На правах рукописи

ТОРОПОВА Александра Витальевна

МЕТОДЫ И АЛГОРИТМЫ ОБРАБОТКИ НЕОПРЕДЕЛЕННОСТИ ДАННЫХ ПРИ ОЦЕНИВАНИИ ИНТЕНСИВНОСТИ ПУАССОНОВСКОГО ПРОЦЕССА ПО ОГРАНИЧЕННОМУ ОБЪЕМУ ДОСТУПНЫХ НАБЛЮДЕНИЙ

Специальность 2.3.1 – Системный анализ, управление и обработка информации, статистика

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата технических наук

Санкт-Петербург – 2022

Работа выполнена в Федеральном государственном бюджетном образовательном учреждении высшего образования «Санкт-Петербургский государственный университет» (СПбГУ) и в лаборатории теоретических и междисциплинарных проблем информатики Федерального государственного бюджетного учреждения науки «Санкт-Петербургский Федеральный исследовательский центр Российской академии наук» (СПб ФИЦ РАН).

Научный руководитель: **ТУЛУПЬЕВ Александр Львович**
доктор физико-математических наук, профессор, главный научный сотрудник лаборатории теоретических и междисциплинарных проблем информатики Федерального государственного бюджетного учреждения науки «Санкт-Петербургский Федеральный исследовательский центр Российской академии наук», профессор кафедры информатики Федерального государственного бюджетного образовательного учреждения высшего образования «Санкт-Петербургский Государственный Университет».

Официальные оппоненты: **УТКИН Лев Владимирович**
доктор технических наук, профессор, директор Института компьютерных наук и технологий Федерального государственного автономного образовательного учреждения высшего образования «Санкт-Петербургский политехнический университет Петра Великого».

МОШКИН Вадим Сергеевич
кандидат технических наук, доцент кафедры информационных систем Федерального государственного бюджетного образовательного учреждения высшего образования «Ульяновский государственный технический университет».

Ведущая организация: Федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики».

Защита диссертации состоится **06 декабря 2022г. в 14 часа 00 минут** на заседании диссертационного совета 24.1.206.01, созданного на базе Федерального государственного бюджетного учреждения науки «Санкт-Петербургский Федеральный исследовательский центр Российской академии наук» (СПБ ФИЦ РАН) по адресу: 199178, Санкт-Петербург, 14-я линия В.О., 39, каб. 401, e-mail: dc@spcras.ru. Факс: (812) 328-44-50, тел: (812) 328-33-11.

С диссертацией и авторефератом можно ознакомиться в отделе аспирантуры (каб. 402а) Федерального государственного бюджетного учреждения науки «Санкт-Петербургский Федеральный исследовательский центр Российской академии наук» (СПБ ФИЦ РАН) по адресу: 199178, Санкт-Петербург, 14-я линия В.О., 39 и на сайте http://www.spiiras.nw.ru/dissovet/

Автореферат разослан **«14» октября 2022 года**.

Ученый секретарь
диссертационного совета 24.1.206.01
кандидат технических наук

АБРАМОВ
Максим Викторович

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования. Пуассоновский процесс традиционно используется для моделирования эпизодического поведения человека, являющегося предметом исследования во многих областях: маркетинге, социологии, медицине и здравоохранении (в частности, эпидемиологии), кибербезопасности. Оценка числовых характеристик эпизодического поведения, таких как частота реализации эпизодов на временной оси или интенсивность, используется при принятии решений в перечисленных областях.

В силу различных причин, таких как правовые нормы, ограниченность временных и финансовых ресурсов часто невозможно организовать прямое длительное наблюдение за поведением индивида, что ведет к необходимости использования данных из самоотчетов о поведении. Ограниченность числа наблюдений и неточность информации, предоставляемой на естественном языке, обусловливает актуальность задачи разработки инструментов получения косвенной оценки параметров процесса поведения с использованием математических моделей. Пуассоновский процесс выступает простейшей моделью эпизодического поведения индивида. Параметр интенсивности этого процесса отражает частоту реализации эпизодов процесса на временной оси, т.е. отношение количества эпизодов процесса за период исследования к количеству временных единиц, составляющих этот период исследования.

В диссертационном исследовании решается **научная задача** моделирования и обработки неопределенности, связанной с ответами респондентов.

Важность и значимость решаемой задачи обусловлены возможностью применения полученных результатов в различных социоориентированных областях, в которых требуется моделирование эпизодического поведения человека по ограниченному объему доступных наблюдений.

Степень разработанности темы. В работах А.Л. Тулупьева, С.И. Николенко, Т.В. Тулупьевой, А.Е. Пащенко и соавторов для минимизации неточности информации, предоставляемой респондентом, было предложено использовать данные о нескольких последних последовательных эпизодах процесса и рекордных интервалах. Дальнейшее развитие этот подход получил в работах А.В. Суворовой, А.В. Сироткина, В.Ф. Столяровой. Для получения оценки интенсивности эпизодов поведения на временной оси с учетом возникающей неопределенности получаемой информации об эпизодах было предложено использовать байесовские сети доверия. Однако предложенные модели не в полной мере учитывают неопределенность,

связанную с получаемыми данными: не учитывается некорректность получаемых от респондента данных об эпизодах поведения, несогласованность таких данных или же некорректное задание момента окончания исследования.

Заделом для диссертационного исследования в том числе послужили работы Н.Г. Ярушкиной и С.М. Ковалева, посвященные анализу сверхкоротких нечетких временных рядов; М.К. Кано-Лосано и Дж.А. Вольфсон, отражающие подходы к оцениванию интенсивности определенных видов поведения; Р.А. Рехфельдта, описывающие методы сбора информации о поведении; С.А. Потрясаева, А.Л. Ронжина, Б.В. Соколова, С.В. Микони, Р.М. Юсупова и А.В. Смирнова, предлагающие подходы к моделированию сложных объектов и систем.

Цель диссертационного исследования заключается в повышении качества классификации при оценивании интенсивности пуассоновского процесса как математической модели эпизодического поведения индивида за счет разработки методов и алгоритмов обработки неопределенности данных, предоставляемых респондентами.

Для достижения поставленной цели был сформулирован ряд **задач**:

1. Разработать метод и алгоритм оценивания согласованности информации о последних эпизодах и рекордных интервалах пуассоновского процесса в модели оценивания интенсивности пуассоновского процесса, основанной на байесовской сети доверия.

2. Разработать алгоритм обработки некорректности информации, полученной от респондентов, при оценивании интенсивности пуассоновского процесса, выступающего математической моделью поведения индивида, для улучшения оценки с точки зрения показателей качества классификации.

