является мало изученным и потребуется выполнение дополнительных исследований со стороны независимых исследователей, чтобы получить достаточную экспертную

оценку стойкости предложенной схемы ЭЦП к атакам с использованием квантового вычислительного устройства.

Использование необратимого вектора N в описанной схеме ЭЦП связано с предотвращением потенциальных атак, основанных на гомоморфном отображении КНАА в поле $GF(p)$,которые рассмотрены в [21].

Другая новая форма задания скрытой ЗДЛ может быть сформулирована для случая использования КНАА без глобальной двухсторонней единицы (см. раздел 5) следующим образом:

1. Выбрать случайный вектор $N = (a, b, c, d)$ большого простого порядка η, удовлетворяющий условию $ac = bd$.

2. Сформировать случайные векторыU,G, T, H и L, такие, что выполняются условия $U \circ G = E_1$, $T \circ H = E_2$ и $U \circ L \circ H = E_3$, где E_1, E_2 и E_3 — локальные единицы произвольных типов.

3. Сгенерировать случайное число $x < \eta$.

Личный секретный ключ подписанта представляет собой целое число x и тройку векторов N, G, и T.Открытый ключ подписанта представляет собой тройку векторов Y, Qи L,в которой первые два вычисляются по формулам (28).Постквантовая схема ЭЦП описывается следующими двумя алгоритмами.

Алгоритм генерации ЭЦП к некоторому заданному электронному документу M:

1. Сгенерировать случайное число $k < \eta$.

2. Вычислить вектор $W = G \circ N^k \circ T$.

3. Вычислить первый e и второй элементы ЭЦП по формулам $e = F_h(M, W)$ и $s = k - ex \bmod \eta$.

Алгоритм проверки подлинности ЭЦП (e, s) к документу M:

1.Вычислить вектор $\tilde{W} = Y^e \circ L \circ Q^s$.

2. Вычислить значение $\tilde{e} = F_h(M, \tilde{W})$.

3. Если $\tilde{e} = e$, то подпись (e, s) признается подлинной. В противном случае подпись отклоняется.

Доказательство корректности работы второй схемы ЭЦП выполняется следующим образом:

$$\tilde{W} = Y^e \circ L \circ Q^s = \left(G \circ N^x \circ U\right)^e \circ L \circ \left(H \circ N \circ T\right)^s =$$

$$= G \circ N^{ex} \circ U \circ L \circ H \circ N^s \circ T = G \circ N^{ex} \circ E_3 \circ N^s \circ T =$$

$$= G \circ N^{ex+s} \circ T = G \circ N^{ex+(k-ex)} \circ T = G \circ N^k \circ T =$$

$$= W \quad \Rightarrow \quad \tilde{e} = e\,.$$

Предложенные формы задания ЗДЛ в скрытой группе и алгоритмы ЭЦП на их основе заслуживают внимания криптографов, поскольку при подтверждении их стойкости к квантовым атакам устраняются недостатки (большой размер подписи и открытого ключа, ограниченное число документов, которые могут быть подписаны при регистрации одного открытого ключа), присущие постквантовым схемам ЭЦП, предложенным в рамках конкурса НИСТ по разработке постквантовых криптосхем с открытым ключом.

Представляет интерес рассмотрение возможности разработки на основе предложенных схем ЭЦП протоколов слепой цифровой подписи по аналогии с протоколами слепой подписи, основанными на ЗДЛ и описанными в работах[24, 25].

7. Заключение. В данной статье предложены две новые 4-мерные КНАА и общий способ построения КНАА произвольной размерности $m \geq 2$. Одна из предложенных алгебр является кольцом с глобальной двухсторонней единицей и представляет интерес в качестве нового носителя ЗКЛ. Для этой алгебры предложены два новых варианта задания ЗДЛ в скрытой группе, отличные от ЗКЛ. Вторая четырехмерная алгебра не содержит глобальной двухсторонней единицы и для нее предложен третий вариант ЗДЛ в скрытой группе. Другие две новые формы задания скрытой ЗДЛ предложены и использованы для построения постквантовых алгоритмов ЭЦП.

Алгоритмы ЭЦП, основанные на вычислительной трудности скрытой ЗДЛ, разработаны впервые. Предложенные КНАА и новые формы задания скрытой ЗДЛ представляют существенный интерес для разработки протоколов открытого согласования общего секретного ключа, коммутативного шифрования и ЭЦП, стойких к атакам с использованием квантовых вычислителей.

