

КОМПЛЕКСНЫЙ ПОДХОД К ФОРМИРОВАНИЮ НАБОРА ДАННЫХ ДЛЯ ПРОГНОЗИРОВАНИЯ УЩЕРБА ОТ ЧРЕЗВЫЧАЙНЫХ СИТУАЦИЙ И ОПТИМИЗАЦИИ СИЛ И СРЕДСТВ

Рыбаков А.В., Постернак Е.В.

Цель. Разработать единый интегрированный набор данных для прогнозирования ущерба от чрезвычайных ситуаций и последующего анализа необходимых сил и ресурсов. Сформировать комплексный подход, позволяющий преодолеть разрозненность исходных сведений об ущербе, антропогенных показателях и задействованных инженерно-технических мер (далее – ИТМ) (сил и средств).

Методы. В работе использовалась последовательная схема агрегирования и предварительной обработки данных из различных источников, включающая фильтрацию пропусков и унификацию форматов. Для восполнения недостающих сведений и повышения репрезентативности выборки применялись алгоритмы синтетической генерации на основе вариационных автокодировщиков, копул и генеративно-состязательных сетей. На заключительном этапе введена простая нейронная сеть, которая по известному значению ущерба дополняет отсутствующие данные о силах и средствах.

Результаты. Эксперименты показали, что наилучший баланс между точностью и сохранением исходного распределения признаков продемонстрировала модель CTGAN, достигшая высших показателей формы распределения столбцов и общего балла качества данных. Применение нейронного модуля с функцией активации ReLU в выходном слое позволило избежать отрицательных значений при прогнозировании сил и ресурсов. Совокупное использование синтетической генерации и нейронных предсказаний увеличило полноту итогового набора данных, сохранив статистическую согласованность признаков. Итоговый массив данных дает более детальное представление о связях между ущербом, социально-демографическими характеристиками территории и объемами ИТМ, что обеспечивает возможности для дальнейших исследований по повышению эффективности ликвидации последствий ЧС.

Область применения исследований. Сформированный набор данных может быть использован в качестве обучающей базы для моделей машинного обучения, ориентированных на прогнозирование материальных потерь, а также для разработки систем поддержки принятия решений при планировании и распределении ресурсов в чрезвычайных ситуациях. Результаты исследования представляют научно-практический интерес для специалистов в области управления рисками, а также для разработчиков информационных систем, интегрирующих аналитические модули по оценке ущерба и оптимизации сил и средств.

Ключевые слова: чрезвычайные ситуации, синтетические данные, антропогенные показатели, нейронная сеть, ущерб, силы и средства, машинное обучение.

(Поступила в редакцию 25 марта 2025 г.)

Введение

Чрезвычайные ситуации (далее – ЧС) формируются под воздействием природных и антропогенных факторов и зачастую сопровождаются серьезными социально-экономическими последствиями. Для оценки возможного ущерба необходимо учитывать широкий круг показателей, отражающих характеристики происшествия (например, сила ЧС), социально-демографический профиль пострадавшей территории (численность и плотность населения, доля детей, пожилых, безработных), а также фактические потери (погибшие, пострадавшие, разрушения) [1]. Однако такие сведения обычно представлены в отдельных массивах, где отсутствует сквозная связь с данными о задействованных силах и используемых средствах [2].

Разрозненность исходных материалов затрудняет сопоставительный анализ и генерацию целостного набора данных, пригодного для последующего применения в задачах прогнозирования [3]. Нередко наблюдаются несоответствия во временных промежутках, форматах и классификациях, а также пропуски и дублирующие записи [4]. Отсутствие унифицированной базы с корректными значениями основных показателей не позволяет оперативно оценивать возможные масштабы ущерба и потребности в ресурсах [5]. В результате повышается риск принятия решений, основанных на неполной либо неточной информации.

Цель настоящего исследования состоит в создании интегрированного набора данных, включающего как параметры, описывающие характеристики ЧС (страна, год, сила происшествия и др.), так и социально-демографические и экономические показатели, а также сведения об ущербе, задействованных силах и средствах [5]. Сопоставление показателей масштабов трагедий и задействованных мер защиты на основе корректной аналитической базы способно существенно повысить точность прогнозов и снизить ущерб за счет более взвешенного планирования [6; 7]. Итоговый набор данных должен быть пригоден для последующего обучения и тестирования моделей машинного обучения, специализирующихся на прогнозировании ущерба и определении оптимального состава инженерно-технических мер ИТМ.

Для достижения сформулированной цели необходимо:

- систематизировать разрозненные источники, содержащие сведения о произошедших ЧС, с учетом различий в формате, глубине и качестве первичных данных;
- агрегировать ключевые параметры: сила ЧС, временные рамки, социально-демографические характеристики, фактические показатели ущерба, а также данные о применяемых силах и средствах;
- применить методы генерации синтетических данных для дополнения пропусков и увеличения репрезентативности выборки при сохранении статистических свойств исходных массивов;
- проверить согласованность, полноту и достоверность сформированного набора, используя описательные статистики и сравнение распределений признаков, чтобы обеспечить корректность будущей аналитики;
- подготовить итоговую структуру набора данных, которая позволит исследователям в дальнейшем разрабатывать и обучать различные модели прогнозирования ущерба и планирования реагирования.

Методы машинного обучения и искусственного интеллекта демонстрируют высокую эффективность, если исходные данные обладают достаточной полнотой и репрезентативностью [8]. Следовательно, создание обобщенного набора данных с включением генерированных синтетических наблюдений является востребованным условием для разработки передовых технологий анализа и предупреждения последствий ЧС. Таким образом, формирование качественной и унифицированной базы данных становится определяющим фактором успешной интеграции современных аналитических инструментов в практику управления рисками [9].

Основная часть

Обзор литературы. Прогнозирование ущерба от ЧС основано на анализе разнообразных факторов, среди которых выделяют характеристики происшествия (тип, масштаб, географическое положение), социально-экономические параметры затронутой территории (численность населения, уровень экономического развития) и структуру инфраструктуры [10]. В современных исследованиях наблюдается тенденция к интеграции нескольких методологических направлений [11]:

- классические статистические методы. К ним относят регрессионные модели и временные ряды, применяемые к историческим данным для выявления закономерностей в распределении и динамике ущерба. Однако в ряде случаев они оказываются недостаточно гибкими, особенно при наличии пропусков и неоднородных признаков;
- модели машинного обучения. Сюда входят методы на основе деревьев решений (Random Forest, Gradient Boosting), нейронные сети различной архитектуры, а также байесовские модели. Их преимущество заключается в способности работать с большими массивами данных и находить нетривиальные зависимости, однако при недостатке репрезентативных наблюдений или наличии значительных пропусков результат может оказаться искаженным;
- комбинированные подходы и экспертные системы. В ряде случаев прогнозирование ущерба дополняется экспертными оценками и контекстными знаниями, особенно если фактическая база данных неполна либо фрагментирована. Данное решение повышает точность и надежность выводов, но требует согласованного подхода к верификации экспертных предположений.

Для успешного применения аналитических и прогностических методов необходима высокая степень полноты и структурированности исходных данных. При сильной неоднородности информации, а также отсутствии структуры, отражающей ключевые параметры ЧС, аналитические модели не достигают требуемого уровня точности. Поэтому возникает

потребность в формировании согласованного и насыщенного набора данных, который станет базисом для корректного применения соответствующих математических инструментов.