3. Разработать метод и алгоритм обработки неопределенности задания конца исследуемого периода при оценивании интенсивности пуассоновского процесса по ограниченному объему доступных наблюдений для улучшения оценки с точки зрения показателей качества классификации.

4. Разработать архитектуру и прототип комплекса программ, реализующих предложенные методы и алгоритмы, для их апробации, вычислительных экспериментов и решения практических задач.

Объектом исследования являются сверхкороткие, неполные и неточные данные о последовательных эпизодах и рекордных интервалах пуассоновского процесса.

Предметом исследования являются методы, алгоритмы и модели обработки неопределенности в задаче оценивания интенсивности пуассоновского процесса по сверхкоротким, неполным и неточным

данным о последовательных эпизодах поведения и рекордных интервалах между эпизодами поведения за определенный промежуток времени.

Научная новизна диссертационной работы определяется тем, что:

1. Предложены новые метод и алгоритм оценивания согласованности информации о последних эпизодах и рекордных интервалах пуассоновского процесса, отличающиеся применением расширенной, по отношению к используемой ранее, байесовской сети доверия с дополнительными узлами принятия решений, обеспечивающие возможность работы с данными необходимой степени согласованности.

2. Предложен новый алгоритм обработки некорректности информации, полученной от респондентов, при оценивании интенсивности пуассоновского процесса, выступающего математической моделью поведения индивида, отличающийся использованием новых скрытых переменных в байесовской сети доверия, отвечающих истинным длинам интервалов, обеспечивающий повышение показателей качества классификации по сравнению с предложенными ранее подходами.

3. Предложены новые метод и алгоритм обработки неопределенности задания конца исследуемого периода при оценивании интенсивности пуассоновского процесса по ограниченному объему доступных наблюдений на основе байесовской сети доверия, отличающиеся внедрением вершины, характеризующей интервал между последним эпизодом пуассоновского процесса и эпизодом, произошедшим после окончания периода исследования, обеспечивающие повышение показателей качества классификации по сравнению с предложенными ранее подходами при наличии ретроспективных данных.

4. Разработаны архитектура и прототип комплекса программ для оценивания интенсивности пуассоновского процесса по ограниченному объему доступных наблюдений, отличающиеся наличием инструментов обработки неопределенности данных, связанной с ответами респондентов, обеспечивающие работу с предложенными в диссертационной работе алгоритмами.

Теоретическая и практическая значимость работы. В рамках диссертации были разработаны математические модели на основе байесовских сетей доверия, которые могут быть использованы для оценивания интенсивности поведения и адаптированы к использованию информации со значительной долей неопределенности. Такие модели являются новыми и представляют теоретическую значимость.

Разработанные методы и алгоритмы создают основу для обработки неопределенности и некорректности информации, полученной от респондентов, при построении оценок интенсивности эпизодического поведения человека по ограниченному объему доступных наблюдений.

Кроме того, предложенные методы и алгоритмы обеспечивают возможность автоматизации решения задачи построения оценок и составляют основу для создания систем поддержки принятия решений в социоориентированных областях знаний.

Методология исследования. Задача обработки неопределенности данных при оценивании числовых параметров пуассоновского процесса по ограниченному объему доступных наблюдений является мультидисциплинарной. Методология системного анализа позволяет представить рассматриваемую задачу как общую задачу разработки специального математического и алгоритмического обеспечения системы принятия решений и обработки информации в условиях информационного дефицита в различных социоориентированных областях.

Методы исследования. В аналитической части работы используются методы теории вероятностей для построения математической модели поведения (пуассоновский процесс) и методы байесовской статистики и машинного обучения для обработки возникающей неопределенности, в частности теория байесовских сетей доверия. Для получения данных были использованы методы поиска и сбора информации из социальных сетей, также были использованы методы синтеза данных и имитационное моделирование зашумленных эпизодов пуассоновского процесса со случайной интенсивностью.

Положения, выносимые на защиту:

1. Метод и алгоритм оценивания согласованности информации о последних эпизодах и рекордных интервалах пуассоновского процесса в модели оценивания его интенсивности, основанной на байесовской сети доверия.
2. Алгоритм обработки некорректности информации об эпизодах поведения, полученной от респондентов, при оценивании интенсивности пуассоновского процесса, выступающего математической моделью поведения индивида.
3. Метод и алгоритм обработки неопределенности задания конца исследуемого периода при оценивании интенсивности пуассоновского процесса по ограниченному объему доступных наблюдений.
4. Архитектура и прототип комплекса программ, реализующие разработанные методы и алгоритмы.

Соответствие диссертации научной специальности. Представленные результаты соответствуют специальности 2.3.1 — «Системный анализ, управление и обработка информации, статистика».

Успешная апробация результатов предложенной диссертации на научных конференциях различного уровня, в том числе международных и

российских, публикации в рецензируемых изданиях, согласованность результатов, качественный анализ тематики и корректное использование приведенных моделей и математических методов обусловили высокую **степень достоверности** полученных в исследовании результатов.

Апробация результатов. Основные научные мероприятия, в ходе которых были представлены и обсуждались результаты предлагаемого диссертационного исследования: Научная сессия НИЯУ МИФИ-2015 (Москва, 2015 г.); VIII Международная научно-техническая конференция «Интегрированные модели и мягкие вычисления в искусственном интеллекте» (Коломна, 2015 г.); III, IV Международная школа-семинар по искусственному интеллекту для студентов, аспирантов и молодых ученых «Интеллектуальные системы и технологии: современное состояние и перспективы» — ISYT (Тверь, 2015 г.; Санкт-Петербург, 2017 г.); XVIII–XXI, XXIII–XXV Международная конференция по мягким вычислениям и измерениям — SCM (Санкт-Петербург, 2015–2018, 2020–2022 гг.); III Всероссийская Поспеловская конференция с международным участием «Гибридные и синергетические интеллектуальные системы» (Светлогорск, 2016 г.); I, III International Scientific Conference «Intelligent Information Technologies for Industry» — IITI (Sochi, 2016, 2018 гг.); Всероссийская научная конференция по проблемам информатики — СПИСОК (Санкт-Петербург, 2016, 2017 гг.); VII,VIII всероссийская научно-практическая конференция «Нечеткие системы, мягкие вычисления и интеллектуальные технологии» — НСМВИТ (Санкт-Петербург, 2017 г.; Смоленск, 2020 г.); The 11th Conference of the European Society for Fuzzy Logic and Technology organized jointly with the IQSA Workshop on Quantum Structures (Prague, 2019); XVIII Национальная конференция по искусственному интеллекту КИИ–2020 (Москва, 2020 г.); VI Международная научно-практическая конференция ICIT–2020 (Саратов, 2020 г.).

