ИНФОРМАТИКА

УДК 004.9

СРЕДСТВА И ПРИМЕРЫ ИНТЕЛЛЕКТУАЛЬНОЙ ОБРАБОТКИ ДАННЫХ ДЛЯ ГЕОЛОГИЧЕСКИХ МОДЕЛЕЙ

В.Б. Таранчук

Белорусский государственный университет, *Минск*

TOOLS AND EXAMPLES OF INTELLIGENT DATA PROCESSING FOR GEOLOGICAL MODELS

V.B. Taranchuk

Belarusian State University, *Minsk*

Обсуждаются вопросы разработки, инструментального наполнения, использования интегрированного программного комплекса составителя цифровых геологических моделей. Отмечены возможности интерактивной графической визуализации и сравнения результатов. Представлены и обсуждаются результаты применения искусственных нейронных сетей в анализе и интерпретации геопространственных данных.

***Ключевые слова**: цифровая геологическая модель*, *система компьютерной алгебры Mathematica*, *интерактивная графическая визуализация*, *искусственные нейронные сети.*

The problems of development, tool filling, and usages of the integrated program complex of the composer of digital geological models are considered. Possibilities of interactive graphics visualization and comparison of results are marked. The results of application of artificial neural networks in the analysis and interpretation of geospatial data are presented and discussed.

Keywords: *digital geological model*, *Computer Algebra System Mathematica*, *interactive graphics visualization*, *artificial neural network.*

Введение

Цифровые геологические, геоэкологические модели в настоящее время являются обязательной составляющей экспертизы во многих сферах деятельности. Геологическое моделирование представляет самостоятельное направление, включающее в себя развитие математических методов и алгоритмов; разработку компьютерных программ, обеспечивающих цикл построения моделей; создание баз данных, их наполнение и сопровождение. Основными этапами информационного обеспечения геологических моделей являются загрузка из различных источников и предварительная обработка данных, корреляция, создание цифровых кубов характеристик среды, интерактивный анализ данных, визуализация с помощью графики, картографирование.

Данные, используемые в геологических и геоэкологических моделях, являются представительной частью геоданных, которые интегрируют и обобщают информацию о процессах и явлениях на земной поверхности, включают классифицированные и интегрированные в единую систему группы данных [1]. Геоданные содержат пространственно-временные характеристики территории, предметов, построек. Существенно, что геоданные, как обобщение накопленной информации, включают сведения не только из области наук о Земле, но и из других, таких как транспорт, экономика, экология, управление, образование, анализ, искусственный интеллект. Геоданные, и в частности геологические данные, дополняют и интегрируют другие данные, чем обеспечивают решение многих задач.

Технологические, системные, информационные особенности геоданных отмечены в [1]. Технологическая особенность геоданных состоит в том, что их не получают на основе непосредственных измерений, а они формируются в результате постобработки измеренной информации, могут иметь разную точность, храниться с использованием разных единиц измерений. Системная особенность заключается в том, что после их формирования геоданные представляют собой интегрированную в единый комплекс совокупность параметров и описаний разных типов и структур, отражают различные характеристики и свойства, описывают реально существующие пространственные отношения с учетом временного и тематического факторов. Информационная особенность обусловлена тем, что геоданные представляют собой новый информационный ресурс, при этом данные группируют по трём характеристикам: месту, времени, теме. Также особенностью является наличие (реализуемое автоматически) взаимовлияния графических и атрибутивных данных – изменение атрибутивных данных предполагает замену графической информации, уточнение пространственного положения, требует изменений координат, пространственных отношений. Отмеченное взаимовлияние обеспечивает надёжную основу для пространственного визуального анализа и управления.

