

УДК 004.9

DOI 10.18413/2411-3808-2019-46-3-511-522

МЕТОДЫ И ПРИМЕРЫ ИНТЕЛЛЕКТУАЛЬНОЙ ОБРАБОТКИ ДАННЫХ ДЛЯ ГЕОЛОГИЧЕСКИХ МОДЕЛЕЙ

METHODS AND EXAMPLES OF INTELLIGENT DATA PROCESSING FOR GEOLOGICAL MODELS

В.Б. Таранчук**V.B. Taranchuk**

Белорусский государственный университет, Беларусь, 220030, г. Минск, пр. Независимости, 4

Belarusian State University, 4 Nezavisimosti Ave, Minsk, 220030, Belarus

E-mail: taranchuk@bsu.by

Аннотация

В работе обсуждаются результаты разработки, инструментального наполнения, использования интегрированного программного комплекса составителя цифровых геологических моделей. Отмечены возможности формирования эталонных и восстанавливаемых разными методами аппроксимации цифровых полей, интерактивной графической визуализации, подготовки наборов представительных данных для возможных интерпретаций экспертов. Описаны методы, варианты получения оценок, методические решения и программные инструменты, которые позволяют подтвердить обоснованность интерпретаций, визуализировать и получить числовые значения погрешностей, рассчитываемых разными методами результатов интеллектуальной обработки данных, включаемых и используемых в компьютерных геологических моделях.

Abstract

The paper discusses the results of the development, instrumental filling, the use of an integrated software package of the compiler of digital geological models. The developed complex provides the possibility of manipulating the original data, their intellectual analysis, comparison of different interpretations. The developed complex provides the possibility of manipulating the primary data, their intellectual analysis, comparison of different interpretations. The possibilities of formation of reference and recoverable by different methods of approximation of digital fields, interactive graphic visualization, preparation of sets of representative data for possible interpretations of experts are marked. Several variants of analysis and evaluation, methodological solutions and software tools that allow confirming the validity of interpretations, visualize and obtain numerical values of errors calculated by different methods of intellectual data processing results included and used in computer geological models are described. The discussed results, examples of processing and visualization of spatial data, methods of tuning the tools of artificial neural networks are a confirmation of the wide possibilities of the considered technology of intelligent data processing.

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

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 [Савиных, Цветков, 2014]. 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. 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 [Савиных, Цветков, 2014]. 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 [Tsvetkov, Lobanov, 2014]), в том числе автоматизированного интеллектуального анализа данных (ИАД). В работе [Шайтура, 2015] 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 [Шайтура, 2015] «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 работе [Шайтура, 2015] 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 статье [Дышленко, 2016]. К обязательным относят: репрезентативность, содержательность, прагматизм, достаточность, точность, актуальность, устойчивость, сертификат безопасности, надежность.

Ряд перечисленных вопросов по анализу, оценкам качества пространственных данных можно решать с использованием компьютерного комплекса «Генератор геологической модели залежи» – ГГМЗ [Таранчук, 2017; Taranchuk, 2017]. В настоящей работе обсуждаются дополнительные возможные варианты получения оценок, методические решения и программные инструменты, которые позволяют подтвердить обоснованность интерпретаций, визуализировать и получить числовые значения погрешностей, рассчитываемых разными методами результатов интеллектуальной обработки данных, включаемых и используемых в компьютерных геологических моделях.

Программная платформа. Компоненты компьютерного комплекса ГГМЗ

Представленные ниже результаты получены с использованием компьютерного комплекса ГГМЗ. 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 [WOLFRAM MATHEMATICA], язык Wolfram Language ([Таранчук, 2015а; Таранчук, 2015б]), геоинформационная система Golden Software Surfer [SURFER]. При программировании в системе Mathematica модулей графики реализованы технические решения, описанные в работе [Таранчук, 2015в]. 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 вебсервере, 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 работах [Таранчук, 2014; Taranchuk, Kulinkovich, 2015а; Taranchuk, Kulinkovich, 2015б].

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 комплекса ГГМЗ состоит в предоставлении эксперту инструментов выбора и обоснования метода обработки пространственных данных путем сопоставления эталонного цифрового поля и восстановленного по «наблюдениям». Эталонное цифровое поле рассматривается как явно задаваемая аналитическая функция $z(x,y)$, определяемая в области с двумя независимыми переменными. В случае дискретного представления это сеточная функция – множество, массив значений $z_{i,j}$ в узлах сетки (x_i, y_j) . В настоящем изложении вместо термина «цифровое поле» часто используется «поверхность», как наиболее понятное для пояснений. Но следует понимать, что с позиций математического описания речь идет о явном задании функции, определяемой в области с двумя независимыми переменными. Так можно моделировать любое распределение,



например: концентрация примеси или плотность материала слоя, насыщенность продуктивного пласта нефтью, пористость, гидропроводность или мощность пропластка.

Программные модули группы «Конструктор цифрового поля» (КЦП) обеспечивают с сопутствующей графической визуализацией в режиме интерактивной работы конструирование из типовых элементов математического описания (аналитической функции) модели поверхности – совокупность фрагментов разных форм поверхности. Каждый эксперт при этом определяет и включает, позиционирует, масштабирует, ориентирует в модель типичные фрагменты.

