УДК 517:316.4

ИНФОРМАЦИОННАЯ СИСТЕМА ДЛЯ АНАЛИЗА И МОДЕЛИРОВАНИЯ СОЦИАЛЬНОГО И ЭКОНОМИЧЕСКОГО РАЗВИТИЯ РЕГИОНА

Романчуков Сергей Викторович
аспирант отделения Информационных технологий Инженерной школы информационных технологий и робототехники (ИШИТР),
e-mail: inoytomsk@yandex.ru,
Лызин Иван Александрович
аспирант отделения Информационных технологий ИШИТР,
e-mail: i-lyzin@mail.ru,
Марухина Ольга Владимировна
доцент, к.т.н. отделения Информационных технологий ИШИТР,
e-mail: marukhina@tpu.ru,
Россия, 634050, г. Томск, пр-т Ленина 30, Томский политехнический университет.

Аннотация. Предложен подход к разработке информационной модели, описывающей взаимосвязь социальных и экономических факторов инновационного развития региона РФ. В процессе исследования применялись диалектический подход, методы системного анализа, статистических группировок, факторного и кластерного анализа, информационного моделирования на основе информационных сетей, машинного обучения, нейросетевых моделей, нечёткой логики и др. Разработанная математическая модель и программный комплекс могут быть использованы в работе департаментов инвестиций и целевых программ региональных администраций с целью повышения эффективности проводимой в стране государственной социальной политики.

Ключевые слова: Информационная система, многомерные данные, анализ данных, социально-экономическое развитие, компьютерное моделирование, логические правила, нейронные сети.

Цитирование: Романчуков С.В., Лызин И.А., Марухина О.В. Информационная система для анализа и моделирования социального и экономического развития региона // Информационные и математические технологии в науке и управлении. 2020. № 3 (19). С. 96-104 DOI:10.38028/ESI.2020.19.3.010.

Введение. Информационные технологии уже достаточно давно проникли в науки об обществе. В настоящее время в социальных исследованиях широко используются различные математико-статистические методы обработки информации, реализованные в программных приложениях с применением современных информационных технологий. На этапе обработки результатов исследования наиболее популярными являются пакеты статистического анализа данных, такие, как SPSS, Statistica, STATGRAPHICS.

Проблема информационной поддержки социальных и экономических исследований актуальна не первый год, и по этой теме опубликовано достаточно большое количество материалов, однако даже их авторы зачастую отмечают тот факт, что разработанные в

рамках информационных технологий компьютерные приёмы решения социологических задач остаются неизвестными большинству социологов и психологов и потому зачастую не внедряются в реальную каждодневную практику исследовательских групп. Кроме того, акценты в большинстве подобных публикаций смещены в сторону задач анализа и обработки социологических данных, которые является важными и значимыми, но отнюдь не единственными проблемами.

Таким образом, междисциплинарный подход, предполагающий совместное рассмотрение социальных и экономических факторов с более широким использованием методов, предлагаемых IT-сферой, имеет право на существование. Перспективными представляются разработка нейросетевой модели, описывающей состояние целевых регионов, и создание автоматизированной системы, объединяющей потоки данных из разных статистических источников и использующей новые данные для прогнозирования будущего состояния регионов и постоянного дообучения моделей [6].

Объект исследования – социально-экономическая структура региона как единая система, представленная рядом статистических и экономических показателей, оценок уровня инновационного развития, социальное самочувствие региона, представленное т.н. "портретом региона" и рядом статистических показателей, сформированных по выборкам региональных социологических опросов.

Предметом исследования являются достоверные взаимосвязи между показателями социального самочувствия и инновационного развития региона.

Исследование заключается в разработке информационной модели, описывающей взаимосвязь социальных и экономических факторов инновационного развития региона РФ, обладающей достаточной точностью, предсказательной силой и пригодной для проведения вычислительных экспериментов с достоверными (статистически значимыми) результатами.

Формирование и подготовка массива статистических данных. Экспертные мнения и оценки позволили систематизировать источники статистических данных и места поиска в Интернете, связанные с инновационным развитием исследуемой отрасли и динамикой регионов – объектов исследования, включая официальные источники, такие, как:

- материалы, представленные на сайтах Федеральной службы государственной статистики (и её региональных отделений);
- материалы Всероссийского центра изучения общественного мнения (ВЦИОМ);
- рейтинги и материалы Ассоциации инновационных регионов России (АИРР);
- матрицы данных проекта “Социокультурный портрет региона” и других исследовательских коллективов, занимающихся региональной компаративистикой[1] [3];
- ряд зарубежных источников, включая материалы Всемирного Банка (уступают российским в детализации точности, но небезынтересны для сопоставления);
- ежегодные сборники опросов Левада-центра (оппозиционный источник добавлен в пул наравне с официальными государственными структурами для обеспечения политической непредвзятости исследования).