Рассмотренные ТУФБВ могут быть применены также и для задания КНАА над конечными полями, отличными от $GF(p)$, в частности над полями $GF(2^s)$. При этом в последнем случае самостоятельный интерес представляет использование в качестве степени расширения двоичного поля s простого числа, равного степени Мерсенна, за счет чего можно добиться формирования конечных циклических групп, (являющихся подмножествами задаваемых КНАА), которые обладают простым значением порядка. Для данного варианта построения КНАА применимы все предложенные варианты задания ЗДЛ в скрытой подгруппе.

Скрытая ЗДЛ представляется перспективной в качестве кандидата на универсальный постквантовый криптографический примитив, на основе которого могут быть разработаны постквантовые двухключевые алгоритмы и протоколы различного типа.

Литература

1. *Sirwan A., Majeed N.* New Algorithm for Wireless Network Communication Security // International Journal on Cryptography and Information Security. 2016. vol. 6. no. 3/4. pp. 1–8.
2. *Feng Y., Yang G., Liu. J.K.* A new public remote integrity checking scheme with user and data privacy // International Journal of Applied Cryptography. 2017. vol. 3. no 3. pp. 196–209.
3. *Chiou S.Y.* Novel Digital Signature Schemes based on Factoring and Discrete Logarithms // International Journal of Security and Its Applications. 2016. vol. 10. no. 3. pp. 295–310.
4. *Poulakis D.* A variant of Digital Signature Algorithm // Designs, Codes and Cryptography. 2009. vol. 51. no. 1. pp. 99–104.
5. *Yan S.Y.* Quantum Computational Number Theory // Springer. 2015. 252 p.
6. *Yan S.Y.* Quantum Attacks on Public-Key Cryptosystems // Springer. 2014. 207 p.
7. *Shor P.W.* Polynomial-time algorithms for prime factorization and discrete logarithms on quantum computer // SIAM Journal of Computing. 1997. vol. 26. pp. 1484–1509.
8. *Smolin J.A., Smith G., Vargo A.* Oversimplifying quantum factoring // Nature. 2013. vol. 499. no. 7457. pp. 163–165.
9. *Hamdi S.M., Zuhori S.T., Ffiroz M., Biprodip P.* A Compare between Shor's quantum factoring algorithm and General Number Field Sieve // 2014 International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT). 2014. pp. 1–6.
10. Federal Register. The Daily Journal of the United States Goverment URL: https://www.gpo.gov/fdsys/pkg/FR-2016-12-20/pdf/2016-30615.pdf (дата обращения: 06.03.2018).
11. *Verma G.K.* A Proxy Blind Signature Scheme over Braid Groups // International Journal of Network Security. 2009. vol. 9. no 3. pp. 214–217.
12. *Hiranvanichakorn P.* Provably Authenticated Group Key Agreement based on Braid Groups − The Dynamic Case // International Journal of Network Security. 2017. vol. 19. no. 4. pp. 517−527.
13. *Myasnikov A., Shpilrain V., Ushakov A.* A Practical Attack on a Braid Group Based Cryptographic Protocol // Annual International Cryptology Conference. 2005. vol. 3621. pp. 86−96.
14. *Chaturvedi A., Lal S.* An Authenticated Key Agreement Protocol Using Conjugacy Problem in Braid Groups // International Journal of Network Security. 2008. vol. 6. no. 2. pp. 181–184.
15. *Verma G.K.* Probable Security Proof of a Blind Signature Scheme over Braid Groups // International Journal of Network Security. 2011. vol. 12. no. 2. pp. 118−120.
16. *Kuzmin A.S. et al.* Cryptographic Algorithms on Groups and Algebras // Journal of Mathematical Sciences. 2017. vol. 223. no. 5. pp. 629−641.
17. *Moldovyan D.N., Moldovyan N.A., Shcherbacov V.A.* Non-commutative finite associative algebras of 3-dimensional vectors // Quasigroups and related systems. 2018. vol. 26. no. 1. pp. 109−120.
18. *Кузьмин А.С. и др.* Криптографические алгоритмы на группах и алгебрах // Фундаментальная и прикладная математика. 2015. Т. 20. № 1. С. 205−222.
19. *Глухов М.М.* К анализу некоторых систем открытого распределения ключей, основанных на неабелевых группах // Математические вопросы криптографии. 2010. Т. 1. № 4. С. 5−22.
20. *Moldovyan D.N., Moldovyan N.A.* Cryptoschemes over hidden conjugacy search problem and attacks using homomorphisms // Quasigroups and Related Systems. 2010. vol. 18. pp. 177−186.