Синтетическая генерация данных представляет собой процесс создания искусственных наблюдений, сохраняющих статистические закономерности исходной выборки [12]. Целью данного метода является расширение существующего набора данных и восполнение пропусков без утраты основных характеристик распределения. В контексте анализа ЧС это актуально, поскольку реальные события могут встречаться относительно редко и сопровождаться частичными записями.

Обучаемые вероятностные модели используют алгоритмы, моделирующие взаимосвязь между признаками на основе подходов вроде байесовских сетей или моделей на основе копул. Например, Gaussian Copula позволяет описать совместное распределение переменных при условии некоторого набора допущений о корреляционных связях.

Генеративно-сопоставительные сети (далее – CTGAN) строятся на взаимодействии двух нейронных сетей – генератора и дискриминатора. Генератор пытается воспроизвести признаки исходных данных, а дискриминатор – различить реальные и синтетические экземпляры. Такие модели могут работать с категориальными и числовыми признаками, что делает их перспективными для задач, связанных с многокомпонентными данными по ЧС.

Готовые фреймворки, такие как Synthetic Data Vault (далее – SDV), предлагают набор алгоритмов и утилит для автоматизированной генерации синтетических данных, в том числе для сложных разнородных наборов [13].

Основная проблема при синтетической генерации состоит в контроле качества полученных выборок. Кроме проверки сохранения базовых статистик (средние значения, дисперсии, корреляции) особое внимание уделяется отсутствию копирования реальных записей. Выявить такие ошибки можно через метрики сравнения распределений или визуальный анализ многомерных проекций. Успешное применение подобных методов обеспечивает восполнение пропусков и увеличение общего объема выборки, не нарушая согласованности исходных данных.

Правильное распределение ИТМ при возникновении ЧС является ключевым условием эффективной ликвидации последствий [6]. Задача оптимизации включает выбор и расстановку персонала, техники, финансовых и материальных средств в условиях ограниченного времени и часто неполной информации. Современный подход к планированию опирается:

- на математические модели распределения ресурсов. Наиболее известны методы линейного программирования, динамического программирования и теории игр, где формализуются связи между потребностями и возможностями различных подразделений, а также учитываются эвакуационные маршруты и логистические схемы [14];

- системы поддержки принятия решений. Пакеты, основанные на экспертных правилах, моделируют вероятные сценарии развития ЧС и показывают, как корректировать набор задействованных ресурсов в зависимости от меняющихся условий. Такие системы обычно применяются в интеграции с геоинформационными системами (далее – ГИС) и другими информационными модулями;

- комбинации исторических данных и прогнозных моделей. Предполагается анализ прошлых происшествий для выявления корреляций между масштабами ЧС и необходимым объемом сил. Такая стратегия эффективна лишь при условии, что собранные данные содержат достаточное количество сопоставимых записей о ресурсах и последствиях.

Таким образом, ключевым условием для успешного моделирования и оптимизации мер реагирования становится наличие систематизированной и достоверной информации об ущербе, особенностях ЧС и антропогенных характеристиках территории. При отсутствии систематизированной информации о ранее задействованных ресурсах и полученном результате (включая потери и эффект от мер реагирования) невозможно строить надежные прогнозы и давать эффективные рекомендации. Поэтому корректное формирование набора данных, описывающих ЧС вместе с показателями ущерба и ресурсного обеспечения, служит важной предпосылкой для дальнейших исследований в области оптимизации сил и средств.

Описание исходных данных. Исходные сведения о последствиях ЧС представляют собой совокупность показателей, характеризующих урон, понесенный населением и объектами инфраструктуры. В большинстве случаев основное внимание уделяется величине ущерба, выраженной в финансовых единицах, а также демографическим параметрам, отражающим уязвимость территории. К числу ключевых полей относятся следующие:

- обозначение страны или региона, где произошло событие;
- календарный год (иногда указывается дата или интервал), в течение которого имело место происшествие;
- условный показатель силы ЧС;
- плотность населения в зоне воздействия;
- совокупное число жителей пострадавшей территории;
- численность или доля в общей структуре населения соответствующих возрастных групп;
- показатель безработицы или число граждан, лишенных работы в регионе;
- числовые характеристики последствий: число погибших, раненых, пострадавших (включая эвакуацию), а также оставшихся без жилья;
- итоговая оценка материального ущерба, учитывающая прямые потери и косвенные издержки.

Существенным моментом является то, что в некоторых источниках данные представлены в различных масштабах (город, провинция, страна) или используют неодинаковые критерии выделения социально уязвимых групп. Например, численное значение для пожилого населения может определяться по разным возрастным порогам, что требует согласования или нормализации при объединении.

Правильная структура такого набора данных должна обеспечивать однозначную идентификацию события, а также удобство дальнейшей интеграции с дополнительными признаками. При наличии значительных пропусков в исходных источниках возникает необходимость их логического или статистического восстановления, поскольку отсутствие ключевых переменных (например, точного числа погибших или оценки ущерба) способно существенным образом исказить будущие результаты анализа.

В задаче оценки ЧС важнейшими вспомогательными данными служат сведения об объеме ИТМ. Отсутствие информации затрудняет анализ взаимосвязи ущерба и распределения сил, поскольку невозможно напрямую сопоставить затраченные ресурсы и полученный эффект при ликвидации последствий.

Сложность создания целостного набора данных, объединяющего информацию об ущербе, демографической структуре и параметрах сил и средств, связана с несколькими факторами:

- в исходных документах могут применяться разные шкалы для оценки силы ЧС, а категории ресурсов и формирований также не всегда согласованы. Для корректного сопоставления необходима единая система кодов или механизм сопоставления, способный учесть все несоответствия;
- даже после согласования понятий часто остаются пропущенные значения, особенно если документы составлялись в условиях нехватки времени и недостаточного взаимодействия ведомств. В результате формируется явная или скрытая несбалансированность по ряду признаков: одни события представлены более детально, другие – фрагментарно;
- организации, ведающие статистикой ЧС, могут использовать собственные форматы хранения данных, не предусматривая централизованной передачи или публикации. Такая автономность затрудняет доступ и интерпретацию, а автоматическое объединение требует дополнительной валидации.

Преодоление данных трудностей требует проведения тщательной проверки качества собранной информации, унификации форматов, а при необходимости и применения статистических методов восполнения пропусков. Лишь после выстраивания единого формата данных возможен корректный переход к созданию комплексного набора данных, пригодного для решения задач прогнозирования ущерба и оптимизации объемов ИТМ.