Полученные в диссертации результаты, являются частью научно-исследовательских проектов, поддержанных следующими грантами РФФИ: «Машинное обучение и структурные особенности байесовской сети доверия со скрытыми переменными как модели социально-значимого поведения» № 19-37-90120, 2019–2021 (грант «Аспиранты»); «Методы идентификации параметров социальных процессов по неполной информации на основе вероятностных графических моделей» № 16-31-00373, 2016–2017; «Комбинированный логико-вероятностный графический подход к представлению и обработке систем знаний с неопределенностью: алгебраические байесовские сети и родственные модели» № 15-01-09001, 2015–2017.

Полученные результаты были использованы при проведении исследовательских работ СПб ФИЦ РАН, в учебном процессе СЗИУ РАНХиГС, а также при разработке подходящего для клиента режима физических нагрузок в ООО «Хоум Фитнес», получены соответствующие акты внедрения.

Публикации. Соискателем было сделано 42 публикации и научная работа, к ним приравненная. В это число входят: 1 монография; 11 публикаций в изданиях, индексируемых Scopus/WoS; 4 статьи в изданиях из «Перечня рецензируемых научных изданий, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук» (из которых 2 единоличных); 23 доклада и тезиса на научных конференциях (из которых 9 единоличных); получено 1 свидетельство о регистрации программ для ЭВМ (Роспатент).

Личный вклад соискателя в опубликованных работах отражают содержание диссертации и основные положения, выносимые на защиту. Публикация полученных результатов проводилась совместно с научным руководителем А.Л. Тулупьевым и членами лаборатории теоретических и междисциплинарных проблем информатики СПб ФИЦ РАН, причем вклад соискателя был существенным. Представленные к защите результаты получены лично автором.

Структура и объем диссертации. Диссертация включает в себя введение, четыре главы, заключение, список сокращений, список литературы (255 источников), списки иллюстративного материала и таблиц, приложения. Общий объем диссертации — 184 страницы, включая 35 таблиц, 52 рисунка и 3 приложения.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность исследуемой научной задачи, сформулированы цель и задачи исследования, определена научная новизна, представлены теоретическая и практическая значимость работы, перечислены основные положения, выносимые на защиту.

В **первой главе** обоснованы актуальность диссертационного исследования, использование байесовских сетей доверия для задачи оценивания интенсивности пуассоновского процесса, выступающего моделью эпизодического поведения. Проведен обзор сфер применения байесовских сетей доверия и инструментария для работы с ними. Рассмотрены возможные подходы к обучению байесовских сетей доверия.

Во многих областях науки возникают задачи, связанные с моделированием и оцениванием параметров точечных случайных процессов. В случае невозможности или большой ресурсозатратности работы с реальными объектами или системами, используются их

математические модели. В данном исследовании рассматривается однородный пуассоновский процесс, параметр которого отражает интенсивность реализации эпизодов на временной оси. При этом в ряде случаев возникает ситуация, когда о процессе известен лишь ограниченный объем доступных наблюдений.

«Золотым стандартом» получения данных об эпизодах процесса поведения является прямое наблюдение. В силу различных причин (правовые нормы, ограниченность ресурсов) его может быть трудно или даже невозможно реализовать. Поэтому для получения оценки интенсивности эпизодического поведения приходится прибегать к построению математических моделей. Полученные при помощи математических моделей оценки интенсивности поведения являются косвенными.

Дневниковый метод (последовательная и подробная запись действий респондента за определенный промежуток времени) позволяет получить достаточно точную информацию об эпизодах поведения, однако является ресурсозатратным.

Для получения информации об эпизодах поведения в рамках интервью или опроса наиболее часто используются прямые вопросы типа «Как часто Вы совершали определенное действие за последний месяц?», ответы на которые подразумевают либо число, либо относительную словесную характеристику («редко, часто, иногда, всегда»). Ответы на такие прямые вопросы часто некорректны, не отражают истинную ситуацию. С целью извлечения наиболее точной числовой информации об эпизодах поведения в работах Т.В. Тулупьевой и соавторов был предложен метод последних эпизодов и рекордных интервалов. Этот метод опирается на вопросы: «Когда Вы в последний / предпоследний / предпредпоследний раз совершали …?» и «Каков минимальный / максимальный временной интервал между …?». Ответы на такие вопросы обладают рядом преимуществ перед ответами на прямые вопросы: являются численными описаниями конкретной характеристики (момента реализации некоторого действия). Однако получаемая таким способом информация может быть подвержена некоторой неопределенности, а именно ответы респондентов могут быть некорректны, неточны или не согласованы, также момент интервью может быть некорректно задан.

Формулировка исследуемой задачи оценивания интенсивности процесса схожа с задачами из области сверхкоротких временных рядов, однако для применения этого метода необходимо больше данных об эпизодах процессов. Регрессионный анализ используется для построения оценок интенсивности поведения по ограниченным данным об эпизодах поведения, однако этот метод не приспособлен к учету значительной

неопределенности входных данных. Байесовские сети доверия являются удобным инструментом для учета некоторых типов неопределенности, связанной с ситуацией сбора самоотчетов респондентов об их поведении. Байесовские сети доверия обладают высокой интерпретируемостью и обширным инструментарием для работы с ними, что определило их выбор в данном исследовании как основного инструмента.

Вторая глава посвящена описанию теоретических результатов, послуживших предпосылками для данного диссертационного исследования и создающих основу для решения поставленных задач. Описаны основы теории байесовских сетей доверия. Представлен разработанный ранее подход к оцениванию интенсивности пуассоновского процесса как модели эпизодического поведения индивида по сверхкороткому набору наблюдений с использованием байесовских сетей доверия. Также описываются используемые метрики качества моделей и структура исследования. Данная глава не включает результаты данного диссертационного исследования, а предназначена для введения единой системы обозначений и описания предложенного ранее подхода к моделированию пуассоновских процессов, в развитии которого заключается суть этой работы.

Байесовская сеть доверия — это вероятностная графическая модель, которая представляет собой ациклический направленный граф, с вершинами которого ассоциированы случайные элементы, а ребра обозначают причинно-следственные связи между элементами. С ребрами сети связаны тензоры условной вероятности.

В диссертационном исследовании используются байесовские сети доверия с многозначными случайными элементами в узлах сети. Таким образом, задача оценивания интенсивности пуассоновского процесса сводится к задаче классификации, поэтому для определения качества моделей используется матрица ошибок и выводимые из нее метрики: точность и средняя точность.