Объемы геоданных (огромные массивы имеющейся и вновь собираемой информации) растут с очень большой скоростью. Имеет место информационный бум. Соответственно, естественным является применение технологий «больших данных» (конкретика для геоданных в [2]), в том числе автоматизированного интеллектуального анализа данных (ИАД). В [3] акцентируется одна из главных целей ИАД – обнаружение в «сырых» (первичных) массивах данных ранее неизвестных, нетривиальных, практически полезных и понятных интерпретации знаний; отмечена специфика геоданных. В формулировке автора [3] «интеллектуальный анализ данных не исключает человеческое участие в обработке и анализе, но значительно упрощает процесс поиска необходимых данных из сырых данных, делая его доступным для широкого круга аналитиков, не являющихся специалистами в статистике, математике или программировании. Человеческое участие выражается в когнитивных аспектах участия и применения информационных когнитивных моделей».

Практические реализации, программные системы и комплексы для формирования и организации геоданных – предмет отдельного рассмотрения. Средства интеллектуального анализа геоданных такие же, как для обычных данных. Основой являются теории, методы, алгоритмы прикладной статистики, баз данных, искусственного интеллекта, распознавания образов. Различных действующих и применяемых программных средств интеллектуального анализа данных много, например, в [3] выделены следующие классы систем ИАД: «Индустриальные системы, Предметно-ориентированные аналитические системы, Статистические пакеты, Искусственные нейронные сети, Пакеты, основанные на деревьях решений, Системы рассуждений на основе аналогичных случаев, Генетические алгоритмы, Эволюционное программирование». Разнообразие предлагаемых методик и программных средств обуславливает необходимость оценки качества геоданных, определения их основных характеристик. Критерии оценок качества геоданных обсуждаются в [4]; к обязательным относят: репрезентативность, содержательность, прагматизм, достаточность, точность, актуальность, устойчивость, сертификат безопасности, надежность.

В настоящей работе обсуждаются возможные варианты получения оценок, методические решения и соответствующие программные инструменты, которые позволяют подтвердить обоснованность интерпретаций, получить числовые значения погрешностей получаемых разными методами результатов интеллектуальной предобработки данных, включаемых и используемых в компьютерных геологических моделях.

1 Программная платформа

Представленные ниже результаты получены с использованием компьютерного комплекса «Генератор геологической модели залежи» – ГГМЗ [5], [6]. Назначение комплекса – тестирование, оценки точности настраиваемых геологических моделей на основе применения СКА, ГИС, «умных» методов адаптации моделей в процессе их эксплуатации, «самонастройки» моделей с учётом дополняемых данных фактического развития процессов. Платформа разработки комплекса – система компьютерной алгебры *Mathematica* [7], язык Wolfram Language [8], [9], геоинформационная система Golden Software Surfer [10]. При программировании в системе *Mathematica* модулей графики реализованы технические решения, описанные в [11]. Предусмотрены возможности, когда программный комплекс в конкретной конфигурации может эксплуатироваться после сборки и сохранения в формате вычисляемых документов CDF. Расчеты, работа пользователя с CDF версией приложения возможна на любом персональном компьютере. При просмотре CDF версии, размещенной на web-сервере, программа просмотра автоматически подгружается в виде плагина браузера. Автономная работа с ПК возможна после инсталляции свободно распространяемого компонента CDF Player. Варианты дополнительных настроек, обеспечивающих интерактивность CDF версии, изложены в [12], [13].

2 Компоненты комплекса ГГМЗ

В изложении ниже упомянуты компоненты, фактически являющиеся автономными программными модулями. Их также можно позиционировать, как составные части автоматизированного рабочего места специалиста, который в вычислительных экспериментах отрабатывает приемы адаптации используемых при построении геологических моделей цифровых полей. Следует отметить важное техническое решение – все этапы работы с модулями комплекса поддерживаются функциями импорта и экспорта результатов с несколькими вариантами настроек форматов. Это обеспечивает пользователя дополнительными возможностями выполнения аналогичных расчетов в разных (в том числе других) приложениях, сопоставления результатов.