Методика формирования единого набора данных. Процесс объединения разрозненных данных по ущербу, антропогенным характеристикам и сведениям об объемах ИТМ предполагает выполнение нескольких этапов:

- каждая запись о ЧС в наборе с ущербом должна быть связана с соответствующими записями о ресурсах и структуре сил, если такие записи существуют. В качестве первичных ключей обычно выступают географические признаки (страна или регион) и временные метки (год или более детальная дата). Когда точные соответствия отсутствуют, применяются эвристические методы (например, близость по дате или региону);

– выделяются единые справочники для кодирования типов ЧС ресурсов и категорий пострадавшего населения. Все поля приводятся к согласованному формату (например, единые единицы измерения ущерба, указания возраста для групп детей и пожилых);

– неисправимые записи с экстремальными либо противоречивыми значениями удаляются или корректируются, если их незначительное количество не оказывает существенного влияния на полноту. В противном случае детально анализируется первоисточник, чтобы уточнить природу расхождений;

– при наличии отсутствующих значений по отдельным демографическим или социальным признакам выполняется их аппроксимация на основе средних или медианных величин либо же применяются более сложные методы регрессии и интерполяции. Для критически важных полей, в частности ущерба или силы ЧС, допускается исключение записей с недостающими данными, если их доля мала.

Результатом агрегирования становится единый массив, где каждая строка или запись отражает конкретный случай ЧС с систематизированной информацией о месте, времени, параметрах происшествия, размере ущерба и основных социально-демографических показателях. Алгоритм объединения данных приведен на рисунке 1.

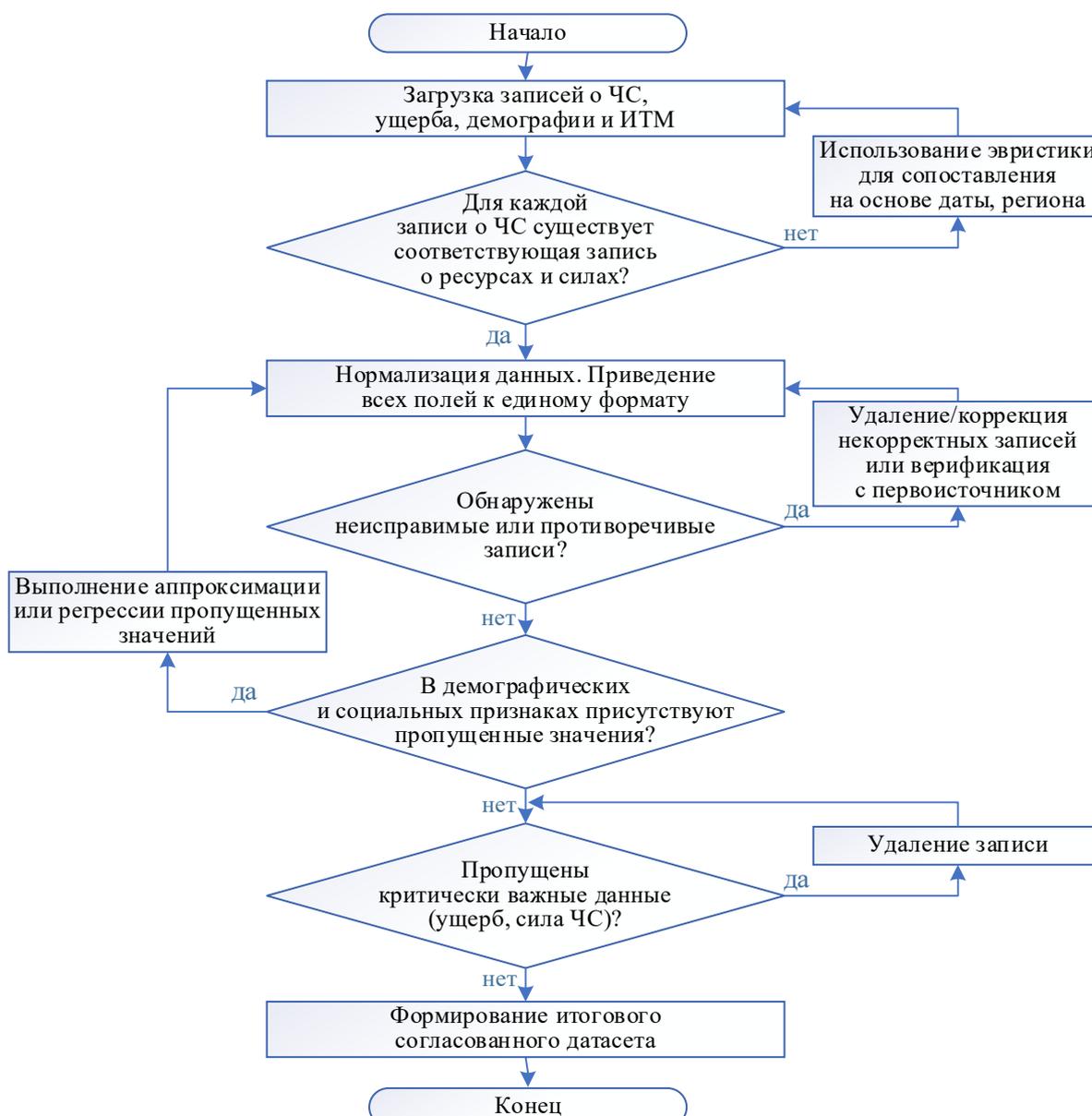


Рисунок 1. – Алгоритм объединения записей о ЧС (показатели ЧС, ущерб, антропогенные характеристики, объемы ИТМ)

При наличии существенных пробелов в реальной статистике либо при недостаточном объеме данных для корректного обучения будущих моделей прогнозирования прибегают к генерации синтетических записей [15; 16]. В настоящем исследовании для этой цели применяется пакет SDV, включающий несколько алгоритмов генерации, в том числе:

- копула модели;
- генеративные сети;
- механизмы проверки качества синтетических данных.

Алгоритм работы с SDV в задаче формирования набора данных для ЧС включает загрузку исходных таблиц, настройку выбранной модели, обучение и генерацию требуемого количества записей. Полученный синтетический массив либо объединяется с реальными данными, восполняя пропуски, либо формирует расширенную выборку для более детализированного анализа. Особое внимание уделяется проверке консистентности с точки зрения характерных интервалов и логических ограничений (например, финансовые затраты не должны быть отрицательными). Алгоритм генерации синтетических данных по ЧС приведен на рисунке 2.

Помимо синтетического восполнения пропусков по демографическим показателям и ущербу, в ряде случаев возникает потребность в уточнении или генерации данных, относящихся к силам и средствам [17]. Поскольку эти параметры часто фиксируются в неполном объеме либо представлены только для отдельных происшествий, для их корректного заполнения используется нейронная модель, обучаемая на доступных совмещенных записях [18].

В рамках данного исследования не создается модель для финального прогнозирования ущерба. Однако для оценки взаимосвязей между признаком ущерба и ИТМ можно использовать многослойный перцептрон или другой простой нейронный подход [19]. Обучение происходит на подмножестве событий, где зафиксированы достоверные данные по силам и средствам.

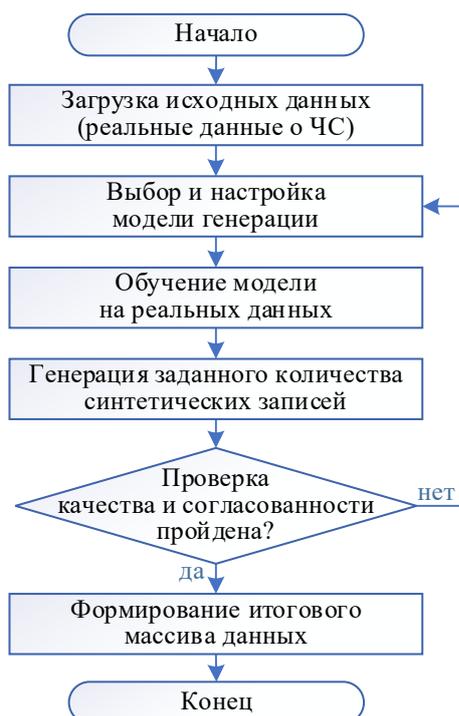


Рисунок 2. – Алгоритм генерации синтетических данных по ЧС

и предсказанные данные, что облегчает дальнейшее построение и тестирование более сложных моделей прогнозирования.