Затем инструментами комплекса в компоненте «Генератор профиля наблюдения» (ГПН) выполняются «наблюдения», имитируется снятие замеров по эталонному распределению, причем геометрия точек замеров и их точность также определяются пользователем комплекса и должны соответствовать типичным исходным данным предметной области. Пользователь, используя интерактивные инструменты компонента ГПН, рисует схему профилей, задавая на каждом из них число узлов. В отдельном программном модуле ГГМЗ для каждого узла схемы профилей формируются пять чисел: номера профиля и узла, координаты x и y , вычисленное значение функции поверхности в точке-узле. Упомянутые пятерки чисел экспортируются в файл, пользователь выбирает формат из предлагаемых вариантов; в том числе это может быть экспорт в .XLS (XLSX, CSV). Схема размещения точек с замерами не обязательно должна быть регулярной. В результате этого этапа пользователь получает рассеянное (нерегулярное) множество точек, набор данных «наблюдений», основными из которых являются координаты узла и значение параметра в нем. Такая имитация получения данных по форме поверхности (рельефу) дает точное значение уровня эталонной поверхности в точке замера. Отдельно в компоненте ГПН есть инструменты «искажения результатов наблюдений» (имитация погрешности замеров на каждом отдельном профиле) наложением шумов с использованием наборов разных функций с осцилляциями и генераторов случайных чисел [Таранчук, 2015б].

Следующим этапом является выбор алгоритма обработки полученного набора данных («восстановления» цифрового поля) выполнением интерполяции и экстраполяции на регулярную сетку [Таранчук, 2006]. Сопоставление результатов «восстановления» и эталона подскажут эксперту метод обработки, геометрию точек наблюдения.

Методы и инструменты анализа погрешностей

В статьях [Таранчук, 2017; Taranchuk, 2017] изложены и проиллюстрированы разными примерами методические и технические решения задач построения, анализа эталонных и аппроксимирующих цифровых полей распределений параметров, характеризующих геологические объекты. Отмечены более десяти вариантов 1D, 2D и 3D графики, в том числе модули интерактивного получения и оформления карт и вставок на них, графиков на профилях, 3D визуализации возможных разрезов (вертикальные, горизонтальные), простых и сложных отсечений. Презентованы результаты, рассмотрены примеры получения и наглядного сопоставления разных решений.

Ниже обсуждаются дополнительные возможности и программные инструменты, которые обеспечивают более полный анализ качества формируемых моделей, позволяют визуализировать и получать числовые значения погрешностей, рассчитываемых разными методами результатов интеллектуальной обработки пространственных данных. При подготовке данной работы в расчетах используются аналитические выражения (5) и (6) статьи [Таранчук, 2017], обсуждаются новые результаты для распределений $z_{SurfA}(x,y)$ и $z_{SurfB}(x,y)$, базовые иллюстрации которых приведены в работе [Таранчук, 2017] на рис. 1–7 (функции определены в прямоугольнике $[xMin,xMax] \times [yMin,yMax]$, $xMin = 50$, $xMax = 250$, $yMin = 0$, $yMax = 160$). Заметим, что подобные распределения (поверхности) можно трактовать, как кровлю и подошву слоя (в частности, водоносного горизонта, нефтеносного пласта, слоя осадочной породы). Описанные в упомянутой статье и используемые ниже варианты распределений не являются примерами лучшего «восстановления» и получения

адаптированного аппроксимирующего цифрового поля. Для иллюстраций в изложении специально применяется простейший метод интерполяции, чтобы получать значительные погрешности, наглядно отметить инструменты ГМЗ для анализа и сопоставления результатов расчетов.

На рис. 1 ниже показаны результаты сопоставления эталонных цифровых распределений и восстановленных полей по двум наборам узлов. Изображение слева иллюстрирует результаты для $zSurfA$ (рис. 8 статьи [Таранчук, 2017]), справа – для $zSurfB$. На рисунках показаны изолинии и схемы точек с замерами – 97+10+12 точек. Начальная схема профилей – одна из предложенных экспертом, которому была дана предварительная информация о форме базовой поверхности, положении фрагментов-возмущений (только центры и словесное описание формы). Узлы начальной схемы на профилях показаны разными примитивами синего цвета (их 97, треугольники, звездочки, ромбики, плюсики, квадратики), примитивы опорных точек фрагментов-возмущений – красные кружки (их 10). Дополняющие начальную схему узлы второй сети точек наблюдений (алгоритм адресного уплотнения для схемы 97 точек) показаны черными примитивами-кольцами (12 узлов), новые узлы дополняют предыдущие, причем они размещаются на первоначальных профилях, новые направления дополнять не предусмотрено. Сплошными линиями выведены изолинии эталонного цифрового поля, построенного по аналитически заданному распределению; уровни изолиний подобраны и заданы так, чтобы площадь на участках не плато была покрыта изолиниями относительно равномерно. Пунктирными утолщенными линиями показаны изолинии тех же уровней, которые получены для сеточной функции, сформированной в системе Mathematica методом интерполяции второго порядка на множестве начальной схемы 97+10 точек. Штрихпунктирными линиями показаны изолинии тех же уровней, полученные на множестве 97+10+12 точек, где дополнительные 12 точек сформированы алгоритмом адресного уплотнения для поверхности слева. Результат справа подтверждает, что для каждого варианта нужен отдельный подбор дополнительных точек – другие уровни, другая форма возмущений. Поэтому для поверхности $zSurfB$ этот алгоритм применялся повторно, уже для схемы 97+10+12 точек. Соответствующие результаты, рассчитанные на множестве 97+10+20 точек (8 дополнительных узлов – ромбики), показаны на рис. 2.