21. *Moldovyan D.N., Moldovyan N.A.* A New Hard Problem over Non-Commutative Finite Groups for Cryptographic Protocols // International Conference on Mathematical Methods, Models, and Architectures for Computer Network Security. 2010. vol. 6258. pp. 183−194.
22. *Moldovyan D.N.* Non-Commutative Finite Groups as Primitive of Public-Key Cryptoschemes // Quasigroups and Related Systems. 2010. vol. 18. pp. 165−176.
23. *Moldovyan A.A., Moldovyan N.A., Shcherbacov. V.A.* Non-commutative finite associative algebras of 2-dimension vectors // Computer Science Journal of Moldova. 2017. vol. 25. no. 3(75). pp. 344−356.
24. *Camenisch J.L., Piveteau J.-M., Stadler M.A.* Blind Signatures Based on the Discrete Logarithm Problem //Workshop on the Theory and Application of Cryptographic Techniques. 1994. pp. 428–432.
25. *Pointcheval D., Stern J.* Security Arguments for Digital Signatures and Blind Signatures // Journal of Cryptology. 2000. vol. 13. no. 3. pp. 361−396.

Молдовян Александр Андреевич — д-р техн. наук, профессор, главный научный сотрудник, лаборатория кибербезопасности и постквантовых криптосистем, Федеральное государственное бюджетное учреждение науки Санкт-Петербургского института информатики и автоматизации Российской академии наук (СПИИРАН). Область научных интересов: компьютерная безопасность, криптография, безопасность компьютерных сетей, управление политиками безопасности, разграничение доступа, аутентификация, анализ защищенности, обнаружение компьютерных атак, межсетевые экраны, защита от вирусов и сетевых червей, анализ и верификация протоколов безопасности и систем защиты информации, защита программного обеспечения от взлома. Число научных публикаций — 200. maa1305@yandex.ru; 14-я линия В.О., 39, 199178, Санкт-Петербург, Российская Федерация; р.т.: +7(812)328–5185.

Молдовян Николай Андреевич — д-р техн. наук, профессор, главный научный сотрудник, лаборатория кибербезопасности и постквантовых криптосистем, Федеральное государственное бюджетное учреждение науки Санкт-Петербургского института информатики и автоматизации Российской академии наук (СПИИРАН). Область научных интересов: компьютерная безопасность, криптография, симметричные и асимметричные криптосистемы, электронная цифровая подпись, аутентификация, блочные шифры, псевдовероятностные шифры. Число научных публикаций — 250. nmold@mail.ru; 39, 14-я линия В.О., 199178, Санкт-Петербург, Российская Федерация; р.т.: +7(812)328–5185.

Поддержка исследований. Работа выполнена при частичной финансовой поддержке РФФИ (проект № 18-07-00932-а).

DOI 10.15622/sp.18.2.504-530

A.A. MOLDOVYAN, N.A. MOLDOVYAN

NEW FORMS OF DEFINING THE HIDDEN DISCRETE LOGARITHM PROBLEM

Moldovyan A.A., Moldovyan N.A. **New Forms of Defining the Hidden Discrete Logarithm Problem.**

Abstract. Novel variants of defining the discrete logarithm problem in a hidden group, which represents interest for constructing post-quantum cryptographic protocols and algorithms, are proposed. This problem is formulated over finite associative algebras with non-commutative multiplication operation. In the known variant this problem, called congruent logarithm, is formulated as superposition of exponentiation operation and automorphic mapping of the algebra that is a finite non-commutative ring. As it has been shown before, congruent logarithm problem defined in the finite quaternion algebra can be reduced to discrete logarithm in the finite field that is an extension of the field over which the quaternion algebra is defined. Therefore further reseaches of the congruent logarithm problem as primitive of the post-quantum cryptoschemes should be carried out in direction of finding new carriers. This paper presents novel associative algebras possessing significantly different properties than quaternion algebra, in particular they contain no global unit. This difference demanded a new definition of the discrete logarithm problem in a hidden group, which is different from the congruent logarithm. Several variants of such definition, in which the notion of the local unite is used, are proposed. Right, left, and bi-side local unites are considered. Two general methods for constructing the finite associative algebras with non-commutative multiplication operation are proposed. The first method relates to defining the algebras having dimension value equal to a natural number $m > 1$, and the second one relates to defining the algebras having arbitrary even dimensions. For the first time, the digital signature algorithms based on computational difficulty of the discrete logarithm problem in a hidden group have been proposed.

Keywords: Cryptography, Public-Key Ciphers, Post-Quantum Cryptoschemes, Discrete Logarithm Problem, Congruence Logarithm, Commutative Ciphers, Public Encryption, Digital Signature.