Архитектура и реализация нейронной модели. Формирование синтетических данных для прогнозирования ущерба от чрезвычайных ситуаций и сопутствующих показателей потребовало последовательной проверки нескольких подходов к генерации. Первоначально опробовались модели на основе вариационных автокодировщиков (далее – TVAE) и копул (далее – Gaussian Copula), а также CTGAN.

Входными признаками служат материальные показатели (плотность населения, общее число пострадавших, совокупный ущерб и т.д.), а целевыми переменными выступают числовые значения сил и ресурсов. Модель оптимизируется с помощью стандартных процедур (например, Adam), а точность оценивается на основе метрик для регрессии (MAE, RMSE) [20–22].

После достижения удовлетворительной точности на контрольной выборке модель применяется для событий, в которых поля силы и средства отсутствуют. Предсказанные значения записываются в соответствующие столбцы, формируя более полную картину о распределении необходимых мер реагирования. Алгоритм восстановления пропущенных значений по ИТМ приведен на рисунке 3.

Использование нейронного метода дополнения сил и ресурсов придает формируемому набору данных дополнительные возможности: анализ взаимосвязи между последствиями ЧС и привлеченными ресурсами становится доступным даже для тех записей, которые первоначально были неполными. В итоге получается расширенный массив, включающий как реальные, так

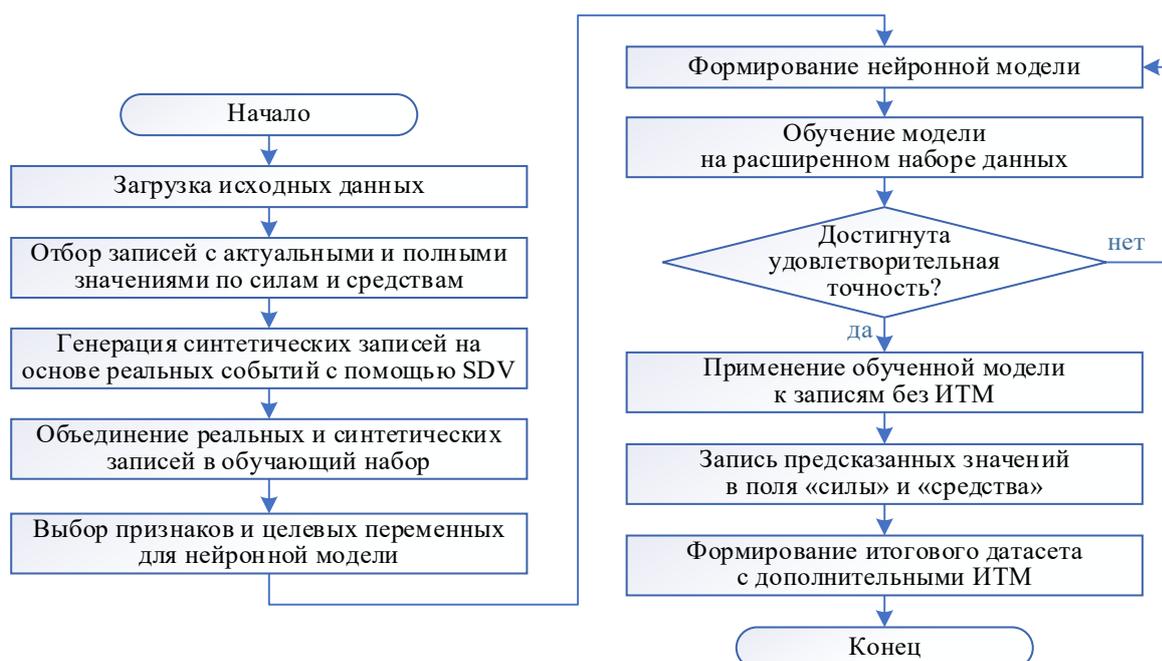


Рисунок 3. – Алгоритм восстановления пропущенных объемов ИТМ

Для работы использовались исходные данные МЧС России¹, данные с ресурсов Статистической базы данных ЕЭК ООН², данные по ЧС с ресурса «Центр исследований эпидемиологии катастроф»³:

Для оценки моделей использовались показатели:

- оценка формы распределения столбцов – сравнивает распределение каждого отдельного столбца в синтетических данных с оригинальным;
- общий балл качества данных – обобщенная оценка качества всех сгенерированных данных.

Состав набора данных для генерации синтетических данных с антропогенными показателями:

- код страны;
- сила ЧС;
- год ЧС;
- плотность населения;
- численность;
- численность детей;
- численность пожилого населения;
- процент безработицы;
- количество погибших при ЧС;
- количество раненых при ЧС;
- количество пострадавших при ЧС;
- количество людей, которые лишились жилья;
- ущерб.

Результаты экспериментов показывают, что наибольшее среднее качество продемонстрировал CTGAN с параметром эпох 4000. Данная конфигурация достигла оценки формы распределения столбцов 78,65 % и общего балла качества данных 79,58 %, что превосходит показатели альтернативных решений. Подробные результаты приведены в таблице 1.

¹ Министерство Российской Федерации по делам гражданской обороны, чрезвычайным ситуациям и ликвидации последствий стихийных бедствий. Итоги деятельности МЧС России. – URL: <https://mchs.gov.ru/deyatelnost/itogi-deyatelnosti-mchs-rossii> (Дата обращения: 25.03.2025).

² Европейская экономическая комиссия ООН. Статистическая база данных ЕЭК ООН. – URL: <https://w3.unece.org/PXWeb/ru/> (Дата обращения: 25.03.2025).

³ Центр исследований эпидемиологии катастроф. Международная база данных о чрезвычайных ситуациях EM-DAT. – URL: <https://www.emdat.be/> (Дата обращения: 25.03.2025).

Таблица 1. – Оценка качества моделей для генерации синтетических данных по ущербу

Модель	Количество эпох, шт.	Оценка формы распределения столбцов, %	Общий балл качества данных, %
TVAE	1000	53,99	46,92
TVAE	4000	49,61	43,35
Gaussian Copula	1000	76,47	77,91
Gaussian Copula	4000	76,50	77,92
CTGAN	1000	77,03	78,98
CTGAN	4000	78,65	79,58

Аналогичный процесс применялся к расширенному массиву, включавшему параметры ущерба, сил и средств. И в этом случае лучший результат достигался при использовании CTGAN с 4000 эпохами обучения. Благодаря таким показателям стало возможным достоверно восполнить пропущенные значения и расширить объем обучающих данных для последующего анализа. Подробные результаты приведены в таблице 2.