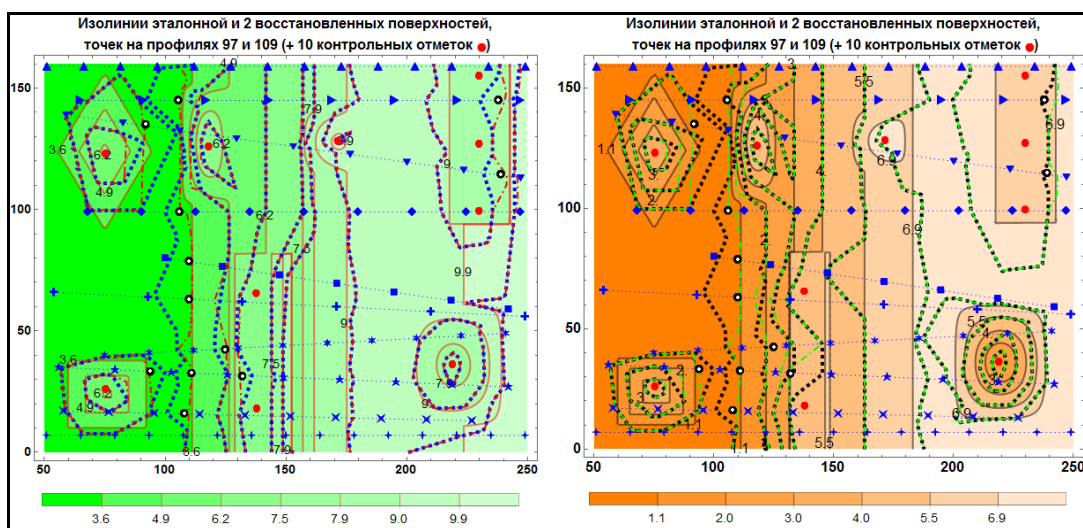


Рис. 1. Изолинии эталонных и восстановленных полей для $zSurfA$, $zSurfB$
 Fig. 1. Isolines of the reference and reconstructed fields for $zSurfA$, $zSurfB$

На рис. 2 справа показан фрагмент $[50,150] \times [90,160]$ карты, выведенной слева, что иллюстрирует возможности детализации при изучении контролируемых распределений на выбранных участках.

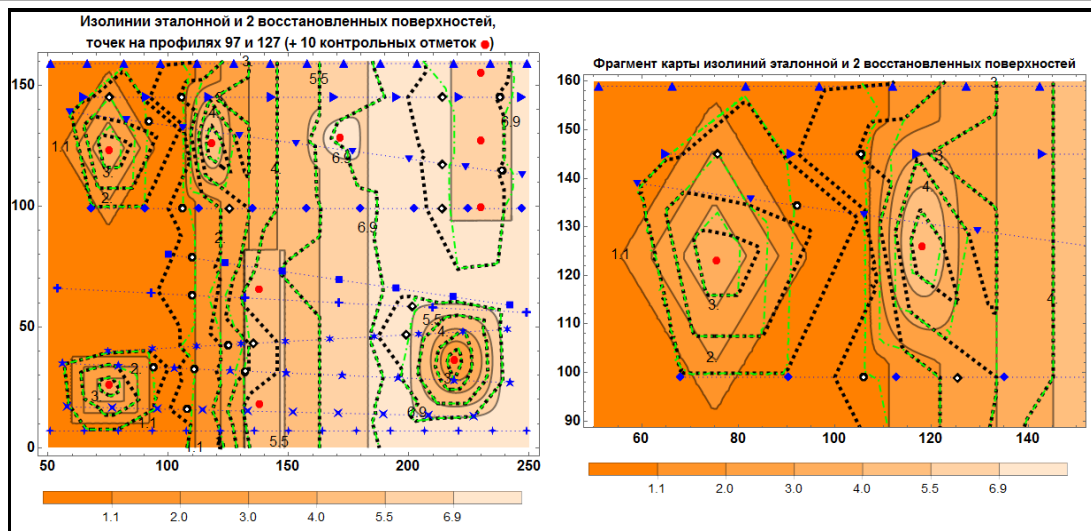


Рис. 2. Изолинии эталонного и двух восстановленных по 107 и 127 узлам полей для z_{SurfB}
 Fig. 2. Isolines of the reference and two reconstructed fields on 107 and 127 nodes for z_{SurfB}