В программном комплексе ГГМЗ реализованы следующие средства:

– инструменты и шаблоны для подготовки эталонной модели цифрового поля, отвечающего оговоренным свойствам («Конструктор цифрового поля»);

– средства и несколько вариантов модулей «искажения» эталонной модели;

– инструменты имитации «съема» данных, которые используются в практике моделирования («Генератор профиля наблюдения»);

– модули расчета, визуализации, сопоставления аппроксимирующих цифровых полей несколькими разными методами (компонент «Аппроксимация»);

– инструменты и модули адаптации («доводки») формируемой цифровой модели (компонент «Адаптация»).

Основная идея и цель разработки настоящего комплекса состоит в выборе метода обработки исходных данных путем сопоставления эталонного цифрового поля и восстановленного по «наблюдениям». Эталонное распределение для прямоугольной области формируется с использованием математических описаний, каждый эксперт при этом определяет и включает в модель типичные фрагменты. Затем инструментами комплекса выполняются «наблюдения», имитируется снятие замеров по эталонному распределению, причем геометрия точек замеров и их точность также определяются пользователем комплекса и примерно должны соответствовать исходным данным предметной области. Схема размещения точек с замерами не должна быть регулярной. В результате этого этапа пользователь получает набор данных «наблюдений», основными из которых являются координаты точки и значение в ней. Следующим этапом является выбор алгоритма обработки полученного набора («воспроизведения» цифрового поля) выполнением интерполяции и экстраполяции. Сопоставление результатов «воспроизведения» и эталона подскажут эксперту метод обработки, геометрию точек наблюдения.

3 Этапы подготовки типовой эталонной модели

Конструктор цифрового поля (КЦП). Программные модули этой группы обеспечивают в режиме интерактивной работы конструирование из типовых элементов с сопутствующей визуализацией математического описания (аналитической функции) модели поверхности, интерпретируемой, как рельеф - совокупность фрагментов разных форм поверхности. Возможности, шаги формирования подробно описаны и проиллюстрированы разными примерами в [5], [6]. Пользователем задаются границы области определения и ограничения поверхности по высоте. В комплекте (библиотеке) составных частей формируемой функции есть математические выражения (элементы), обеспечивающие воспроизведение участков поведения, характерных для рельефа местности. Пользователь на первом этапе формирования эталонной поверхности составляет кусочно-заданную функцию, базовый профиль – ленту заданной ширины и длины, имитирующую типы рельефа с элементами плато, склон, откос, обрыв. Составление средствами КЦП базового профиля из фрагментов возможно с переходом «фрагмент – добавленный фрагмент» непрерывным образом, гладким переходом, скачком (имитация разлома). В случае непрерывного, гладкого переходов параметры склейки кусочно-заданной функции определяются автоматически программным модулем. Простейший вариант определения функции и задания базового профиля записан в выражении (3.1) [5]. В приведенном примере модель базовой поверхности является квазитрехмерной, уровень z (высота) не зависит от y (ширины ленты). Базовая поверхность (лента) составлена из 3-х типовых участков: плоский горизонтальный (плато), плоский быстрого возрастания уровня (откос), плоский медленного возрастания (пологий склон). Стыковка участков – непрерывно. Переход «плато – откос» осуществлен под заданным углом, переход «откос – пологий склон» осуществлен непрерывным и гладким. Следующий этап конструирования - использование инструментов программного модуля для дополнения базового профиля возмущениями, фрагментами типовых элементов рельефа. Библиотека шаблонов включает элементы, которые соответствуют возмущениям (участкам искажений базовой поверхности) разной геометрической формы. При подключении шаблонов предусмотрены возможности интерактивного задания их положения и размеров.

В основной комплект комплекса включены математически описываемые элементы, которые имитируют следующие формы рельефа: холм, насыпь, яма, выемка, траншея, канал, карьер, овраг, впадина. Аналитическое выражение для сформированной и рассматриваемой далее поверхности – (3.5) [5]. На рисунке 3.1 показаны два вида поверхности, с которой далее проводятся вычислительные эксперименты. Слева дан вид поверхности целиком, справа – с простым профильным разрезом по направлению {0,10} – {100,20}. Рассмотрен конкретный профиль 2 (в [5] иллюстрирован одномерными графиками на рисунках 4.1 и 4.2).