Пуассоновский процесс является естественной моделью для рассматриваемого эпизодического поведения, отвечающего следующим предположениям: эпизоды поведения происходят в непрерывном времени, за конечный промежуток может произойти только конечное число эпизодов, эпизоды не могут происходить одновременно, время реализации эпизода для каждого индивида не зависит от времени предыдущих эпизодов, и интенсивность поведения индивида остается постоянной во времени. Гамма-пуассоновская модель поведения возникает при учете неоднородности популяции в плане эпизодического поведения: интенсивность поведения варьируется между индивидами и моделируется гамма–распределенной случайной величиной.

На рисунке 1 представлена обобщенная модель байесовской сети доверия $M = (G(V,L), \mathbf{P})$. Ее структура выражена графом $G(V,L)$, где $V = \{t_{01}, t_{12}, t_{23}, t_{\min}, t_{\max}, \lambda, n\}$ — множество вершин, $L = \{(u,v) : u,v \in V\}$ — множество направленных связей между вершинами.

λ — это гамма–распределенная случайная величина, описывающая интенсивность эпизодов поведения. Ее распределение задается гамма-распределением вероятности. t_{ij} — случайная величина, которая соответствует длине интервала между *i*-ым и *j*-ым эпизодами поведения (здесь и далее будем полагать, что i может принимать значения 1 и 2, а $j = i+1$). В рамках гамма–пуассоновской модели поведения, t_{ij} имеют экспоненциальное распределение. Интервал между моментом получения информации от респондента и последним эпизодом t_{01} не является интервалом между эпизодами поведения. $t_{\min}$ и $t_{\max}$ — минимальный и максимальный интервалы за определенный промежуток времени T. n — скрытая переменная, которая соответствует числу эпизодов процесса за отрезок времени *T*. Все непрерывные случайные величины, входящие в модель, полагаются дискретизированными.

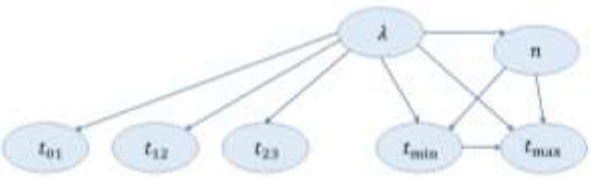

Рисунок 1 — Модель оценивания интенсивности пуассоновского процесса

При оценивании интенсивности пуассоновского процесса по неполным и неточным данным самоотчетов респондентов об их поведении в работе рассматривается неопределенность следующих видов: 1. Несогласованность ответов респондентов. 2. Некорректность ответов респондентов. 3. Некорректность задания момента окончания исследования. Под некорректностью ответов респондентов, понимается их расхождение с действительным временем реализации эпизодов, а под несогласованностью информации понимаются ситуации, когда респондент дает ответы, противоречащие друг другу. При создании и анализе моделей обработки таких типов неопределенности были пройдены этапы моделирования, предложенные С.В. Микони, Б.В. Соколовым и Р.М. Юсуповым.

Формальная постановка задачи выглядит следующим образом.

Дано: исходная модель $M = \left(G(V,L), \mathbf{P}\right)$, сведения респондентов о последних эпизодах и рекордных интервалах пуассоновского процесса, выступающего математической моделью эпизодического поведения, то есть поступает свидетельство $E = \left\{t_{01}, t_{12}, t_{23}, t_{\min}, t_{\max}\right\}$, в котором данные могут быть ошибочны. **Требуется:** построить такие модели, что их средняя точность классификации статистически выше, чем у исходной модели. Задача моделей оценивания интенсивности пуассоновского процесса правильно отнести величину λ к соответствующему классу дискретизации, а $r^0 = \max\limits_{i=1,\ldots,m} \mathrm{P}(\lambda_i \mid E)$ — это оценки величины λ исходной модели с учетом поступившего свидетельства E. Требуется предложить модель $M^* = \left(G^*(V^*, L^*), \mathbf{P}^*\right)$, оценки $r^* = \max\limits_{i=1,\ldots,m} \mathrm{P}(\lambda_i \mid E)$ которой будут такими, что средняя точность r^* будет больше средней точности r^0.

Третья глава представляет собой описание полученных соискателем теоретических результатов. Предложены метод и алгоритм оценивания согласованности информации о последних эпизодах и рекордных интервалах пуассоновского процесса в модели оценивания его интенсивности, основанной на байесовской сети доверия; алгоритм обработки некорректности информации, полученной от респондентов, при оценивании интенсивности пуассоновского процесса, выступающего математической моделью поведения индивида; метод и алгоритм обработки неопределенности задания конца исследуемого периода при оценивании интенсивности пуассоновского процесса по ограниченному объему доступных наблюдений. Рассмотрена модель оценивания интенсивности пуассоновского процесса, расширенная скрытыми переменными, отражающими истинную информацию о времени последних эпизодов и рекордных интервалах процесса за определенный промежуток времени, со структурой, обученной на синтетических данных. Описаны модели оценивания интенсивности постинга в социальных сетях, расширенные за счет объективных данных о пользователе. Также рассмотрены возможные варианты дискретизации непрерывных величин, входящих в модели.

Для того, чтобы определить, насколько согласованы между собой данные, представляемые респондентом, предлагается метод, основанный на расширении модели узлами, которые отражают согласованность информации о последних эпизодах и рекордных интервалов (рис. 2).

Вероятность согласованности означиваний случайных величин t_{12} и t_{23} со значением случайной величины $t_{\min}$ рассчитывается следующим

образом: если данные респондента противоречат друг другу (указанный интервал меньше минимального или больше максимального), то она равна 0, если данные согласованы — 1. Согласованность t_{01} и $t_{\min}$ не рассматривается в связи с тем, что t_{01} не является интервалом между эпизодами поведения.

Вершина $c_{t_{ij},\min}$, определяющая оценку согласованности информации респондента, может иметь одно из трех значений: свидетельства t_{ij} и $t_{\min}$ согласованы (обозначим $c^{+}_{t_{ij},\min}$), не согласованы ($c^{-}_{t_{ij},\min}$), и согласованность не определена ($c^{?}_{t_{ij},\min}$). Последнее значение $c_{t_{ij},\min}$ принимает, если значения t_{ij} и $t_{\min}$ относятся к одному интервалу.