Иллюстрации на рис. 1, 2 дают общее представление об отличиях форм поверхностей. Детали можно анализировать на карте невязки, что не представляет трудностей, так как в ГМЗ имеем аналитическое представление эталона, и средствами Wolfram Mathematica можно получить аналитическое выражение для любой восстановленной сеточной функции. Примеры иллюстраций распределения погрешности восстановленной (в сравнении с эталонной) функции, полученные выводом изолиний (ContourPlot) невязки с цветовым заполнением интервалов разницы цифровых полей, приведены для нескольких типичных вариантов в работах [Таранчук, 2018; Taranchuk, 2018]. Дополнительно отметим, что можно получать числовые значения для любой принятой нормы отклонений, рассчитывать их для всего объекта или любой его части. Соответствующие инструменты в комплексе ГМЗ есть, их обеспечивают подключаемые из системы Mathematica функции Volume, Integrate, NIntegrate, Total.

Приведем конкретные значения для рассматриваемых примеров. Оценки выполнялись по объемам, считая, что имеются разные тела, границы которых определены в прямоугольнике $[x_{\text{Min}}, x_{\text{Max}}] \times [y_{\text{Min}}, y_{\text{Max}}]$ плоскостью $z = 0$ и поверхностями z_{Surf} . Для пространственного объекта с границей z_{SurfA} (эталонное поле нижней поверхности) имеем рассчитанный по аналитической функции уравнения поверхности объем $v_{\text{SurfA}} = 135163$. Далее сопоставление проводится с сеточными функциями, поэтому также приведем значения v_{SurfAi} и относительные погрешности e_{SurfAi} , рассчитанные по значениям на квадратных сетках. В расчетах с шагом сетки h_i , используя соответствующий модуль компонента «Аппроксимация», получаем: $h_1 = 1,0$, $v_{\text{SurfA1}} = 136777$, $e_{\text{SurfA1}} = -1,12$ %; $h_2 = 0,5$, $v_{\text{SurfA2}} = 135900$, $e_{\text{SurfA2}} = -0,54$ %; $h_3 = 0,25$, $v_{\text{SurfA3}} = 135570$, $e_{\text{SurfA3}} = -0,30$ %.

Учитывая приведенные результаты, далее оценки погрешностей рассчитываются с $h = 0,5$. Результаты расчетов объемов v_{Appr} и относительных погрешностей e_{Rr} на эталонной сеточной функции следующие: для варианта восстановления по системе 97+10 точек $v_{\text{Appr1}} = 136031$, $e_{\text{Rr1}} = -0,64$ %; по системе 97+10+12 точек $v_{\text{Appr2}} = 135318$, $e_{\text{Rr2}} = 0,52$ %; по системе 97+10+20 точек $v_{\text{Appr3}} = 135381$, $e_{\text{Rr3}} = -0,26$ %.

Отдельно надо отметить, что модули компонента «Аппроксимация» позволяют рассчитывать погрешности на любом выбранном участке или фрагменте по разным признакам или характеристикам цифрового поля, например, вблизи локальных экстремумов, на автоматически определяемых участках роста. В частности, для примеров выше следует обратить внимание, что приведенные погрешности e_{Rr} (по уровням они находятся в пределах точности воспроизведения эталона) на самом деле не являются

удовлетворительными. Пояснение – превышения и занижения «компенсируют» итог. А по факту, например, для варианта $97+10+20$ точек имеем объем превышений 2498,27, объем занижения 2979,33. В сумме для всего объема получаем $2498,27-2979,33 = -481,06$, что и дает приведенное выше $eRr3 = -0,26$ %.

Примеры анализа данных с помощью искусственных нейронных сетей

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 вычислительных экспериментов. Приведенные ниже результаты и примеры анализа данных с помощью искусственных нейронных сетей получены в комплексе ГГМЗ с использованием соответствующих функций, которые доступны в версиях 11 и 12 Mathematica [WOLFRAM MATHEMATICA].