4 Подготовка данных для численных экспериментов

Wolfram *Mathematica*, как система интеллектуальных вычислений, предоставляет пользователю не только средства математических преобразований, точных и приближенных вычислений, но и инструменты машинного обучения. Они могут быть использованы при интерпретации и обработке входных данных и результатов моделирования. Приведенные ниже результаты и примеры анализа данных с помощью нейронных сетей получены с использованием соответствующих функций, которые доступны в версии 11 *Mathematica* [7].

Рассмотрим средства и несколько вариантов применения модулей «искажения» эталонной модели, примеры предобработки, вывода графиков профилей с использованием инструментов

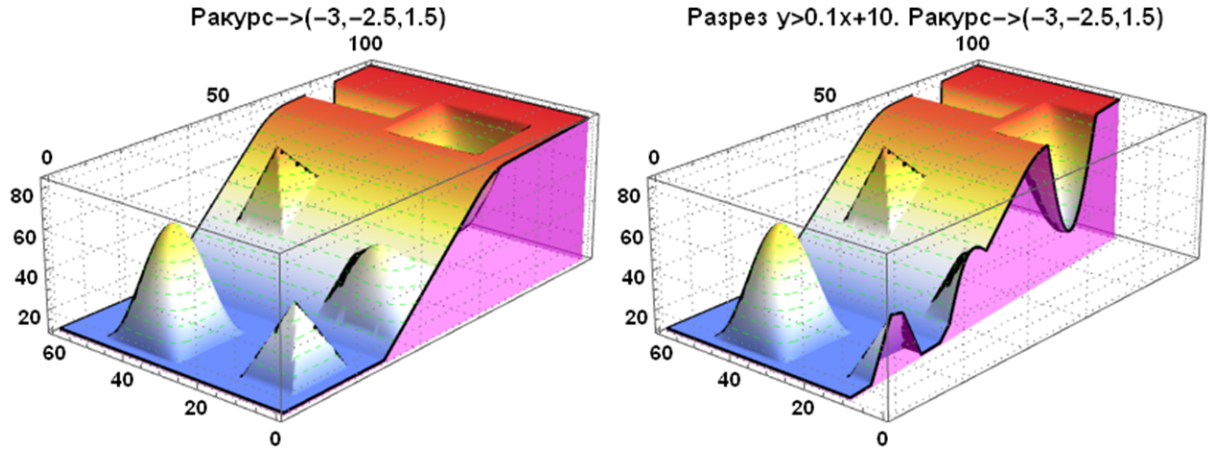

Рисунок 3.1 – Вид поверхности в полном варианте и с профильным разрезом

компонента комплекса «Генератор профиля наблюдения». Имитируем наблюдения и получим типичные наборы данных замеров. Иллюстрируем варианты применения инструментов ГГМЗ, в частности, модуля «искажения» эталонной модели для того, чтобы показать на примере профилей по эталонным поверхностям несколько методов имитации наблюдений и получения исходных данных средствами инструментов модуля «Генератор профиля наблюдения». Для этой цели используем рассчитанные из аналитических выражений на упомянутом профиле значения, но добавим «искажения», используя разные генераторы шума. Графики представлены на рисунке 4.1. На графиках сплошная синяя линия иллюстрирует аналитическую функцию распределения высоты поверхности на выбранном направлении разреза; кружочками отмечены значения в «узлах наблюдений» (если бы замеры были точными). Окружностями и ромбиками отмечены значения, которые имитируют замеры. На графике слева окружностями отмечены «данные наблюдений», когда исходные значения в узлах (обозначены кружочками) искажаются с использованием выборки, генерируемой функциями *Mathematica* RandomVariate (реализация случайной переменной) и NormalDistribution[0,3] (нормальное распределение со средним 0 и стандартным отклонением 3). На графике справа ромбиками отмечены «данные наблюдений», когда исходные значения искажаются с использованием выборки, генерируемой функциями RandomVariate и UniformDistribution[{–6,6}] (равномерное статистическое распределение, дающее значения от –6 до 6).