При этом источники данных ранжируются в зависимости от своего авторитета, регулярности, надёжности и частоты обновлений. Ключевая роль отводится данным

[1] Компаративи́стика (от лат. comparativus – сравнительный) – исследования социально-экономических систем через их сравнение.

Федеральной службы государственной статистики и ВЦИОМ, остальные источники служат на вспомогательных ролях для уточнения данных, заполнения временных промежутков, пропущенных по тем или иным причинам, если эти данные ложатся на общую линию тренда [1].

При наличии противоречащих данных об одном и том же временном интервале приоритет отдаётся официальным российским источникам, поддержанным РАН, РФФИ, РГНФ и иными научными, а также государственным структурам РФ, осуществляющим наблюдения на продолжительном отрезке времени и с достаточной регулярностью (как минимум – ежегодно).

Формирование исходного массива данных проходило в несколько этапов.

1. Получение исходных таблиц. Исходные данные достаточно разнородные:
    a. Разные источники.
    b. Разный формат отчётов.
    c. Разная гранулярность.
    d. Разные шкалы.
2. На их основе сформированы более обобщённые сводные таблицы:
    a. Единая гранулярность: регион + временная метка.
    b. Снижение размерности.
    c. Подготовка к последующей обработке.

К накопленным данным были применены методы разведочного корреляционного и кластерного анализа, а также методы анализа, построенные на алгоритмах визуализации временных рядов, разработанные в кооперации с коллективом исполнителей проекта РФФИ №18-07-00543. Проведение разведочного анализа позволило сократить размерность сформированного массива данных, выделив в исходной выборке наборы линейно-зависимых переменных и латентные переменные (факторы). Помимо этого, многие алгоритмы чувствительны к наличию в массиве данных коррелированных переменных, что потребовало их удаления. Первая итерация была проведена полностью вручную – начиная с запроса данных на сайтах-источниках и заканчивая их трансформациями.

Для последующего повторного (и регулярного) извлечения информации из выбранных источников, автоматического отслеживания обновлений была необходима разработка соответствующих алгоритмов парсинга, ориентированных на структуру и особенности выбранных сайтов, и последующее построение ETL-процесса (extract-transform-load – извлечение, трансформация и сохранение) по автоматизированному извлечению данных из выбранных источников, в случае обновления последних. С этой целью было создано приложение-парсер, осуществляющее семантический анализ соответствующих веб-страниц, извлечение данных, соответствующих заданным формальным признакам, сопоставление их с уже имеющимися в таблицах и внесение корректив в таблицы. Для решения этой задачи в рамках договора гражданско-правового характера был привлечён сторонний специалист с квалификацией программиста-разработчика [8].

Построение ETL-процесса. Парсинг источников данных, как уже было сказано, даёт нам набор разнородных таблиц. Для формирования итогового массива данных необходимо построение полноценного ETL-процесса, заполнение промежуточных таблиц и формирование итогового массива, согласно ранжированию источников данных (в соответствии с вышеупомянутыми соображениями авторитетности, регулярности и

новизны). Традиционный процесс ETL, в котором вы переносите и обрабатываете данные партиями из исходных баз данных в хранилище данных, следуя передовым методикам ETL, включает в себя:

1. Подготовительный этап. Создание набора данных, который определяет диапазон допустимых значений.
2. Извлечение данных. Основой успеха последующих шагов ETL является правильное извлечение данных и объединение данных из нескольких исходных систем, каждая из которых имеет свою собственную организацию и формат. Данные преобразуются в единый формат для стандартизированной обработки.
3. Валидация данных. Проверка, лежат ли данные, полученные из источников, в ожидаемых диапазонах и форматах. Отклонение новых данные, если они не соответствуют заданной структуре.
4. Преобразование данных. Очистка, повторная проверка и агрегирование данных.
5. Стейджинг. Заполнение промежуточной базы данных, диагностика устранение проблем с данными перед загрузкой преобразованных данных в целевое хранилище данных.
6. Загрузка данных непосредственно в целевые таблицы.