Moldovyan Alexandr Andreevich — Ph.D., Dr. Sci., Professor, Chief Researcher of Laboratory of Information Systems Security, St. Petersburg Institute for Informatics and Automation of Russian Academy of Sciences (SPIIRAS). Research interests: computer security, cryptography, network security, including security policy management, access control, authentication, network security analysis, intrusion detection, firewalls, deception systems, malware protection, verification of security systems; The number of publications — about 200. maa1305@yandex.ru, http://www.spiiras.nw.ru; SPIIRAS, 39, 14-th Line V.O., St. Petersburg, 199178, Russia; office phone +7(812)328–5185.

Moldovyan Nikolay Andreevich — Ph.D., Professor, Chief Researcher of Laboratory of Information Systems Security, St. Petersburg Institute for Informatics and Automation of Russian Academy of Sciences (SPIIRAS). Research interests: computer security, cryptography, symmetric and asymmetric cryptosystems, digital signature, authentication, block ciphers, pseudo-probabilistic ciphers. The number of publications — more 250. nmod@mail.ru, http://www.spiiras.nw.ru; SPIIRAS, 39, 14-th Line V.O., St. Petersburg, 199178, Russia; office phone +7(812)328–5185.

Acknowledgements. This research is supported by the Russian Foundation for Basic Research (project No. 18-07-00932-a).

References

1. Sirwan A., Majeed N. New Algorithm for Wireless Network Communication Security. *International Journal on Cryptography and Information Security.* 2016. vol. 6. no. 3/4. pp. 1–8.
2. Feng Y., Yang G., Liu. J.K. A new public remote integrity checking scheme with user and data privacy. *International Journal of Applied Cryptography*. 2017. vol. 3. no 3. pp. 196–209.
3. Chiou S.Y. Novel Digital Signature Schemes based on Factoring and Discrete Logarithms. *International Journal of Security and Its Applications.* 2016. vol. 10. no. 3. pp. 295–310.
4. Poulakis D. A variant of Digital Signature Algorithm. *Designs, Codes and Cryptography*. 2009. vol. 51. no. 1. pp. 99–104.
5. Yan S.Y. Quantum Computational Number Theory. Springer. 2015. 252 p.
6. Yan S.Y. Quantum Attacks on Public-Key Cryptosystems. Springer. 2014. 207 p.
7. Shor P.W. Polynomial-time algorithms for prime factorization and discrete logarithms on quantum computer. *SIAM Journal of Computing*. 1997. vol. 26. pp. 1484–1509.
8. Smolin J.A., Smith G., Vargo A. Oversimplifying quantum factoring. *Nature*. 2013. vol. 499. no. 7457. pp. 163–165.
9. Hamdi S.M., Zuhori S.T., Ffiroz M., Biprodip P. A Compare between Shor's quantum factoring algorithm and General Number Field Sieve. 2014 International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT). 2014. pp. 1–6.
10. Federal Register. The Daily Journal of the United States Government. Available at: https://www.gpo.gov/fdsys/pkg/FR-2016-12-20/pdf/2016-30615.pdf (accessed: 06.03.2018).
11. Verma G.K. A Proxy Blind Signature Scheme over Braid Groups. *International Journal of Network Security.* 2009. vol. 9. no 3. pp. 214–217.
12. Hiranvanichakorn P. Provably Authenticated Group Key Agreement based on Braid Groups − The Dynamic Case. *International Journal of Network Security*. 2017. vol. 19. no. 4. pp. 517−527.
13. Myasnikov A., Shpilrain V., Ushakov A. A Practical Attack on a Braid Group Based Cryptographic Protocol. Annual International Cryptology Conference. 2005. vol. 3621. pp. 86−96.
14. Chaturvedi A., Lal S. An Authenticated Key Agreement Protocol Using Conjugacy Problem in Braid Groups. *International Journal of Network Security*. 2008. vol. 6. no. 2. pp. 181–184.
15. Verma G.K. Probable Security Proof of a Blind Signature Scheme over Braid Groups. *International Journal of Network Security.* 2011. vol. 12. no. 2. pp. 118−120.
16. Kuzmin A.S. et al. Cryptographic Algorithms on Groups and Algebras. *Journal of Mathematical Sciences.* 2017. vol. 223. no. 5. pp. 629−641.
17. Moldovyan D.N., Moldovyan N.A., Shcherbacov V.A. Non-commutative finite associative algebras of 3-dimensional vectors. *Quasigroups and related systems.* 2018. vol. 26. no. 1. pp. 109−120.
18. Kuzmin A.S. et al. [Cryptographic Algorithms on Groups and Algebras]. *Fundamentalinaya i prikladnaya matematika – Fundumental and applied mathematics*. 2015. Issue 20. vol. 1. pp. 205−222. (In Russ.).
19. Glukhov M.M. [On analysis of some public key distribution systems based on non-abelian groups]. *Matematicheskie voprosy kriptografii – Mathematical Items of Cryptography.* 2010. Issue 1. vol. 4. pp. 5−22. (In Russ.).
20. Moldovyan D.N., Moldovyan N.A. Cryptoschemes over hidden conjugacy search problem and attacks using homomorphisms. *Quasigroups and Related Systems*. 2010. vol. 18. pp. 177−186.