Отдельно следует обратить внимание на специально подобранный для экспериментов профиль и задаваемые величины отклонений. Исходная функция имеет область изменения от 10 до, примерно, 80, при этом есть небольшие возмущения от базового уровня (ленты) вблизи $x = 40$ и существенные отклонения от уровня ленты вблизи $x = 10$ и $x = 80$. Величины этих отклонений соразмерны величине изменения функции. Относительно шумов – есть незначительные и есть такие, которые можно классифицировать как «выбросы».

Данные с «искажениями» далее будем использовать для двух разных целей, соответственно будут применяться разные методы предобработки. Важно следующее – обсуждаемые ниже методы интеллектуального анализа «не знают» первоисточник, эталонную функцию. Задача эксперта – максимально воспроизвести (угадать) оригинал по имеющимся данным «замеров». Кстати, распределение, названное в тексте эталонным, может и не быть оригиналом.

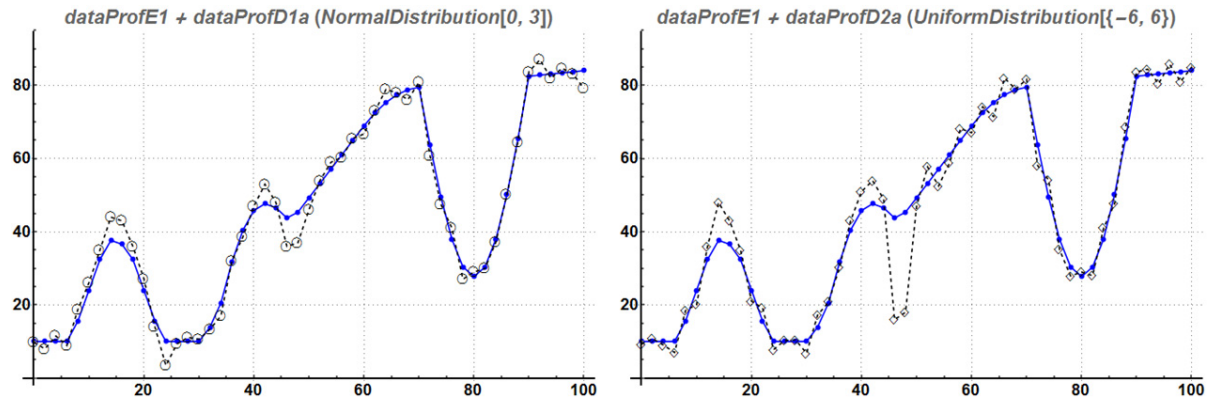

Рисунок 4.1 – Данные на профиле с «искажениями», dataProfD1a, dataProfD2a

5 Варианты интеллектуального анализа данных

Как отмечено выше, в настоящее время много методик и программных средств для интеллектуального анализа данных базируются на достижениях искусственных нейронных сетей. Рассмотрим несколько результатов, иллюстрации которых даны на рисунках 5.1, 5.2.

При решении задач математического моделирования исходные уравнения записываются в дифференциальной форме, поэтому и исходные данные (оснастка модели) должны быть непрерывными, более того, как правило, распределения должны быть гладкими функциями. Другими словами, распределение наблюдаемого параметра вдоль профиля надо передавать в исходные данные компьютерной теоретической модели в виде гладкой функции. Заметим, что и эталонные данные рисунка 4.1 требованию гладкости не отвечают, а данные с «искажениями» вовсе непригодны для численных моделей – такие исходные данные немедленно повлекут «болтанку» в решениях (вычислительная неустойчивость) и результаты будут непригодны.

Рассмотрим варианты предобработки данных, подразумевая получение аппроксимирующей гладкой функции.