Рассмотрим средства и несколько вариантов применения модулей «искажения» эталонной модели, примеры предобработки представительных данных, визуализации профилей с использованием инструментов компонента ГПН. Проиллюстрируем варианты работы с модулем «искажения» эталонной модели. На примерах профилей по рассмотренным выше эталонным поверхностям покажем несколько методов имитации наблюдений и получения представительных данных. Для этой цели используем рассчитанные из аналитических выражений на выбранном профиле значения, к которым добавим «шумы», используя разные генераторы случайных чисел [Таранчук, 2015б]. Приведенные ниже результаты вычислительных экспериментов выполнялись для одного из типичных профильных разрезов, визуализация дана на рис. 3, 4. На рис. 3 слева показана часть объема для верхней поверхности (кровли) с отсечением по профилю $\{50,140\}-\{250,0\}$ (вариант визуализации с профильным разрезом). На иллюстрации справа показан объем для нижней поверхности (подошвы) с отмеченными по такому же профилю точками замеров. Напомним, что такой разрез рассматривался в статье [Таранчук, 2017] – иллюстрация рис. 7 «Разрез $y > 0.7 \times x + 175$ ». Чтобы детализировать уровни объемов, на рис. 3 применена послынная чередующаяся раскраска, в обоих случаях принято разбиение на 8 интервалов по высоте. Одномерные графики, иллюстрирующие уровни поверхностей, показаны на рис. 4, точками (соединены сплошными линиями) отмечены рассчитанные по эталонным поверхностям (когда замеры без погрешностей) значения уровней в «узлах наблюдений». Треугольниками отмечены данные, интерпретируемые наблюдениями с погрешностями – они получены с добавками генерируемых соответствующими датчиками шумов [Таранчук, 2015б]. Для поверхности кровли $zSurfB$ базовые значения в узлах искажаются с использованием выборки, генерируемой функциями Mathematica `RandomVariate` (реализация случайной переменной) и `NormalDistribution [0, 0,7]` (нормальное распределение со средним 0 и стандартным отклонением 0,7); отметки показаны треугольниками с вершинами вверх, комплект данных – `dataProfB`. Для поверхности подошвы $zSurfA$ начальные значения искажаются с использованием выборки, генерируемой функциями `RandomVariate` и `UniformDistribution [{-0,8, 0,8}]` (равномерное статистическое распределение, дающее значения от -0,8 до 0,8), отметки показаны треугольниками с вершинами вниз, комплект данных – `dataProfA`.

Отдельно следует обратить внимание на специально подобранный для экспериментов профиль и задаваемые величины искажений. Исходные эталонные функции изменяются на профиле в следующих диапазонах: $zSurfA$ – от 1 до примерно 7, и $zSurfB$ – от 3 до примерно 10. При этом есть небольшие возмущения от базового уровня (ленты) вблизи $x=150$ и существенные отклонения от уровня ленты вблизи $x=75$ и $x=210$. Величины этих отклонений соразмерны диапазонам изменения исходных функций. Относительно искажения (шумов) – есть незначительные, и есть такие, которые можно классифицировать как «выбросы».

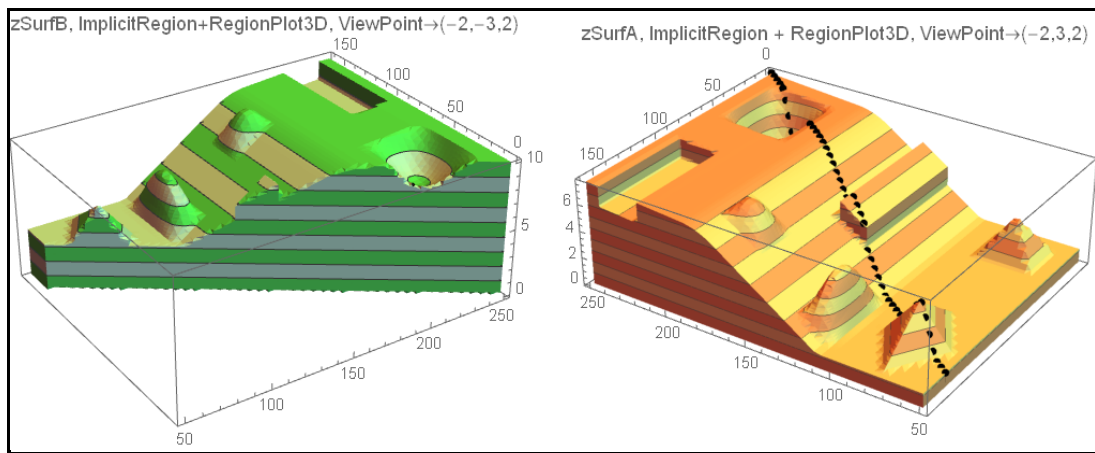


Рис. 3. Изображения моделируемых объемов и профилей замеров
 Fig. 3. Images of simulated volumes and profiles of measurements