Уже на начальном этапе происходит первичное преобразование данных – оно включает в себя назначение метаданных для записей о принадлежности к одной из трех групп источников, адрес источника и необходимость обновления данных, добавление оценки достоверности/надежности источника по десятибалльной шкале, добавление/обновление временных меток, нормализация переменных.

Последний термин должен быть проиллюстрирован на примере различных исследований, опубликованных в разных источниках; оценивали такую переменную, как «уровень доверия к сотрудникам правоохранительных органов». В отчётах, опубликованных в этих источниках, данная переменная выражена в разных шкалах (пятибалльной и десятибалльной, в одной шкале более высокая оценка соответствует большей степени достоверности, и наоборот в другой). Перед сохранением и использованием данных на этом этапе необходимо привести их к единому стандарту, в котором эти данные будут записываться в своей собственной базе данных, и определить правило переноса информации из различных источников в этот стандарт [7].

После выполнения этой процедуры нормализованные данные загружаются во временное (буферное) хранилище. Отсюда они извлекаются для агрегации – формирования из нескольких переменных одного типа одной итоговой, которая впоследствии будет загружена в основное хранилище данных и будет пригодна для дальнейшего анализа [10].

После этого преобразования данные загружаются в основное хранилище данных и могут использоваться для дальнейшего анализа, однако ETL-процесс заканчивается сохранением данных в конечном хранилище. В нашем случае в качестве такого хранилища выступают сервисы Google. На следующем этапе поверх настроенного ETL-процесса был построен прототип информационной системы для исследования взаимосвязей и моделирования процессов социально-экономического развития региона, для которого сейчас проводятся эксперименты, опирающийся на программные продукты, активно используемые зарубежным бизнесом и наукой:

- использован модуль построения карт отношений между переменными и деревьями решений среды Answer Miner, позволяющий легче обнаруживать и интерпретировать

взаимосвязи в системе – эта среда имеет хорошие инструменты визуализации результатов и очень удобна для восприятия человеком;

- прогнозирование настроено с использованием нейросетевых (deepnet) моделей в среде BigML; обе эти платформы поддерживают импорт и автоматическое обновление данных из облачной среды Google, используемой нами как промежуточное хранилище.

Автоматизированное обучение нейронной сети. Для обучения непосредственно нейросетевой составляющей была выбрана среда BigML – эта онлайн-платформа существенно облегчает процесс не только конфигурации и обучения нейронной сети, но и последующего развёртывания полученной модели, выделения ресурсов для её функционирования и обеспечения непрерывной доступности в сети Интернет. Гибкий и достаточно мощный API среды обеспечивает взаимодействие с внешними программными решениями, что позволяет автоматизировать загрузку данных в среду и получение результатов [9].

Подбор параметров нейронной сети: общеизвестна чувствительность глубоких нейронных сетей к выбранной топологии и алгоритму, используемому для оптимизации их параметров. Эта чувствительность означает, что ручная настройка топологии и алгоритма оптимизации может быть сложной и трудоемкой, поскольку число вариантов, которые приводят к неудачным решениям, как правило значительно превосходит число вариантов, которые приводят к построению хороших нейросетевых моделей [4].

Для решения этой проблемы допустимо использование средств автоматического поиска топологии сети и оптимизации параметров, основанный на алгоритмах гиперполосы (eng. “hyperband algorithm”), однако вместо случайного отбора кандидатов для оценки мы используем метод сбора данных, основанный на байесовской оптимизации параметров [2].

Автоматизированное построение предположений о структуре: BigML предлагает быстрый метод, который также может дать качественные результаты. Суть его состоит в формализации общих правил о построении нейросетей, предположениях о том, что делает одну сетевую структуру лучше, чем другая для того или иного набора данных. BigML затем автоматически предлагает структуру и набор значений параметров, которые, к структуре предложенного набора данных. Результат можно получить быстро, но он будет единственным и сильно зависящим от корректности указания правил и определения природы входных данных [5].

Автоматический поиск параметров сети: более ресурсозатратный и длительный процесс, гарантирующий, однако, более глубокие результаты. Во время создания Deepnet среда получает задачу обучить и оценить все возможные конфигурации сети, возвращая лучшие сети, найденные решения для заданной проблемы. Компромиссное решение между верхними «n» сетями, найденными в поиске, даёт нам искомое решение. Схематически данный процесс изображён на рис. 1.