На рисунке 5.1 представлены результаты расчетов после обучения искусственной нейронной сети с использованием функции *Mathematica* методом ADAM [14]. В методе Adam программируется оптимизационный алгоритм, который сочетает в себе идею накопления движения и идею более слабого обновления весов для типичных признаков; реализуется метод стохастического градиентного спуска с использованием адаптивной скорости обучения, инвариантной к диагональному масштабированию градиентов. Ключевые конструкции кода, функции и опции системы *Mathematica* включают команды: netA = NetChain[{vectLength, Tanh, vectLength, Tanh, 1}, "Input"->"Scalar", "Output"->"Scalar"]; netD1a = NetTrain[netA, dataProfD1a, Method->"ADAM"]. Полученное сглаженное приближение выводится на рисунке 5.1 пунктирной красной линией, дополнительно в поле графика даны две вставки ситуационного плана (выкопировки), которые детализируют графики.

На рисунке 5.2 представлены результаты для dataProfD2a, полученные после обучения сети на основе функции NetTrain методом RMSProp – стохастический градиентный спуск с использованием адаптивной скорости обучения, полученной из экспоненциально сглаженного среднего значения градиента.

Следует отметить, что для обработки данных dataProfD1a также применялся метод "RMSProp". В рассмотренном примере ключевые конструкции кода были следующими: netAr = NetChain[{vectLength, Tanh, vectLength,Tanh,1}, "Input"->"Scalar", "Output"–> "Scalar"]; netD1ar = NetTrain[netAr,dataProfD1a,Method–>"RMS Prop"]. В расчетах по этому методу было достаточно vectLengt = 10, а полученные результаты аналогичны варианту рисунка 5.1.

На рисунке 5.2 представлены результаты, полученные при net2Ar = NetChain[{vectLength, Tanh, vectLength, Tanh, 1}, "Input"->"Scalar", "Output"-> "Scalar"]; netD2ar = NetTrain[net2Ar, dataProfD2a, Method->"RMSProp"]. Сглаженное приближение выводится штрихпунктирной линией малинового цвета.

Заключение

Описаны компоненты интегрированного программного комплекса составителя цифровых геологических моделей. Разработанный комплекс дает возможности манипулирования исходными данными, интеллектуального анализа, сопоставления интерпретаций и вариантов экспертов, получаемых разными способами результатов и эталонов. Обсуждаемые результаты, примеры обработки и визуализации пространственных данных, методы настройки инструментов искусственных нейронных сетей являются подтверждением широких возможностей рассматриваемой технологии интеллектуальной обработки данных. Но следует понимать, что работа по