Тензоры условной вероятности $p\left(c^{(s)}_{t_{ij},\min} \mid t_{ij}, t_{\min}\right)$ задаются следующим образом: $P\left(c^{(s)}_{t_{ij},\min} \mid t_{ij}, t_{\min}\right) = \begin{cases} \alpha^{(s)}, & t_{ij} > t_{\min}; \\ \beta^{(s)}, & t_{ij} < t_{\min}; \\ 1-\alpha^{(s)}-\beta^{(s)}, & t_{ij} = t_{\min}; \end{cases}$ где

$s \in \{+,-,?\}$, $\alpha^{(s)}, \beta^{(s)} \in [0;1]$, $\alpha^{(s)} + \beta^{(s)} \le 1$, $\sum\limits_{s\in\{+,-,?\}} \alpha^{(s)} = 1$, $\sum\limits_{s\in\{+,-,?\}} \beta^{(s)} = 1$.

Оценки согласованности с интервалом $t_{\max}$ рассчитываются аналогичным образом (при этом t_{01} также рассматривается).

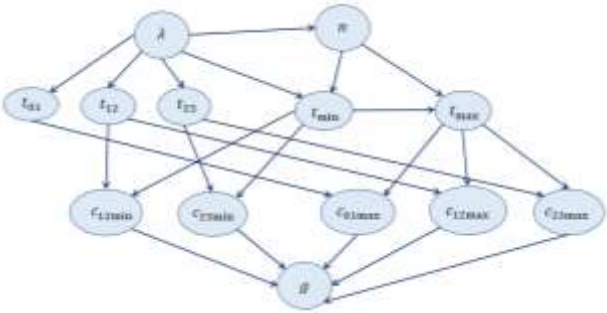

Рисунок 2 — Расширенная модель оценивания интенсивности пуассоновского процесса с диагностикой согласованности информации об эпизодах поведения, получаемой от респондентов

Также для получения общей оценки можно добавить к этой модели вершину g (рис. 2), которая характеризует оценку общей

согласованности данных. Для простоты обозначим $c = \left(c_{t_{12,\min}}, c_{t_{23,\min}}, c_{t_{01,\max}}, c_{t_{12,\max}}, c_{t_{23,\max}} \right)$, тогда $p(g^{(s)} \mid c) = \frac{\sum c^{(s)}}{\sum c}$.

Схема алгоритма получения оценки согласованности информации, полученной от респондентов, представлена на рис. 3.

Предлагается алгоритм обработки некорректности информации, полученной от респондентов, при оценивании интенсивности пуассоновского процесса, выступающего математической моделью поведения индивида. Этот алгоритм основан на добавлении в модель скрытых переменных, отражающих истинную информацию об эпизодах процесса.

В модель добавлены вершины t_{01}^0, t_{12}^0, t_{23}^0, $t_{\min}^0$ и $t_{\max}^0$ (см. рис. 4), представляющие интервалы процесса, полученные из ответов респондентов. А истинное время реализации эпизодов t_{01}, t_{12}, t_{23}, $t_{\min}$ и $t_{\max}$ моделируется скрытыми переменными, наблюдение которых невозможно.

Рисунок 3 — Схема алгоритма оценивания согласованности информации о последних эпизодах и рекордных интервалах пуассоновского процесса в модели оценивания его интенсивности

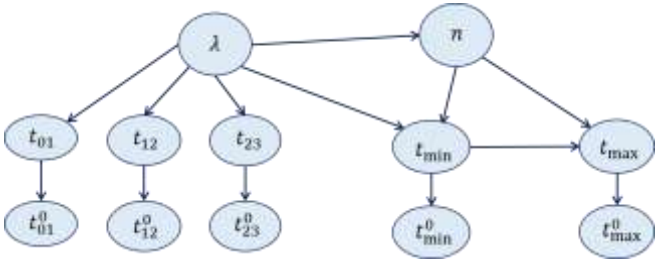

Рисунок 4 — Модель оценивания интенсивности пуассоновского процесса со скрытыми переменными

Тензоры условной вероятности определяются следующим образом:

$$P(t_{ij}^0 \mid t_{ij}) = \begin{cases} c, \mid t_{ij}^0 - t_{ij} \mid = 0, \\ c \cdot q, \mid t_{ij}^0 - t_{ij} \mid = 1, \\ c \cdot q^2, \mid t_{ij}^0 - t_{ij} \mid = 2, , \\ \ldots \\ c \cdot q^n, \mid t_{ij}^0 - t_{ij} \mid = n, \end{cases}$$ где $q \in [0;1)$, а c — это нормализирующая константа. Также тензоры условной вероятности можно получить при обучении модели на статистических данных.

Алгоритм обработки некорректности информации, полученной от респондентов, при оценивании интенсивности пуассоновского процесса, выступающего математической моделью поведения индивида, использующий модель байесовской сети доверия со скрытыми переменными, представлен на рис. 5.

Также предлагается новый метод обработки неопределенности задания конца исследуемого периода при оценивании интенсивности пуассоновского процесса по ограниченному объему доступных наблюдений, использующий при обучении модели сведения о значениях длин интервалов между последними тремя эпизодами, интервале между последним эпизодом и гипотетически «следующим» и рекордных интервалах.

Рисунок 5 — Схема алгоритма обработки некорректности информации об эпизодах поведения, получаемой от респондентов, при оценивании интенсивности пуассоновского процесса

На рис. 6 скрытая переменная t_{next} соответствует интервалу между последним эпизодом, входящим в исследуемый период, и первым эпизодом, который гипотетически произойдет по окончании исследуемого периода.

Использование этой модели может повысить качество оценивания интенсивности пуассоновского процесса в тех случаях, если для обучения модели имеется набор ретроспективных данных, таким образом t_{next} на этапе обучения не будет являться скрытой переменной.

Схема соответствующего алгоритма похожа на схему алгоритма обработки некорректности информации, полученной от респондентов, при оценивании интенсивности пуассоновского процесса на основе модели со скрытыми переменными (рис. 5), единственное отличие заключается в том, что не нужно задавать параметры «зашумленных» данных (в диссертационной работе рис. 3.13).

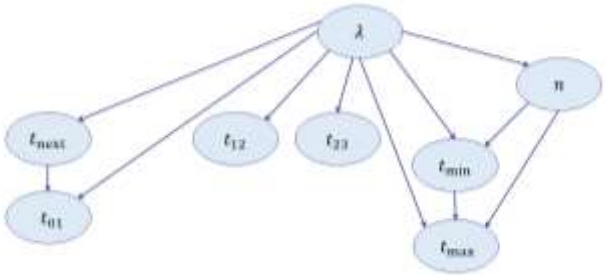

Рисунок 6 — Модель оценивания интенсивности пуассоновского процесса, обрабатывающая неопределенность задания конца исследуемого периода при оценивании интенсивности пуассоновского процесса по ограниченному объему доступных наблюдений

В **четвертой главе** описан прототип комплекса программ, реализующего разработанные методы и алгоритмы. Также описаны собранные данные (синтетические и данные из социальных сетей ВКонтакте и Instagram[1]) для апробации предложенных методов и алгоритмов, результаты этой апробации, а также описывается их применение.