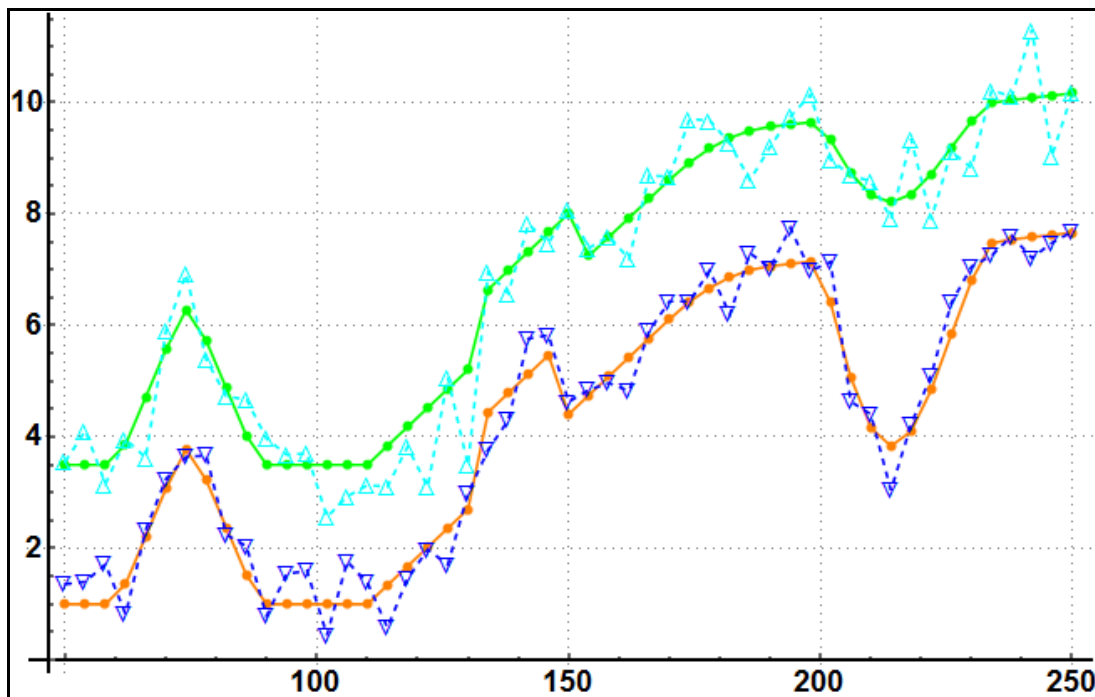


Рис. 4. Базовые и «наблюдаемые» данные на профилях dataProfA, dataProfB
 Fig. 4. Basic and "observed" data on profiles dataProfA, dataProfB

Данные с «искажениями» могут использоваться для разных целей, соответственно будут применяться разные методы предобработки. Важно следующее: обсуждаемые ниже методы интеллектуального анализа «не знают» первоисточник, эталонные функции. Задача эксперта – максимально воспроизвести (угадать) оригинал по имеющимся данным «замеров». Кстати, распределения, названные в тексте эталонными, могут и не быть оригиналом.

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 рис. 5, 6.

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. В частности, распределение наблюдаемого параметра вдоль профиля надо передавать в исходные данные компьютерной теоретической модели в виде гладкой

функции. Заметим, что и эталонные данные рис. 4 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 простейшие варианты предобработки данных, подразумевая получение аппроксимирующей гладкой функции – задача сглаживания [Taranchuk, 2019]. На рис. 5 представлены результаты расчетов после обучения искусственной нейронной сети с использованием функции Mathematica NetTrain и заданием метода ADAM [NetTrain]. В методе Adam запрограммирован 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.

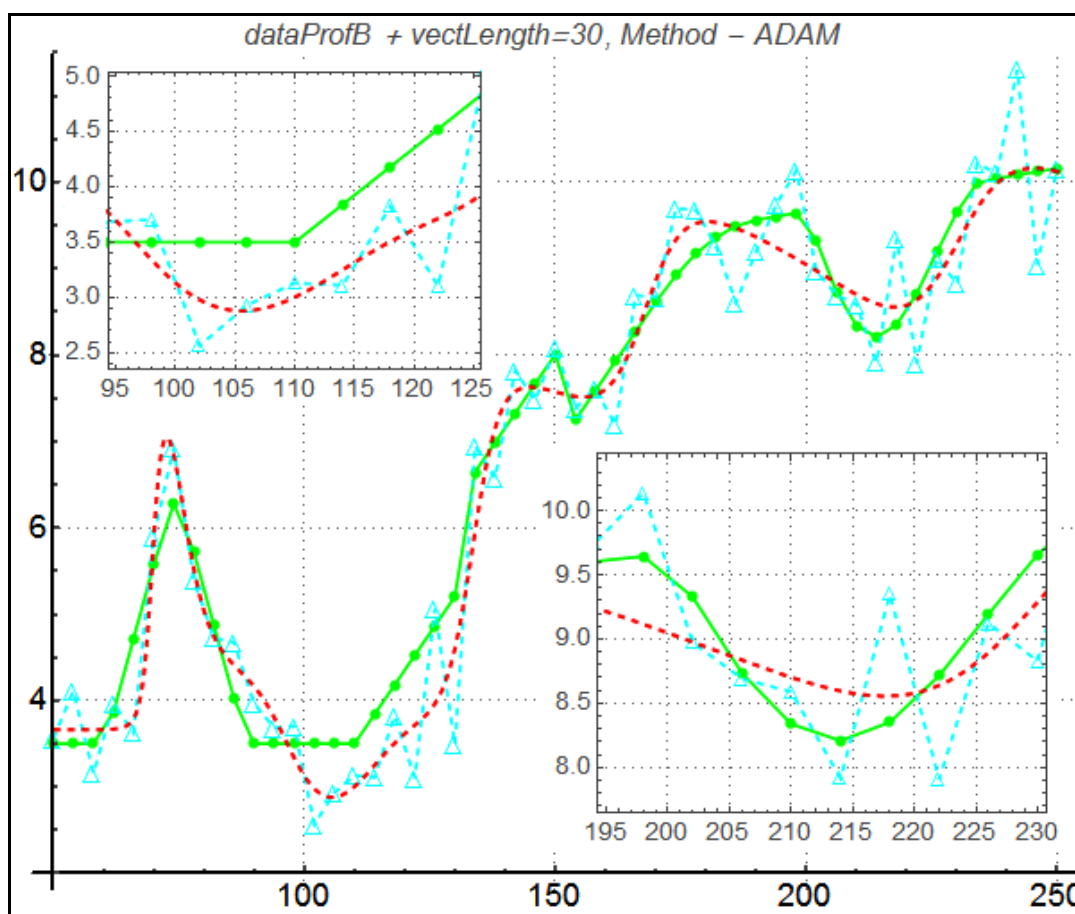


Рис. 5. Результат сглаживания набора данных dataProfB с применением нейронной сети
 Fig. 5. The result of smoothing the dataProfB dataset using a neural network