Таблица 2. – Оценка качества моделей для генерации синтетических данных по силам и средствам

Модель	Количество эпох, шт.	Оценка формы распределения столбцов, %	Общий балл качества данных, %
TVAE	4000	44,76	42,11
Gaussian Copula	4000	75,23	74,22
CTGAN	4000	88,15	90,10

Дополнительно была протестирована простая нейронная модель для генерации сил и средств на основе ущерба. В качестве базовой структуры выбрано четыре полносвязных слоя, где итоговый слой формирует вектор размерности 2. Первоначально применялся линейный выходной слой, что в ряде случаев приводило к отрицательным прогнозам сил и средств. Для устранения этой проблемы итоговый слой был переведен на функцию активации ReLU, тем самым исключив некорректные отрицательные значения. Подобная корректировка структуры продемонстрировала более реалистичные оценки ресурсов, что оправдало выбранный подход.

Перед обучением синтетических моделей, а также при настройке нейронной сети для генерации показателей сил и средств исходные признаки подвергались стандартным процедурам предобработки:

- обработка пропусков и аномальных значений;
- кодирование категориальных переменных;
- нормализация числовых данных;
- разделение записей на обучающие и тестовые.

Применение указанных методов подготовки способствовало выравниванию распределения признаков и улучшило сходимость моделей. Это, в свою очередь, обеспечило более стабильные результаты на этапе синтетической генерации и при обучении нейронной сети, восстанавливающей недостающие показатели сил и средств по известному значению ущерба.

Процесс настройки синтетических генераторов включал серию экспериментов с выбором гиперпараметров:

- количество эпох варьировалось от 1000 до 4000. Оптимальным оказался верхний предел, который позволил достичь максимальных значений метрик качества;
- `batch_size` принималось равным 500 записям;
- прочие внутренние настройки (например, размер скрытого пространства или применяемый оптимизатор) подбирались автоматически согласно средствам SDV.

Следующие параметры определены для нейронной модели:

- архитектура состояла из четырех полносвязных слоев, в том числе входного, двух скрытых и выходного. Размерность скрытых слоев выбиралась в диапазоне 32–64 нейронов;
- функция активации. Первоначально последний слой имел линейную активацию, однако переход на ReLU позволил избежать отрицательных значений прогнозируемых сил и средств;
- оптимизатор Adam, показатель `learning_rate` оставался по умолчанию (0,001);
- функция потерь `mse` (средняя квадратичная ошибка) в сочетании с метрикой `mae` (средняя абсолютная ошибка) для интерпретации точности;

- количество эпох для обучения равно 500, т.к. более длинные циклы не дали существенного прироста качества;
- `batch_size` равен 8, что соответствовало относительно небольшим объемам данных и способствовало лучшей сходимости.

Итоговая схема нейронной сети для восстановления ИТМ приведена на рисунке 4.

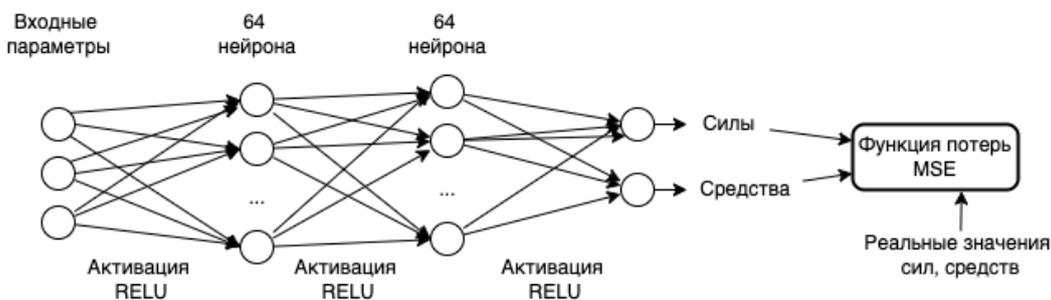


Рисунок 4. – Нейронная сеть для восстановления ИТМ

Результаты и их обсуждение. В ходе экспериментов по формированию синтетических данных для прогнозирования ущерба и сопутствующих показателей рассмотрено несколько подходов, основанных на различных механизмах генерации: вариационных автокодировщиках, копулах и генеративно-сопоставительных сетях. Наилучшую эффективность в целом показала модель CTGAN при 4000 эпохах обучения [23]. Достигнутая точность (79,58 % для набора с антропогенными показателями и 90,10 % для набора, включающего силы и ресурсы) свидетельствует о том, что полученные синтетические записи в наибольшей степени сохраняют статистическое распределение исходных данных и позволяют восполнить пропущенные значения без существенных искажений.

Дополнительно была реализована процедура восстановления недостающих сведений о задействованных силах и ресурсах, основанная на использовании простого нейронного регрессионного модуля. Поскольку в исходных источниках эти данные нередко отсутствовали либо имелись в фрагментарном виде, применялась следующая схема:

- входным признаком выступал показатель ущерба, а выходной слой формировал вектор из двух величин (силы и средства). Такой выбор оправдан тем, что объем постигнутого региона ущерба часто прямо или косвенно коррелирует с масштабом задействованных мер реагирования;
- предварительные эксперименты с линейной активацией выявили проблему отрицательных прогнозируемых значений. Замена функции активации в выходном слое на ReLU устранила эту некорректность, обеспечив реалистичные (неотрицательные) оценки;
- для анализа используемых гиперпараметров модели и корректировки процесса обучения были применены стандартные регрессионные метрики (средняя квадратичная ошибка и средняя абсолютная ошибка). Процедура дополнительной фильтрации помогла избавиться от аномальных случаев, когда предсказанные ресурсы значительно превосходили наблюдавшийся диапазон.

Таким образом, совокупное применение CTGAN (для базового формирования синтетических данных) и нейронной сети (для дообогащения пропущенных полей) позволило создать целостный набор данных, в котором сведены воедино социально-демографические показатели, финансовые потери и необходимые ресурсы реагирования. Полученный массив данных впоследствии может стать базой для обучающих выборок в задачах прогнозирования ущерба и планирования мер по ликвидации последствий ЧС.

Заключение

В ходе работы выполнено создание единого набора данных для дальнейшего использования в задачах прогнозирования ущерба от чрезвычайных ситуаций и сопутствующего анализа необходимых сил и средств. Были учтены особенности разрозненных исходных источников и реализованы следующие шаги:

- агрегация и предобработка фрагментированных данных, что позволило выявить пропуски, несогласованные форматы и неочевидные дубликаты;

– применение различных моделей синтетической генерации (TVAE, Gaussian Copula, CTGAN) и выбор лучшей конфигурации, обеспечивающей наивысшие показатели качества (CTGAN при 4000 эпохах);

– использование нейронной сети для восстановления или уточнения полей, связанных с силами и ресурсами, посредством регрессионных предсказаний по ущербу;

– проверка итогового набора данных с точки зрения сохранения статистических распределений и адекватности полученных значений по ключевым метрикам.

Сформированная база данных открывает перспективы для дальнейших исследований, связанных:

– с разработкой и тестированием полноценных моделей машинного обучения, способных прогнозировать ущерб по сочетанию социально-демографических, инфраструктурных и метеорологических признаков;

– оценкой эффективности планирования мероприятий по защите населения и территорий за счет выявления факторов, наиболее существенно влияющих на показатели ущерба;

– созданием комплексных систем поддержки принятия решений, учитывающих вероятностные сценарии развития ЧС и оптимизирующих распределение сил и ресурсов в условиях ограниченного времени и финансирования.