Поскольку этот вариант создает и обучает несколько сетей, он значительно более требователен к ресурсам, но полученные результаты окупают затраченное время и задействованные вычислительные мощности. Ресурсов, предоставляемых облачной средой BigML достаточно для выбора данного варианта.

Для решения задач классификации и построения регрессии (наш вариант) возможно использование углублённых алгоритмов OptiML. OptiML – это способ автоматизации процессов оптимизации при выборе и параметризации (или гипер-параметризации) модели.

Он также использует байесовскую оптимизацию параметров для выбора модели и настройки коэффициентов. Метод основан на последовательных алгоритмах конфигурации, зависимых от параметров модели (sequential model-based algorithm configuration, SMAC), которые последовательно тестируют группы параметров обучения и оценки моделей (используя методы Монте-Карло для кросс-валидации) и, основываясь на результатах, подбирает новую группу параметров. Когда процесс заканчивается, возвращается список самых эффективных моделей из числа рассмотренных. Основные параметры OptiML определяют, насколько интенсивно будет выполняться поиск, путем установки максимального времени обучения и количества оценок, списка допустимых алгоритмов машинного обучения, которые вообще имеет смысл оценивать [11].

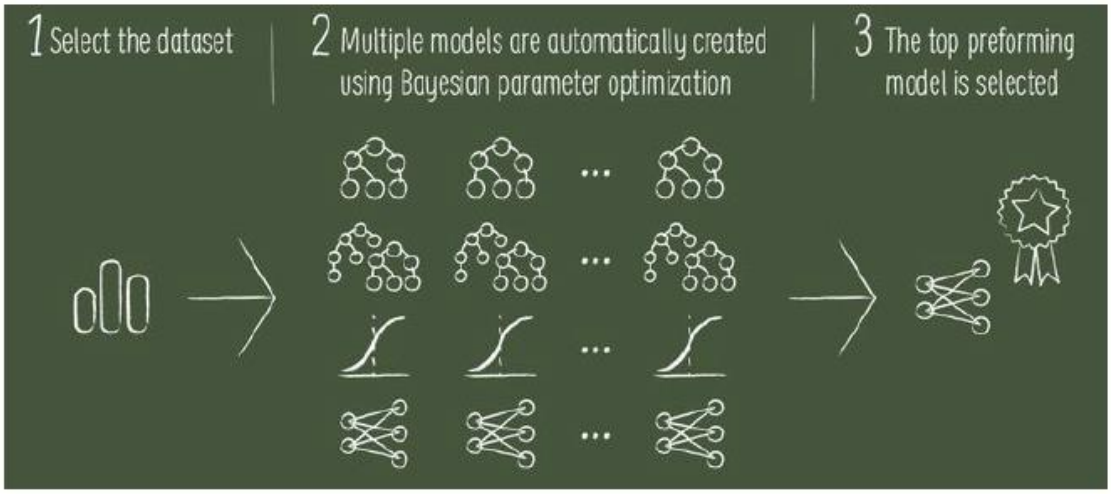

Рис. 1. Стадии автоматизированного создания модели

Для мониторинга прогресса моделей, которые проходят обучение и оценку, на информационной панели будут отображаться: прошедшее время, текущая серия F-мер с указанием оценок и счетчик количества созданных ресурсов, как это показано на рис. 2. Иконки на левой части экрана показывают количество моделей по типам: деревья решений, ансамбли, глубокие нейронные сети, регрессионные модели