На рис. 5 представлены результаты, полученные с минимумом настроек, что сделано специально, так как в данном изложении задачей является демонстрация возможностей как таковых, а не детализация получения лучшего (адаптированного) решения с минимальными затратами времени на расчеты, которые для рассмотренных вариантов требуют только нескольких десятков секунд. В реализациях примеров коды включают задание сети netB конструкцией NetChain[{vectLength, Tanh, vectLength, Tanh, 1}, "Input"->"Scalar", "Output"->"Scalar"] и последующее получение dataProfBn конструкцией NetTrain[netB, dataProfB, Method->"ADAM"]. Полученное сглаженное приближение dataProfBn показано на рис. 5 пунктирной красной линией. Дополнительно в поле графика даны две вставки ситуационного плана (выкопировки), которые детализируют графики.

Расчеты выполнялись с разными значениями `vectLength` из интервала 10–100. Для комплекта данных `dataProfB` лучшие результаты получаются при `vectLength`~30; при значениях меньше 20 и больше 50 получаемые приближения заметно хуже воспроизводят оригинальные возмущения вблизи $x \approx 75, 150, 215$. Также проводились расчеты с использованием метода `RMSProp` – стохастический градиентный спуск с использованием адаптивной скорости обучения, получаемой из экспоненциально сглаженного среднего значения градиента. Для рассматриваемого профиля оптимальным в этом методе оказался вариант `vectLength`=80, график подобен приведенному на рис. 5. Однако следует отметить, что для метода `RMSProp` в сравнении с `ADAM` имеем значительно более «узкий» интервал пригодного `vectLength`; для метода `RMSProp` при `vectLength` меньше 75 и больше 85 воспроизводимость эталона заметно хуже.

На рис. 6 представлены результаты расчетов после обучения искусственной нейронной сети с использованием `NetTrain` методом `ADAM` для данных `dataProfA`. Сглаженное приближение иллюстрируется пунктирной малиновой линией. Расчеты выполнялись с разными значениями `vectLength`. Лучшие результаты получаются при `vectLength`~70, при значениях меньше 60 и больше 80 получаемые приближения заметно хуже воспроизводят оригинальные возмущения-волны. Также проводились расчеты с опцией метода `RMSProp`. Для рассматриваемого профиля оптимальным в этом методе оказался вариант `vectLength`=60.