21. Moldovyan D.N., Moldovyan N.A. A New Hard Problem over Non-Commutative Finite Groups for Cryptographic Protocols. International Conference on Mathematical Methods, Models, and Architectures for Computer Network Security. 2010. vol. 6258. pp. 183−194.
22. Moldovyan D.N. Non-Commutative Finite Groups as Primitive of Public-Key Cryptoschemes. *Quasigroups and Related Systems.* 2010. vol. 18. pp. 165−176.
23. Moldovyan A.A., Moldovyan N.A., Shcherbacov. V.A. Non-commutative finite associative algebras of 2-dimension vectors. *Computer Science Journal of Moldova*. 2017. vol. 25. no. 3(75). pp. 344−356.
24. Camenisch J.L., Piveteau J.-M., Stadler M.A. Blind Signatures Based on the Discrete Logarithm Problem. Workshop on the Theory and Application of Cryptographic Techniques. 1994. pp. 428–432.
25. Pointcheval D., Stern J. Security Arguments for Digital Signatures and Blind Signatures. *Journal of Cryptology*. 2000. vol. 13. no. 3. pp. 361−396.

Signed to print 12.04.2019
Printed in Publishing center GUAP, 67, B. Morskaya, St. Petersburg, 190000, Russia

The journal is registered in Russian Federal Agency for Communications
and Mass-Media Supervision, certificate ПИ № ФС77-41695 dated August 19, 2010 г.
Subscription Index П5513, Russian Post Catalog

Подписано к печати 12.04.2019. Формат 60×90 1/16. Усл. печ. л. 15,4 Заказ № 144.
Тираж 150 экз., цена свободная.
Отпечатано в Редакционно-издательском центре ГУАП, 190000, Санкт-Петербург, Б. Морская, д. 67

Журнал зарегистрирован Федеральной службой по надзору в сфере связи
и массовых коммуникаций,
свидетельство ПИ № ФС77-41695 от 19 августа 2010 г.

Подписной индекс П5513 по каталогу «Почта России»

РУКОВОДСТВО ДЛЯ АВТОРОВ

Взаимодействие автора с редакцией осуществляется через личный кабинет на сайте журнала «Труды СПИИРАН» http://www.proceedings.spiiras.nw.ru. При регистрации авторам рекомендуется заполнить все предложенные поля данных.

Подготовка статьи ведется с помощью текстовых редакторов MS Word 2007 и выше. Объем основного текста – от 20 до 30 страниц включительно. Формат страницы документа – A5 (148 мм ширина, 210 мм высота); ориентация – портретная; все поля – 20 мм. Верхний и нижний колонтитулы страницы – пустые. Основной шрифт документа – Times New Roman, основной кегль (размер) шрифта – 10 pt. Переносы разрешены. Абзацный отступ устанавливается размером в 10 мм. Межстрочный интервал – одинарный. Номера страниц не проставляются.

В основную часть допускается помещать рисунки, таблицы, листинги и формулы. Правила их оформления подробно рассмотрены на нашем сайте в разделе «Руководство для авторов».

AUTHOR GUIDELINES

Interaction between each potential author and the Editorial board is realized through the personal account on the website of the journal "Proceedings of SPIIRAS" http://www.proceedings.spiiras.nw.ru. At the registration the authors are requested to fill out all data fields in the proposed form.

The submissions should be prepared using MS Word 2007 text editor or higher versions, at that, only manuscripts in *.docx format will be considered. The text of the paper in the main part of it should be from 20 – 30 pages of A5 size that is 210 X 148 mm; orientation – portrait; all margins – 20 mm. The font of the main paper text is Times New Roman of 10 pt font size. The pages' headers and footers should be empty; indention – 10 mm; line spacing – single; pages are not numbered; hyphenations are allowed.

Certain figures, tables, listings and formulas are allowed in the main section, and their typography is considered by the paper template in more detail in journal web.

ISSN 2078-9181

Том 18 № 2

ТРУДЫ СПИИРАН
proceedings.spiiras.nw.ru

С-Пб
2019