Реализация представленных подходов и методических рекомендаций будет способствовать совершенствованию управленческих практик в сфере защиты от чрезвычайных ситуаций и повышению уровня защищенности населения за счет более точной и комплексной оценки потенциальных рисков [24].

ЛИТЕРАТУРА

1. Авдоткин, В.П. Оценка ущерба от чрезвычайных ситуаций природного и техногенного характера: монография / В.П. Авдоткин, М.М. Дзыбов, К.П. Самсонов; МЧС России. – М.: ФГБУ ВНИИ ГОЧС (ФЦ), 2012. – 468 с. – ISBN 978-5-93970-082-5. – EDN: PXNONB.
2. Liu, J. The latest progress of data fusion for integrated disaster reduction and intelligent emergency response / J. Liu, M. Konečný, Q. Du, Sh. Xu, F. Ren, X. Che // International Journal of Image and Data Fusion. – 2021. – Vol. 12, No. 4. – P. 265–267. – DOI: 10.1080/19479832.2021.1970931.
3. Горбунов, С.В. Анализ технологий прогнозирования чрезвычайных ситуаций природного и техногенного характера / С.В. Горбунов, Ю.Д. Макиев, В.П. Малышев // Стратегия гражданской защиты: проблемы и исследования. – 2011. – № 1 (1). – С. 43–53. – EDN: OJWKOF.
4. Джумакулова, К.А. Проблемы мониторинга и прогнозирования чрезвычайных ситуаций в республике Узбекистан и пути их решений // Экономика и социум. – 2024. – № 12-1 (127). – 11 с. – URL: <https://cyberleninka.ru/article/n/problemy-monitoringa-i-prognozirovaniya-chrezvychaynyh-situatsiy-v-respublike-uzbekistan-i-puti-ih-resheniy> (дата обращения: 24.03.2025). – EDN: OMGTEJ.
5. Noy, I. Extreme events impact attribution: A state of the art / I. Noy, D. Stone, T. Uher // Cell Reports Sustainability. – 2024. – Vol. 1, No. 5. – Article 100101. – 9 p. – DOI: 10.1016/j.crsus.2024.100101.
6. Фалеев, М.И. Глобальные климатические изменения – фактор активизации природных и антропогенных вызовов населению и окружающей среде / М.И. Фалеев, Н.А. Цыбиков, Т.И. Сидорович // Технологии гражданской безопасности. – 2022. – Т. 19, № 2 (72). – С. 4–10. – DOI: 10.54234/CST.19968493.2022.19.2.72.1.4. – EDN: VCELAQ.
7. Bolan, S. Impacts of climate change on the fate of contaminants through extreme weather events / S. Bolan, L.P. Padhye, T. Jasemizad [et al.] // Science of The Total Environment. – 2024. – Vol. 909. – Article 168388. – DOI: 10.1016/j.scitotenv.2023.168388.
8. Станкевич, Т.С. Разработка метода оперативного прогнозирования динамики развития лесного пожара посредством искусственного интеллекта и глубокого машинного обучения / Т.С. Станкевич // Вестник Иркутского государственного технического университета. – 2018. – Т. 22, № 9 (140). – С. 111–120. – DOI: 10.21285/1814-3520-2018-9-111-120. – EDN: YGGSLR.
9. Steinhäuser, K. Multivariate and multiscale dependence in the global climate system revealed through complex networks / K. Steinhäuser, A.R. Ganguly, N.V. Chawla // Climate Dynamics. – 2012. – Vol. 39, No. 3-4. – P. 889–895. – DOI: 10.1007/s00382-011-1135-9.
10. Шойгу, Ю.С. Прогнозирование и управление социально-психологическими рисками во время чрезвычайной ситуации / Ю.С. Шойгу, Л.Г. Пыжьянова // Вестник Московского университета. Серия 14. Психология. – 2011. – № 4 – С. 76–83. – EDN: OQQWMJ.
11. Godschalk, D.R. Urban hazard mitigation: creating resilient cities / D.R. Godschalk // Natural Hazards Review. – 2003. – Vol. 4, No. 3. – P. 136–143. – DOI: 10.1061/(ASCE)1527-6988(2003)4:3(136).

12. Bauer, A. Comprehensive exploration of synthetic data generation: a survey / A. Bauer, S. Trapp, M. Stenger [et al.]. – arXiv, 2024. – 103 p. – (Preprint / University of Chicago, USA; University of Würzburg, Germany; University of Ulm, Germany; Argonne National Laboratory Lemont, USA; arXiv:2401.02524v2 [cs.LG]). – DOI: 10.48550/arXiv.2401.02524.
13. Zhang, K. Sequential models in the synthetic data vault / K. Zhang, N. Patki, K. Veeramachaneni. – arXiv, 2022. – 17 p. – (Preprint / The Synthetic Data Vault team; arXiv:2207.14406v1 [cs.LG]). – DOI: 10.48550/arXiv.2207.14406.
14. Бухвалов, А.В. Л.В. Канторович и экономико-математическое моделирование: синтез реальности, математики и экономики / А.В. Бухвалов // Российский журнал менеджмента. – 2012. – Т. 10, № 3. – С. 3–30. – EDN: NPVACU.
15. Москвина, Н.В. Применение искусственного интеллекта в системе-112 / Н.В. Москвина // Столыпинский вестник. – 2023. – № 4. – URL: <https://cyberleninka.ru/article/n/primenenie-iskusstvennogo-intellekta-v-sisteme-112> (дата обращения: 24.03.2025). – EDN: ERLYGT.
16. Li, Z. Machine Learning seismic wave discrimination: application to earthquake early warning / Z. Li, M.-A. Meier, E. Hauksson, Z. Zhan, J. Andrews // Geophysical Research Letters. – 2018. – Vol. 45, No. 10. – P. 4773–4779. – DOI: 10.1029/2018GL077870.
17. Ковзель, А.А. Прогнозирование последствий чрезвычайных ситуаций с применением нейросетевых технологий / А.А. Ковзель // Научный Лидер. – 2021. – № 11 (13). – С. 43–45. – URL: <https://scilead.ru/article/183-prognozirovanie-posledstvij-chrezvichajnikh-si> (дата обращения: 24.03.2025). – EDN: GOWGOZ.
18. Рыбаков, А. Нейронные сети для защиты населения и территорий от ЧС / А. Рыбаков, Е. Иванов, В. Нестеров // Системы безопасности. – 2021. – URL: <https://www.secuteck.ru/articles/nejronnye-seti-dlya-zashchity-naseleniya-i-territorij-ot-chs> (Дата обращения: 25.03.2025).
19. Суханова, Н.В. Разработка нейросетевой модели для прогнозирования вероятности землетрясений / Н.В. Суханова // Автоматизация и моделирование в проектировании и управлении. – 2023. – № 2 (20). – С. 40–49. – DOI: 10.30987/2658-6436-2023-2-40-49. – EDN: EDIMKN.
20. Hyndman, R.J. Another look at measures of forecast accuracy / R.J. Hyndman, A.B. Koehler // International Journal of Forecasting. – 2006. – Vol. 22, No. 4. – P. 679–688. – DOI: 10.1016/j.ijforecast.2006.03.001.
21. Chai, T. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature / T. Chai, R.R. Draxler // Geoscientific Model Development. – 2014. – Vol. 7, No. 3. – P. 1247–1250. – DOI: 10.5194/gmd-7-1247-2014.
22. Willmott, C.J. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance / C.J. Willmott, K. Matsuura // Climate Research. – 2005. – Vol. 30, No. 1. – P. 79–82. – DOI: 10.3354/cr030079.
23. Пчелинцев, С. Метод создания синтетических наборов данных для обучения нейросетевых моделей распознаванию объектов / С. Пчелинцев, М. А. Юляшков, О. А. Ковалева // Информационно-управляющие системы. – 2022. – № 3 (118). – С. 9–19. – DOI: 10.31799/1684-8853-2022-3-9-19. – EDN: LBEAQQ.
24. Рыбаков, А.В. Анализ проблемной ситуации в области обоснования объемов инженерно-технических мероприятий при прогнозировании ущерба от природных чрезвычайных ситуаций с учетом антропогенных факторов / А.В. Рыбаков, Е.В. Постернак // Научные и образовательные проблемы гражданской защиты. – 2024. – № 4 (63). – С. 72–83. – EDN: VJANZK.