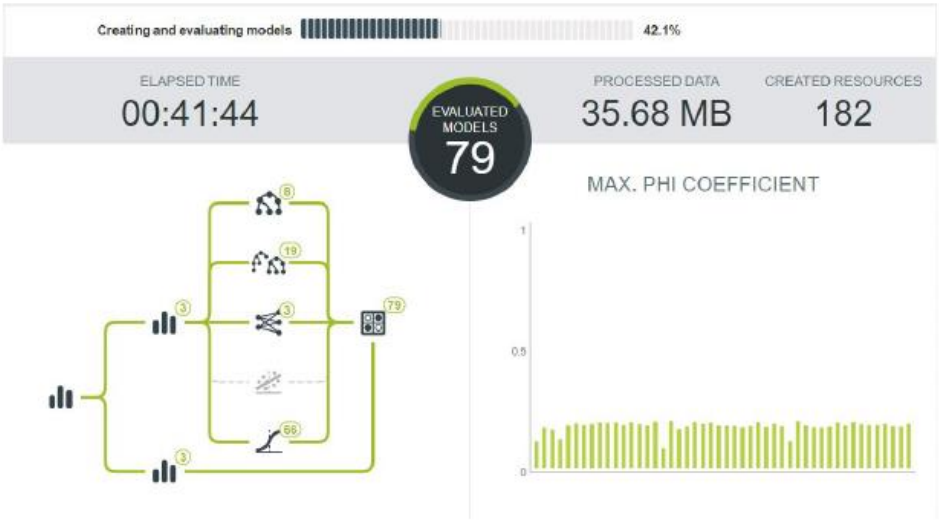

Рис. 2. Процесс автоматизированной оптимизации модели

После завершения процесса OptiML в сводном представлении отображаются общее количество созданных и выбранных ресурсов и моделей, а также общее время и количество обработанных данных. После завершения процесса мы получаем набор моделей, являющихся наиболее эффективными в соответствии с метрикой оценки, первоначально выбранной для анализа.

Заключение. В результате проведенного исследования был решен ряд научно-технических задач, а именно:

- подготовка данных к статистической обработке и глубинному анализу;
- выявление в доступных статистических данных значимых факторов, сокращение размерности факторного пространства;
- проведение разведочного анализа пространства факторов с целью поиска закономерностей и взаимосвязей;
- определение структуры информационной модели и выбор оптимальных технологических решений для её реализации;
- программная реализация информационной модели;
- проведение испытаний разработанного решения.

Основные положения исследования нашли своё применение при реализации проектов, поддержанных РФФИ: “Проблемы социокультурной эволюции России и ее регионов”, “Интеллектуальная система поддержки принятия управленческих решений по инновационному развитию региональных научно-медицинских центров” и “Анализ и моделирование взаимосвязей параметров социального и экономического развития региона”.

Отдельные положения и элементы системы нашли коммерческое применение при планировании региональной стратегии развития российского подразделения международной IT-компании Improvado.

Благодарности. Исследование выполнено при финансовой поддержке РФФИ и МОКНСМ в рамках научных проектов № 18-07-00543, № 20-07-00250-а и № 20-57-44002.

СПИСОК ЛИТЕРАТУРЫ

1. Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Классификация и снижение размерности: справочное издание / Под ред. С.А. Айвазяна. М.: Финансы и статистика. 1989. 607с.
2. Васенков Д.В. Методы обучения искусственных нейронных сетей//Компьютерные инструменты в образовании. 2007. № 1. С. 20-29.
3. Горшков А.В. Региональная компаративистика: теория и практика. Вестгник ЧелГУ. 2002. №1(8). С. 46-50.
4. Марухина О.В., Берестнева О.Г., Шаропин К.А., Берестнева Е.В., Жаркова О.С. Выявление скрытых закономерностей на основе интеллектуального анализа данных // Информационные технологии в науке, образовании и управлении материалы XLIV международной конференции и XIV международной конференции молодых учёных IT S&E`16. под редакцией Е.Л. Глориозова. Москва. 2016. С.143-148.
5. Романчуков С.В., Берестнева О.Г., Петрова Л.А. Обучение нейронной сети, моделирующей социально-экономическое развитие региона // Цифровая социология. 2019. Том №2. С. 34-40.

6. Толстова Ю.Н. Социология и компьютерные технологии//Социологические исследования. 2015. № 8. С. 3-13.
7. Ashurova Z., Tikhomirov A., Trufanov A., Kinash N., Berestneva O., Rossodivita A. Network platform of program governance for e-health service // Proceedings of the 12th international scientific and technical conference on computer sciences and information technologies, CSIT. 2017. Pp. 71-74.
8. Berestneva O.G., Marukhina O.V., Romanchuk S.V., Berestneva E.V. Visualization and Cognitive Graphics in Medical Scientific Research. Lecture Notes in Computer Science, Lecture Notes in Computer Science. 2019. Tom 11466. Pp. 433-444.
9. Berestneva O., Marukhina O., Rossodivita A., Tikhomirov A., Trufanov A. Networkalization of network–unlike entities: how to preserve encoded information // Communications in computer and information science. 2019. Tom: 1083. Pp. 143-151.
10. James G. (2003). Variance and Bias for General Loss Functions//Machine Learning. Режим доступа: http://www-bcf.usc.edu/~gareth/research/bv.pdf.
11. Romanchuk S.V., Berestneva O.G., Ivankina L.I. Population security and social confidence level markers factorisation (based on Tomsk Region studies). 2019. Part 2. Pp. 98-104.