На рис. 7 представлена архитектура прототипа комплекса программ для работы с моделями оценивания интенсивности пуассоновского процесса, учитывающими рассматриваемые типы неопределенности. Этот прототип обеспечивает работу с предложенными методами и алгоритмами для специалистов из различных областей, изучающих поведение человека. Прототип комплекса состоит из трех модулей: модуль для работы с инструментом оценивания согласованности информации о последних эпизодах и рекордных интервалах процесса; модуль для работы с моделью оценивания интенсивности пуассоновского процесса со скрытыми переменными; модуль для работы с моделью

[1] Организация, запрещенная на территории РФ и признанная экстремистской.

оценивания интенсивности пуассоновского процесса с гипотетически «следующим» эпизодом. Изначально пользователь выбирает, с каким именно модулем нужно начать работу, далее пользователь заполняет по порядку необходимые формы, при этом происходят процессы чтения и сохранения данных.

При разработке инструментария для оценивания согласованности ответов респондентов использовались C# и библиотека Smile, для остальных компонент — C#, R, в частности пакет bnlearn, и библиотека RDotNet для взаимодействия между двумя языками.

Код для синтеза данных о последних эпизодах поведения написан на языке R. Программа генерирует «эпизоды поведения» по гамма-пуассоновской модели. Некорректность ответов респондентов реализована в синтетических данных следующим образом: отклонение между окончанием сбора данных и последним эпизодом и от значений длин рекордных интервалов составляет не более четверти; между последними двумя эпизодами — не более половины; отклонение от интервала между предпоследним и предпредпоследним эпизодами максимум вдвое (используется равномерное распределение). Это согласуется с предположением о том, что чем раньше эпизод, тем сложнее его вспомнить, а рекордные интервалы вспоминаются достаточно легко. После того, как данные синтезированы значения непрерывных величин дискретизируются.

Для сбора данных из социальных сетей были написаны специальные программы, извлекающие информацию о времени публикаций пользователя и их количестве за определенный период времени.

Для сбора самоотчетов респондентов об их поведении был разработан опросный инструментарий, который содержит следующие вопросы: имя пользователя, сведения о значениях длин интервалов между последними тремя эпизодами публикации постов и данные о наибольшем и наименьшем значении длины интервалах между публикациями за год. Были собраны данные о 92 пользователях социальной сети Instagram[2]. Эти данные были использованы для тестирования модели со скрытыми переменными, обученной на синтетических данных. По полученным результатам было показано статистически достоверное улучшение точности на 10% (0.254 для исходной модели и 0.355 для предложенной) и средней точности на 2.6% (0.813 для исходной модели и 0.839 для предложенной) по сравнению с предложенным ранее алгоритмом: 95% доверительные интервалы для точности и средней точности, построенные методом бутстрап для 5000 репликаций, составляют (0.1972; 0.2535) и

[2] Организация, запрещенная на территории РФ и признанная экстремистской.

(0.7993; 0.8134) для исходной модели, а (0.2948; 0.4510) и (0.8237; 0.8627) — для предложенной модели оценивания интенсивности пуассоновского процесса со скрытыми переменными.

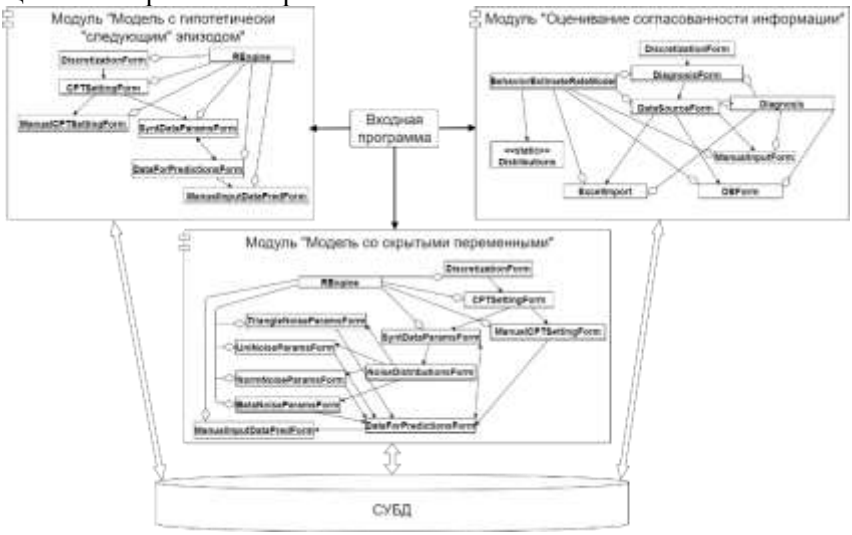

Рисунок 7 — Архитектура прототипа комплекса программ для работы с моделями оценивания интенсивности пуассоновского процесса по ограниченному набору доступных наблюдений

Также были проведены численные эксперименты на синтетических данных, и для разных моделей ошибки (равномерный шум с минимумом равным 0, максимумом — длине интервала, увеличенной вдвое; нормальный шум с математическим ожиданием равным длине интервала и дисперсией распределения равной длине интервала; треугольный шум с минимумом равным половине длины интервала, максимумом — длине интервала, увеличенной в 3 раза, и модой распределения равной длине интервала) было получено статистически достоверное улучшение качества классификации на 1% при использовании алгоритма обработки некорректности информации, полученной от респондентов в задаче оценивания интенсивности пуассоновского процесса, выступающего математической моделью поведения индивида.

Для обработки неопределенности, возникающей при задании конца исследуемого периода, были предложены метод и алгоритм оценивания интенсивности пуассоновского процесса по ограниченному объему доступных наблюдений, в основе которых лежит байесовская сеть доверия с дополнительным интервалом между эпизодами процесса. Были использованы синтетические данные, в которых длина интервала между моментом интервью и последним эпизодом пуассоновского процесса

зашумлена с использованием усеченного нормального распределения со следующими параметрами: математическое ожидание равно длине интервала между моментом интервью и последним эпизодом пуассоновского процесса, дисперсия равна удвоенной длине интервала между моментом интервью и последним эпизодом пуассоновского процесса. Было показано статистически достоверное улучшение средней точности (средняя точность равна 0.904) на 0.6% по сравнению с методом, предложенным ранее (средняя точность равна 0.898): 95% доверительный интервал для средней точности, построенный методом бутстрап для 20000 репликаций, составляет (0.8951; 0.8992) — для исходной модели оценивания интенсивности пуассоновского процесса и (0.9016; 0.9050) — для модели оценивания интенсивности пуассоновского процесса, обрабатывающей неопределенность задания конца исследуемого периода.