Комплексный подход к формированию набора данных для прогнозирования ущерба от чрезвычайных ситуаций и оптимизации сил и средств

A comprehensive approach to forming a dataset for forecasting emergency situations damage and optimizing forces and means

Рыбаков Анатолий Валерьевич

доктор технических наук, профессор
Федеральное государственное бюджетное образовательное учреждение высшего образования «Академия гражданской защиты Министерства Российской Федерации по делам гражданской обороны, чрезвычайным ситуациям и ликвидации последствий стихийных бедствий», факультет комплексной безопасности и основ военной подготовки, кафедра высшей математики, профессор

Адрес: ул. Соколовская, 1А,
141435, мкр. Новогорск,
г. Химки, Московская обл., Россия
Email: a.rybakov@agz.50.mhs.gov.ru
SPIN-код: 8654-3788

Anatoliy V. Rybakov

Grand PhD in Technical Sciences, Professor
Federal State Budget Educational Establishment of Higher Education «Civil Defense Academy of the Ministry of the Russian Federation for Civil Defense, Emergencies and Elimination of Consequences of Natural Disasters», Faculty of Comprehensive Safety and Basics of Military Training, Chair of Higher Mathematics, Professor

Address: Sokolovskaya str., 1A,
141435, Novogorsk microdistrict,
Khimki, Moscow region, Russia
Email: a.rybakov@agz.50.mhs.gov.ru
ORCID: 0000-0002-4037-1231

Постернак Евгений Валерьевич

Федеральное государственное бюджетное образовательное учреждение высшего образования «Российский государственный социальный университет», факультет комплексной безопасности и основ военной подготовки, кафедра экологии и природоохранной деятельности, аспирант

Адрес: ул. Вильгельма Пика, 4, стр. 1,
129226, г. Москва, Россия
Email: kripsy93@yandex.ru
SPIN-код: 7581-4305

Evgeniy V. Posternak

Federal State Budget Educational Establishment of Higher Education «Russian State Social University», Faculty of Comprehensive Safety and Basics of Military Training, Chair of Ecology and Environmental Protection Activities, postgraduate student

Address: Wilhelm Pieck str., 4, building 1,
220118, Moscow, Russia
Email: kripsy93@yandex.ru
ORCID: 0009-0003-9732-4449

A COMPREHENSIVE APPROACH TO FORMING A DATASET FOR FORECASTING EMERGENCY SITUATIONS DAMAGE AND OPTIMIZING FORCES AND MEANS

Rybakov A.V., Posternak E.V.

Purpose. To develop a unified, integrated dataset for forecasting damage from emergency situations and for subsequent analysis of the necessary forces and resources. A comprehensive approach is formulated to overcome the fragmentation of initial data on damage, anthropogenic indicators, and deployed engineering and technical measures (forces and means).

Methods. The study employed a sequential scheme for aggregating and preprocessing data from various sources, including filtering out missing values and unifying formats. To fill data gaps and increase sample representativeness, synthetic generation algorithms based on variational autoencoders, copulas, and generative adversarial networks were applied. In the final stage, a simple neural network was introduced, which uses the known damage value to supplement missing information on forces and means.

Findings. Experiments demonstrated that the CTGAN model achieved the best balance between accuracy and preservation of the original feature distribution, attaining the highest column shape scores and overall quality metrics. The application of a neural module with a ReLU activation function in the output layer prevented negative values when predicting forces and resources. The combined use of synthetic generation and neural predictions enhanced the completeness of the final dataset while maintaining statistical consistency among features. As a result, the dataset provides more detailed insights into the relationships between damage, socio-demographic characteristics of the affected area, and the scope of deployed engineering and technical measures, thus enabling further research aimed at improving the efficiency of disaster response.

Application field of research. The resulting dataset can be utilized as a training base for machine learning models focused on predicting material losses, as well as for developing decision support systems for planning and allocating resources in emergency situations. The study findings are of scientific and practical interest to risk management professionals and to developers of information systems that incorporate analytical modules for damage assessment and optimization of forces and means.

Keywords: emergencies, synthetic data, anthropogenic indicators, neural network, damage, forces and means, machine learning.

(The date of submitting: March 25, 2025)

REFERENCES

1. Avdot'in V.P., Dzybov M.M., Samsonov K.P. Otsenka ushcherba ot chrezvychaynykh situatsiy prirodno i tekhnogennogo kharaktera: monografiya [Damage assessment from natural and technogenic emergencies]: monograph; EMERCOM of Russia. Moscow, FGBU VNII GOChS (FC), 2012. 468 p. (rus). ISBN 978-5-93970-082-5. EDN: PNXONB.
2. Liu J., Konečný M., Du Q., Xu Sh., Ren F., Che X. The latest progress of data fusion for integrated disaster reduction intelligence service. *International Journal of Image and Data Fusion*, 2021. Vol. 12, No. 4. Pp. 265–267. DOI: 10.1080/19479832.2021.1970931.
3. Gorbunov S.V., Makiev Yu.D., Malyshev V.P. Analiz tekhnologii prognozirovaniya chrezvychaynykh situatsiy prirodno i tekhnogennogo kharaktera [Analysis of forecasting technologies of natural and man-made emergencies]. *Civil Protection Strategy: Issues & Research*, 2011. No. 1 (1). Pp 43–53. (rus). EDN: OJWKOF.
4. Dzhumakulova K.A. Problemy monitoringa i prognozirovaniya chrezvychaynykh situatsiy v respublike Uzbekistan i puti ikh resheniy [Problems of emergency monitoring and forecasting of the Republic of Uzbekistan and solutions]. *Economy and Society*, 2024. No. 12-1 (127). 11 p. Available at: <https://cyberleninka.ru/article/n/problemy-monitoringa-i-prognozirovaniya-chrezvychaynykh-situatsiy-v-respublike-uzbekistan-i-puti-ih-resheniy> (accessed: March 24, 2025). (rus). EDN: OMGTEJ.
5. Noy I., Stone D., Uher T. Extreme events impact attribution: A state of the art. *Cell Reports Sustainability*, 2024. Vol. 1, No. 5. Article 100101. 9 p. DOI: 10.1016/j.crsus.2024.100101.
6. Faleev M.I., Tsybikov N.A., Sidorovich T.I. Global'nye klimaticheskie izmeneniya – faktor aktivizatsii prirodnykh i antropogennykh vyzovov naseleniyu i okruzhayushchey srede [Global climate change as a factor of natural and anthropogenic challenges activation to the population and the environment]. *Civil Security Technology*, 2022. Vol. 19, No. 2 (72). Pp. 4–10. (rus). DOI: 10.54234/CST.19968493.2022.19.2.72.1.4. EDN: VCELAQ.