UDK 517:316.4

INFORMATION SYSTEM FOR ANALYSIS AND MODELING OF SOCIAL AND ECONOMIC DEVELOPMENT OF THE REGION

Sergey V. Romanchukov
PhD-student, Department of information technology,
e-mail: inoytomsk@yandex.ru,
Ivan A. Lyzin
PhD-student, Department of information technology,
e-mail: i-lyzin@mail.ru,
Olga V. Marukhina
PhD, researcher, Department of information technology,
e-mail: marukhina@tpu.ru,
Russia, 634050, Tomsk, Lenin avenue 30, Tomsk Polytechnic University.

Abstract. An approach to the development of an information model describing the relationship between social and economic factors of innovative development of the Russian Federation region is proposed. The research applied dialectical approach, methods of system analysis, statistical groupings, factor and cluster analysis, information modeling based on information networks, machine learning, neural network models, fuzzy logic, etc. The developed mathematical model and software package can be used in the work of investment departments and target programs of regional administrations to increase the effectiveness of the state social policy in the country.
Keywords: Information system, multidimensional data, data analysis, socio-economic development, computer simulation, logical rules, neural network.

References

1. Ajvazyan S.A., Buhshtaber V.M., Enyukov I.S., Meshalkin L.D. Prikladnaya statistika. Klassifikaciya i snizhenie razmernosti: spravochnoe izdanie [Applied statistics. Classification and dimensionality reduction: reference edition]/ Pod red. S.A. Ajvazyana. M.: Finansy i statistika = Finance and statistics. 1989. 607p.

2. Vasenkov D.V. Metody obucheniya iskusstvennyh nejronnyh setej [Methods of training artificial neural networks] //Komp'yuternye instrumenty v obrazovanii = Computer tools in education. 2007. № 1. Pp. 20–29.

3. Gorshkov A.V.. Regional'naya komparativistika: teoriya i praktika. [Regional comparative studies: theory and practice]. Vestnik ChelGU = Vestgnik ChelSU. 2002. № 1 (8). Pp. 46-50.

4. Maruhina O.V., Berestneva O.G., Sharopin K.A., Berestneva E.V., ZHarkova O.S. Vyyavlenie skrytyh zakonomernostej na osnove intellektual'nogo analiza dannyh [Identifying hidden patterns based on data mining] // Informacionnye tekhnologii v nauke, obrazovanii i upravlenii materialy XLIV mezhdunarodnoj konferencii i XIV mezhdunarodnoj konferencii molodyh uchyonyh IT S&E`16. pod redakciej E.L. Gloriozova = Information technologies in science, education and management proceedings of the XLIV international conference and the XIV international conference of young scientists IT S&E`16. edited by E. L. Gloriozov, Moskva. 2016. Pp.143-148.

5. Romanchukov S.V., Berestneva O.G., Petrova L.A. Obuchenie nejronnoj seti, modeliruyushchej social'no-ekonomicheskoe razvitie regiona [Training of a neural network that models the socio-economic development of the region]. Cifrovaya sociologiya = Digital sociology. 2019. Vol. №2. Pp. 34-40.

6. Tolstova YU.N. Sociologiya i komp'yuternye tekhnologii [Sociology and computer technology] //Sociologicheskie issledovaniya = Sociological research. 2015. № 8. Pp. 3–13.

7. Ashurova Z., Tikhomirov A., Trufanov A., Kinash N., Berestneva O., Rossodivita A. Network platform of program governance for e-health service // Proceedings of the 12th international scientific and technical conference on computer sciences and information technologies, CSIT. 2017. Pp. 71-74.

8. Berestneva O.G., Marukhina O.V., Romanchuk S.V., Berestneva E.V. Visualization and Cognitive Graphics in Medical Scientific Research. Lecture Notes in Computer Science, Lecture Notes in Computer Science. 2019.Tom 11466. Pp. 433-444.

9. Berestneva O., Marukhina O., Rossodivita A., Tikhomirov A., Trufanov A. Networkalization of network–unlike entities: how to preserve encoded information // Communications in computer and information science. 2019. Vol. 1083. Pp. 143-151.

10. James G. (2003). Variance and Bias for General Loss Functions//Machine Learning. Available at: http://www-bcf.usc.edu/~gareth/research/bv.pdf. (accessed: 20.07.2020).

11. Romanchuk S.V., Berestneva O.G., Ivankina L.I. Population security and social confidence level markers factorisation (based on Tomsk Region studies). 2019. Part 2. Pp. 98-104.