Улучшение качества классификации на 1% важно при проведении популяционных исследований, так как в этом случае даже такое улучшение имеет значительный экономический эффект.

Таким образом, показатели качества оценки в предложенных моделях оценивания интенсивности пуассоновского процесса выше по сравнению с исходными. Кроме того, использование предложенных методов позволяет снизить влияние неопределенности при оценивании интенсивности эпизодического поведения индивидов что немаловажно в областях, где данные получать сложно и дорого.

Можно заключить, что предложенных методов и алгоритмов достаточно для достижения поставленной цели: предлагаются три подхода к обработке трех видов неопределенности, возникающей при оценивании интенсивности пуассоновского процесса по ограниченному объему данных.

Все результаты работы, соответствующие положениям, выносимые на защиту, были внедрены в научно-исследовательской работе СПб ФИЦ РАН по государственному заданию № 0073-2019-0003 «Состояние и перспективы развития информационного общества и цифровой экономики в России», что позволило повысить качество получаемых оценок интенсивности социально-значимого поведения респондентов по данным о последних эпизодах по сравнению с предыдущими результатами проектов, реализованных в лаборатории теоретических и междисциплинарных проблем информатики СПб ФИЦ РАН. Разработанные методы используются в учебном процессе СЗИУ РАНХиГС. Также внедрение модуля разработанного комплекса программ в ООО «Хоум Фитнес» позволило усовершенствовать процесс построения подходящего для клиента режима физических нагрузок.

ЗАКЛЮЧЕНИЕ

В диссертационной работе решена научная задача обработки некоторых типов неопределенности данных при оценивании интенсивности пуассоновского процесса по ограниченному объему доступных наблюдений. Было повышено качество классификации при оценивании интенсивности пуассоновского процесса как математической модели эпизодического поведения индивида за счет разработки методов и алгоритмов обработки неопределенности данных, предоставляемых респондентами. Решенная задача имеет важное значение для совершенствования методов и алгоритмов, используемых в системном анализе, оптимизации, управлении, принятии решений и обработке информации, связанных с моделированием человеческого поведения и улучшением качества оценок его интенсивности.

Итоги исследования включают нижеперечисленные научные результаты:

— Разработаны метод и алгоритм оценивания согласованности информации о последних эпизодах и рекордных интервалах пуассоновского процесса, в основе которых лежит расширенная байесовская сеть доверия с дополнительными узлами принятия решений. С помощью этого инструмента можно оценить, насколько согласована информация, полученная от респондентов, и дальше действовать в зависимости от целей исследования: например, исключить из рассмотрения данные, не удовлетворяющие определенному порогу согласованности.

— Разработан алгоритм обработки некорректности информации, полученной от респондентов, при оценивании интенсивности пуассоновского процесса, выступающего математической моделью поведения индивида, в основе которого лежит использование скрытых переменных в байесовской сети доверия, отвечающих истинным длинам интервалов. Было показано, что использование в модели таких скрытых переменных позволяет улучшить показатели качества классификации по сравнению с предложенным ранее подходом.

— Разработаны метод и алгоритм обработки неопределенности задания конца исследуемого периода при оценивании интенсивности пуассоновского процесса по ограниченному объему доступных наблюдений на основе байесовской сети доверия с дополнительной вершиной, характеризующей интервал между последним эпизодом пуассоновского процесса и ненаблюдаемым эпизодом, произошедшим после окончания периода исследования. При наличии ретроспективных данных для исследования использование этой модели позволяет улучшить

показатели качества классификации по сравнению с предложенным ранее подходом.

— Разработаны архитектура и прототип комплекса программ для работы с предложенными новыми методами и алгоритмами для их апробации, вычислительных экспериментов и решения практических задач.

По итогам исследования даны **рекомендации** по применению полученных результатов в научных исследованиях и прикладных задачах, в которых важно оценивать числовые характеристики поведения человека. Результаты, представленные в диссертации, применимы в качестве инструмента автоматизации оценивания одного из ключевых параметров эпизодического поведения — его интенсивности, с учетом неточности и некорректности данных самоотчетов респондентов. Результаты данного исследования предназначены для использования специалистами, изучающими поведение человека, в условиях ограниченности ресурсов. С помощью разработанных инструментов можно получить довольно высокое качество оценивания интенсивности поведения на основе небольшого количества данных.

Метод оценивания согласованности информации о поведении респондента можно использовать для того, чтобы отфильтровать сведения респондентов ненадлежащего качества. Алгоритм обработки некорректности информации, полученной от респондентов, при оценивании интенсивности пуассоновского процесса, выступающего математической моделью поведения индивида, по ограниченному объему доступных наблюдений, на основе модели байесовской сети доверия со скрытыми переменными, можно использовать в тех случаях, когда особенно важно учитывать то, что респонденты могут предоставить неверные данные. Метод обработки неопределенности задания конца исследуемого периода при оценивании интенсивности пуассоновского процесса по ограниченному объему доступных наблюдений рекомендован для ретроспективных исследований.

Перспективами дальнейшей разработки тематики могут стать исследования, направленные на включение в модели оценивания интенсивности пуассоновского процесса дополнительных параметров, интересующих исследователя, а также учета дополнительных сведений, полученных от экспертов. Еще одним направлением может стать использование в методах оценивания интенсивности поведения синтезированных по данным моделей, а также более детальный подход к созданию синтетических данных для таких задач.

Полученные результаты соответствуют специальности 2.3.1 – «Системный анализ, управление и обработка информации, статистика».

ОСНОВНЫЕ ПУБЛИКАЦИИ СОИСКАТЕЛЯ ПО ТЕМЕ ДИССЕРТАЦИИ

Монографии:

1. Тулупьев, А.Л. Мягкие вычисления и измерения. Модели и методы: монография / А.Л. Тулупьев, Т.В. Тулупьева, А.В. Суворова, М.В. Абрамов, А.А. Золотин, М.А. Зотов, А.А. Азаров, Е.А. Мальчевская, Д.Г. Левенец, **А.В. Торопова**, Н.А. Харитонов, А.И. Бирилло, Р.И. Сольницев, С.В. Микони, С.П. Орлов, А.В. Толстов; под ред. д.т.н., проф. С.В. Прокопчиной. — М.: ИД «Научная библиотека», 2017. — 3 т. — 300 с.

Статьи, опубликованные в журналах из перечня рецензируемых научных изданий, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук:

2. Торопова, А.В. Байесовские сети доверия: инструменты и использование в учебном процессе / А.В. Торопова // Компьютерные инструменты в образовании. — №4. — 2016. — С. 43–53.