7. Bolan S., Padhye L.P., Jasemizad T., Govarathanan M., Karmegam N., Wijesekara H., Amarasiri D., Hou D., Zhou P., Biswal B.K., Balasubramanian R., Wang H., Siddique K., Rinklebe J., Kirkham M.B., Bolan N. Impacts of climate change on the fate of contaminants through extreme weather events. *Science of The Total Environment*, 2024. Vol. 909. Article 168388. DOI: 10.1016/j.scitotenv.2023.168388.
8. Stankevich T.S. Razrabotka metoda operativnogo prognozirovaniya dinamiki razvitiya lesnogo pozhara posredstvom iskusstvennogo intellekta i glubokogo mashinnogo obucheniya [Development of operational prediction method of forest fire dynamics based on artificial intelligence and deep machine learning]. *Proceedings of Irkutsk State Technical University*, 2018. Vol. 22, No. 9 (140). Pp. 111–120. (rus). DOI: 10.21285/1814-3520-2018-9-111-120. EDN: YGGSLR.
9. Steinhäuser K., Ganguly A.R., Chawla N.V. Multivariate and multiscale dependence in the global climate system revealed through complex networks. *Climate Dynamics*, 2012. Vol. 39, No. 3-4. Pp. 889–895. DOI: 10.1007/s00382-011-1135-9.
10. Shoygu Yu.S., Pyzh'yanova L.G. Prognozirovanie i upravlenie sotsial'no-psikhologicheskimi riskami vo vremya chrezvychaynoy situatsii [Foreseeing and managing social and psychological risks during an emergency situation]. *Moscow University Bulletin. Series 14. Psychology*, 2011. No. 4. Pp. 76–83. (rus). EDN: OQQWMJ.
11. Godschalk D.R. Urban hazard mitigation: creating resilient cities. *Natural Hazards Review*, 2003. Vol. 4, No. 3. Pp. 136–143. DOI: 10.1061/(ASCE)1527-6988(2003)4:3(136).
12. Bauer A., Trapp S., Stenger M., Leppich R., Kounev S., Leznik M., Chard K., Foster I. *Comprehensive exploration of synthetic data generation: a survey*. arXiv, 2024. 103 p. Preprint arXiv:2401.02524v2 [cs.LG]. DOI: 10.48550/arXiv.2401.02524.
13. Zhang K., Patki N., Veeramachaneni K. *Sequential models in the synthetic data vault*. arXiv, 2022. 17 p. Preprint arXiv:2207.14406v1 [cs.LG]. DOI: 10.48550/arXiv.2207.14406.
14. Bukhvalov A.V. L.V. Kantorovich i ekonomiko-matematicheskoe modelirovanie: sintez real'nosti, matematiki i ekonomiki [L.V. Kantorovich and economic-mathematical modeling: synthesis of reality, mathematics, and economics]. *Russian Management Journal*, 2012. Vol. 10, No. 3. Pp. 3–30. (rus). EDN: NPVACU.
15. Moskvina N.V. Primenenie iskusstvennogo intellekta v sisteme-112 [Application of artificial intelligence in the system-112]. *Stolypin Bulletin*, 2023. No. 4. Available at: <https://cyberleninka.ru/article/n/primenenie-iskusstvennogo-intellekta-v-sisteme-112> (accessed: March 24, 2025). (rus). EDN: ERLYGT.
16. Li Z., Meier M.-A., Hauksson E., Zhan Z., Andrews J. Machine learning seismic wave discrimination: application to earthquake early warning. *Geophysical Research Letters*, 2018. Vol. 45, No. 10. Pp. 4773–4779. DOI: 10.1029/2018GL077870.
17. Kovzel' A.A. Prognozirovanie posledstviy chrezvychaynykh situatsiy s primeneniem neyrosetevykh tekhnologiy [Forecasting the consequences of emergencies using neural network technologies]. *Nauchnyy Lider*, 2021. No. 11 (13). Available at: <https://scilead.ru/article/183-prognozirovanie-posledstvij-chrezvichajnykh-si> (accessed: March 24, 2025). (rus). EDN: GOWGOZ.
18. Rybakov A., Ivanov E., Nesterov V. Neyronnye seti dlya zashchity naseleniya i territoriy ot ChS [Neural networks for protection of the population and territories from emergencies]. *Sistemy bezopasnosti*, 2021. Available at: <https://www.secuteck.ru/articles/nejronnye-seti-dlya-zashchity-naseleniya-i-territorij-ot-chs> (accessed: March 25, 2025). (rus)
19. Sukhanova N.V. Razrabotka neyrosetevoy modeli dlya prognozirovaniya veroyatnosti zemletryaseniy [Developing a neural network model for predicting the probability of earthquakes]. *Automation and Modeling in Design and Management*, 2023. No. 2 (20). Pp. 40–49. (rus). DOI: 10.30987/2658-6436-2023-2-40-49. EDN: EDIMKN.
20. Hyndman R.J., Koehler A.B. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 2006. Vol. 22, No. 4. Pp. 679–688. DOI: 10.1016/j.ijforecast.2006.03.001.
21. Chai T., Draxler R.R. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 2014. Vol. 7, No. 3. Pp. 1247–1250. DOI: 10.5194/gmd-7-1247-2014.
22. Willmott C.J., Matsuura K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 2005. Vol. 30, No. 1. P. 79–82. DOI: 10.3354/cr030079.
23. Pchelintsev S., Yulyashkov M. A., Kovaleva O. A. Metod sozdaniya sinteticheskikh naborov dannykh dlya obucheniya neyrosetevykh modeley raspoznavaniyu ob"ektov [Method for creating synthetic data sets for training neural network models for object recognition]. *Information and Control Systems*, 2022. No. 3 (118). (rus). DOI: 10.31799/1684-8853-2022-3-9-19. EDN: LBEAQQ.

24. Rybakov A.V., Posternak E.V. Analiz problemnoy situatsii v oblasti obosnovaniya ob"emov inzhenerno-tekhnicheskikh meropriyatiy pri prognozirovanii ushcherba ot prirodnykh chrezvychaynykh situatsiy s uchetom antropogennykh faktorov [Analysis of the problematic situation in the optimization of engineering and technical measures when forecasting damage from natural emergencies, taking into account anthropogenic factors]. *Scientific & Educational Problems of the Civil Protection*, 2024. No. 4 (63). Pp. 72–83. (rus). EDN: BJANZK.

Copyright © 2025 Rybakov A.V., Posternak E.V.

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.