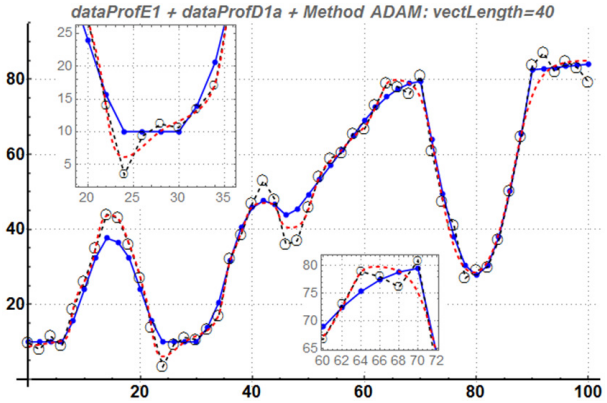

Рисунок 5.1 – Сглаживание dataProfD1a с применением нейронной сети

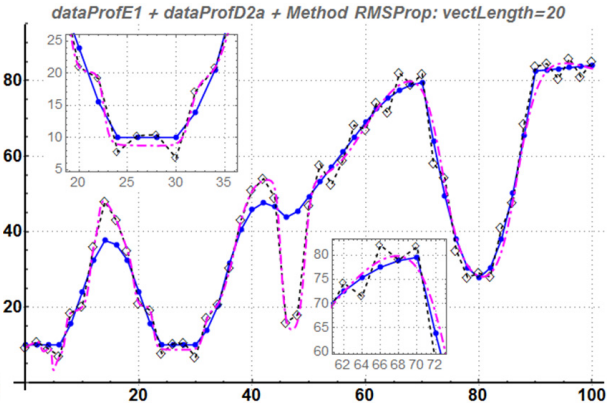

Рисунок 5.2 – Сглаживание dataProfD2a с применением нейронной сети

обучению сети – работа с черным ящиком, с неповторяющимися результатами. Причем, всегда присутствует субъективность, все определяется опытом, когда принимается решение о необходимости наложить какие-то ограничения на функцию потерь. Поэтому отдельно в следующей статье будут приведены классические решения методами прикладной статистики.

ЛИТЕРАТУРА

1. *Савиных*, *В.П.* Геоданные как системный информационный ресурс / В.П. Савиных, В.Я. Цветков // Вестник Российской академии наук. – 2014. – Т. 84, № 9. – С. 826–829.

2. *Tsvetkov*, *V.Ya.* Big Data as Information Barrier / V.Ya. Tsvetkov, A.A. Lobanov // European Researcher. – 2014. – Vol. 78, № 7–1. – P. 1237–1242.

3. *Шайтура*, *С.В.* Интеллектуальный анализ данных геоданных / С.В. Шайтура // Перспективы науки и образования. – 2015. – № 6 (18). – С. 24–30.

4. *Дышленко*, *С.Г.* Анализ и разработка характеристик качества геоданных / С.Г. Дышленко // Перспективы науки и образования. – 2016. – № 2 (20). – С. 23–27.

5. *Таранчук*, *В.Б.* Программный комплекс адаптации геологических моделей. Концепция, решения, примеры реализации / В.Б. Таранчук // Проблемы физики, математики и техники. – 2017. – № 3 (32). – С. 81–90.

6. *Taranchuk*, *V.B.* The integrated computer complex of an estimation and adapting of digital geological models / V.B. Taranchuk // Studia i Materiały. – 2017. – № 2 (14). – P. 73–86.

7. *Wolfram Mathematica. Наиболее полная система для современных технических вычислений в мире* [Электронный ресурс] / Wolfram Computation Meets Knowledge. – Режим доступа: http://www.wolfram.com/mathematica. – Дата доступа: 9.03.2019.

8. *Таранчук*, *В.Б.* Введение в язык Wolfram: учеб. материалы для студентов фак. прикладной математики и информатики спец. 1-31 03 04 «Информатика» / В.Б. Таранчук. – Минск: БГУ, 2015. – 51 с.

9. *Таранчук*, *В.Б.* Основы программирования на языке Wolfram: учеб. материалы для студентов фак. прикладной математики и информатики спец. 1-31 03 04 «Информатика» / В.Б. Таранчук. – Минск: БГУ, 2015. – 49 с.

10. *Explore the depths of your data*. *Surfer* [Электронный ресурс] / Golden Software. – Режим доступа: http://www.goldensoftware. com/products/surfer. – Дата доступа: 9.03.2019.

11. *Таранчук*, *В.Б.* Особенности функционального программирования интерактивных графических приложений / В.Б. Таранчук // Вестник Самарского государственного университета. Естественнонаучная серия, раздел Математика. – 2015. – № 6 (128). – С. 178–189.

12. *Таранчук*, *В.Б.* О создании интерактивных образовательных ресурсов с использованием технологий Wolfram / В.Б. Таранчук // Информатизация образования. – 2014. – № 1 (73). – С. 78–89.

13. *Таранчук*, *В.Б.* Об использовании системы Mathematica при подготовке и распространении интерактивных графических приложений / В.Б. Таранчук, В.А. Кулинкович // Весці БДПУ. Серыя 3. – 2015. – № 2 (84). – С. 58–64.

14. *NetTrain* (*Experimental*) [Электронный ресурс]. – Режим доступа: https://reference. wolfram.com/language/ref/NetTrain.html. – Дата доступа: 9.03.2019.

Поступила в редакцию 10.03.19.