3. Торопова, А.В. Подходы к диагностике согласованности данных в байесовских сетях доверия / А.В. Торопова // Труды СПИИРАН. — 2015. — № 6(43) — С. 156–178.

4. Торопова, А.В. Машинное обучение байесовской сети доверия как инструмента оценки интенсивности процесса по данным из социальной сети / **А.В. Торопова**, М.В. Абрамов, Т.В. Тулупьева // Научно-технический вестник информационных технологий, механики и оптики. — 2021. — Т. 21. — № 5. — С. 727–737. — Doi: 10.17586/2226-1494-2021-21-5-727-737.

5. Столярова, В.Ф. Модель для оценки частоты публикации постов в онлайн социальной сети по неполным данным с учетом объективных детерминант поведения / В.Ф. Столярова, **А.В. Торопова**, А.Л. Тулупьев // Нечеткие системы и мягкие вычисления. — 2021. — Т. 16. — № 2. — С. 77–95. — Doi: 10.26456/fssc81.

Статьи, опубликованные в изданиях WoS/Scopus:

6. Toropova, A.V. Discretization of a Continuous Frequency Value in a Model of Socially Significant Behavior / **A.V. Toropova**, T.V. Tulupyeva // 2022 XXV International Conference on Soft Computing and Measurements (SCM). — St. Petersburg, Russia. — 2022. — P. 28–30. — Doi: 10.1109/SCM55405.2022.9794892.

7. Toropova, A.V., Comparison of Behavior Rate Models Based on Bayesian Belief Network / **A.V. Toropova**, T.V. Tulupyeva // Recent Research in Control Engineering and Decision Making. ICIT 2020. — Studies in Systems, Decision and Control. —2021. — Vol 337. — Springer, Cham. — Doi: 10.1007/978-3-030-65283-8_42.

8. Toropova, A.V. Approbation of the behavior rate model with hidden variables based on respondents' data on recent Instagram posts / **A.V. Toropova**, T.V. Tulupyeva // 2021 XXIV International Conference on Soft Computing and Measurements (SCM). — St. Petersburg, Russia. — 2021. — P. 43–45. — Doi: 10.1109/SCM52931.2021.9507171.

9. Toropova, A.V. Testing Behavior Rate Models on data from Vk.com Social Network / **A.V. Toropova**, T.V. Tulupyeva // CEUR Workshop Proceedings. Russian Advances in Fuzzy Systems and Soft Computing: Selected Contributions to the 8th International Conference on “Fuzzy Systems, Soft Computing and Intelligent Technologies (FSSCIT 2020)”. — Smolensk, Russia. — July, 2020. — Vol. 2782. — P. 258–263.

10. Toropova, A.V. Bayesian Belief Network as a Behavior Intensity Rate Model on the Example of Posting in a Social Network / **A.V. Toropova**, T.V. Tulupyeva // 2020 XXIII IEEE International Conference on Soft Computing and Measurements (SCM). — St. Petersburg, Russia. — 2020. — P. 22–24. — Doi: 10.1109/SCM50615.2020.9198795.

11. Toropova, A.V. Learning Behavior Rate Models on Social Network Data / **A.V. Toropova**, T.V. Tulupyeva // CEUR Workshop Proceedings. Selected Contributions of the "Russian Advances in Artificial Intelligence" Track at RCAI 2020 co-located with 18th Russian Conference on Artificial Intelligence. — Moscow, Russia. — October 10-16, 2020. — Vol. 2648. — P. 200–209.

12. Toropova, A.V. Synthesis and learning of socially significant behavior model with hidden variables / **A.V. Toropova**, T.V. Tulupyeva // Advances in Intelligent Systems and Computing. — 2019. — V. 875. — P. 76–84.

13. Toropova, A.V. Data Coherence Diagnosis in BBN Risky Behavior Model / A.V. Toropova // Proceedings of the First International Scientific Conference «Intelligent Information Technologies for Industry» (IITI’16). — Springer International Publishing. — 2016. — P. 95–102.

14. Toropova, A.V. Analysis of socially significant behavior model with hidden variables / **A.V. Toropova**, A.V. Suvorova // 2017 XX IEEE International Conference on Soft Computing and Measurements (SCM). — IEEE. — 2017. — P. 50–53.

15. Toropova, A.V. Data coherence diagnosis in socially significant behavior model / **A.V. Toropova**, A.V. Suvorova // 2016 XIX IEEE International Conference on Soft Computing and Measurements (SCM). — IEEE. — 2016. — P. 14–17.

16. Toropova, A.V. Evidence coherence estimation in risky behavior / **A.V. Toropova**, A.V. Suvorova, T.V. Tulupyeva // Soft Computing and Measurements (SCM), 2015 XVIII International Conference. — 2015. — IEEE Conference Publications. — P. 27–29. — Doi: 10.1109/SCM.2015.7190401.

Зарегистрированные программы для ЭВМ:

17. Торопова, А.В. Программа для диагностики согласованности исходных данных в модели социально-значимого поведения (Input Data Coherence Diagnostics in Behavior Model, Version 01 (IDCDiBM v.01)) / **А.В. Торопова**, Р.Р. Хайбуллин, А.В. Суворова, А.Л. Тулупьев. — Свидетельство о гос. Регистрации пр. для ЭВМ № 2018615722. — 2018.

Кроме того, были опубликованы 23 доклада и тезиса на научных мероприятиях (из которых 9 единоличных) и 1 статья. Полный перечень публикаций соискателя по теме исследования представлен в приложении Б диссертационной работы.

Автореферат диссертации

ТОРОПОВА
Александра Витальевна

МЕТОДЫ И АЛГОРИТМЫ ОБРАБОТКИ НЕОПРЕДЕЛЕННОСТИ ДАННЫХ ПРИ ОЦЕНИВАНИИ ИНТЕНСИВНОСТИ ПУАССОНОВСКОГО ПРОЦЕССА ПО ОГРАНИЧЕННОМУ ОБЪЕМУ ДОСТУПНЫХ НАБЛЮДЕНИЙ

Текст автореферата размещен на сайтах:
Высшей аттестационной комиссии при Министерстве науки и высшего образования Российской Федерации
https://vak.minobrnauki.gov.ru/
Федерального государственного бюджетного учреждения науки «Санкт-Петербургский Федеральный исследовательский центр Российской академии наук»
http://www.spiiras.nw.ru/dissovet/

Подписано в печать "29" сентября 2022 г.
Формат 60x84 1/16. Бумага офсетная. Печать офсетная.
Усл.печ.л. 1,0. Тираж 100 экз.
Заказ № ___