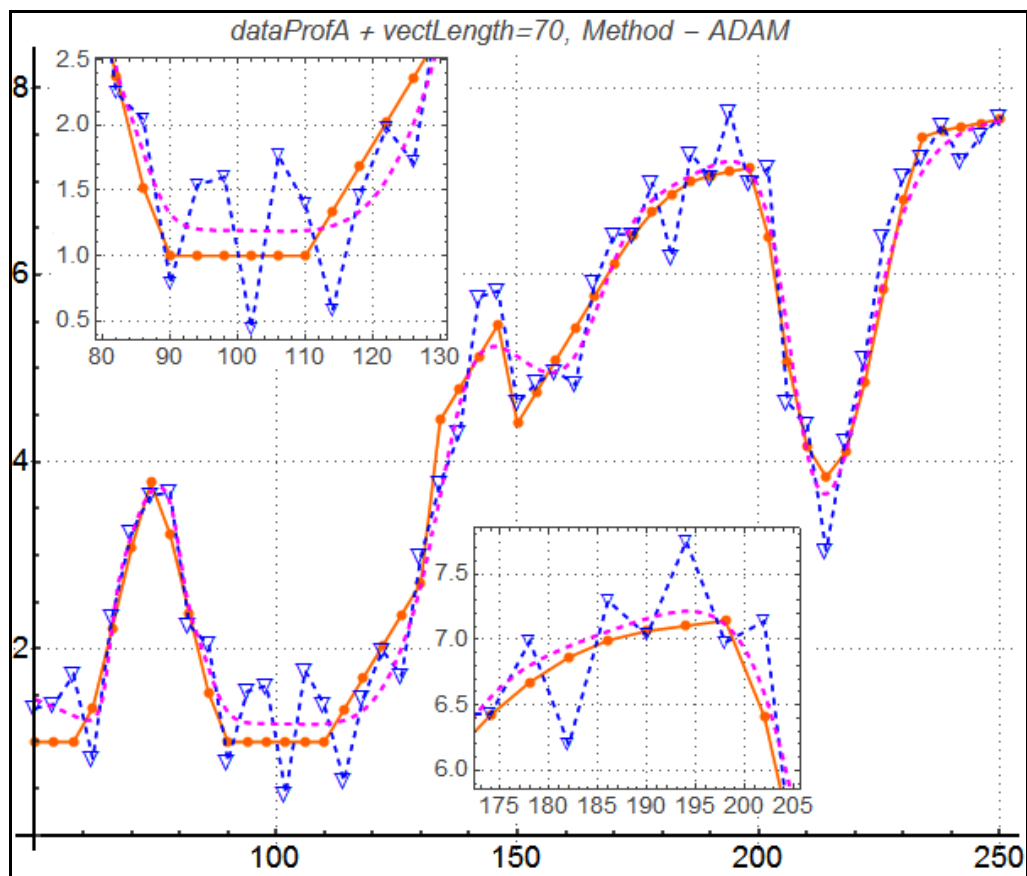


Рис. 6. Результат сглаживания набора данных `dataProfA` с применением нейронной сети
Fig. 6. The result of smoothing the `dataProfA` dataset using a neural network

Заклучение

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, которые были искажены разными способами. Однако следует понимать, что обучение сети – работа с черным ящиком, в каждом конкретном случае при получении лучшего воспроизведения эталона присутствует субъективность, требуется достаточно большая предварительная серия расчетов, относительно продолжительные расчеты. С этой точки зрения для задач типа «сглаживание данных» применение методов, например, скользящего среднего, скользящей медианы [Taranchuk, 2019] обеспечивает повторяющиеся результаты и существенно менее затратно в вычислениях.

Список литературы

1. Дышленко С.Г. 2016. Анализ и разработка характеристик качества геоданных. Перспективы науки и образования. 2(20): 23–27.
2. Савиных В.П., Цветков В.Я. 2014. Геоданные как системный информационный ресурс. Вестник Российской академии наук. 84(9): 826–829.
3. Таранчук В.Б. 2006. Построение цифровых геолого-экологических моделей в системе ГБД-э. Геоэкология Беларуси, Минск, Право и экономика: 72–96.
4. Таранчук В.Б. 2014. О создании интерактивных образовательных ресурсов с использованием технологий Wolfram. Информатизация образования. 1(73): 78–89.
5. Таранчук В.Б. 2015а. Введение в язык Wolfram. Минск, БГУ, 51.
6. Таранчук В.Б. 2015б. Основы программирования на языке Wolfram. Минск, БГУ, 49.
7. Таранчук В.Б. 2015в. Возможности и средства Wolfram Mathematica для разработки интеллектуальных обучающих систем. Научные ведомости Белгородского государственного университета. Серия: Экономика. Информатика. 1(198/33): 102–110.
8. Таранчук В.Б. 2017. Интегрированный программный комплекс тестировщика геологических моделей. Научные ведомости Белгородского государственного университета. Серия: Экономика. Информатика. 43(16): 148–159.
9. Таранчук В.Б. 2018. Методы и инструментарий оценки точности компьютерных геологических моделей. Вестник БрГТУ, 2: 64–69.
10. Шайтура С.В. 2015. Интеллектуальный анализ геоданных. Перспективы науки и образования. 6 (18): 24–30.
11. NetTrain. Available at: <https://reference.wolfram.com/language/ref/NetTrain.html>.
12. SURFER. Available at: <http://www.goldensoftware.com/products/surfer>.
13. Taranchuk V., Kulinkovich V. 2015a. On the preparation and distribution of interactive graphics applications using Mathematica. Computer Algebra Systems in Teaching and Research. Vol. V. University of Natural Sciences and Humanities. Poland: 380–387.
14. Taranchuk V., Kulinkovich V. 2015b. On programming interactive graphic applications in Mathematica system. Computer Algebra Systems in Teaching and Research. Vol. V. University of Natural Sciences and Humanities. Poland: 388–395.
15. Taranchuk V.B. 2017. The integrated computer complex of an estimation and adapting of digital geological models. Studia i Materiały. 2 (14): 73–86.
16. Taranchuk V.B. 2018. New computer technologies, analysis and interpretation of geodata. MATEC Web of Conferences IPICSE-2018. 251, 04059. VI International Scientific Conference “Integration, Partnership and Innovation in Construction Science and Education” (IPICSE-2018): 1–8.
17. Taranchuk V.B. 2019. Examples of the use of artificial neural networks in the analysis of geodata. Open Semantic Technologies for Intelligent Systems: Research Papers Collection. Issue 3: 225–230.
18. Tsvetkov V.Ya., Lobanov A.A. 2014. Big Data as Information Barrier. European Researcher. 78(7): 1237–1242.
19. Wolfram mathematica. Wolfram computation meets knowledge. Available at: <http://www.wolfram.com/mathematica>.

References

1. Dyshlenko S.G. 2016. Analiz i razrabotka harakteristik kachestva geodannyh [Analysis and design of quality attributes of GEODATA]. Perspektivy nauki i obrazovaniya. 2(20): 23–27.
2. Savinyh V.P., Cvetkov V.Ya. 2014. Geodannye kak sistemnyi informacionnyi resurs



[GEODATA as a system information resource]. Vestnik Rossiiskoi akademii nauk. 84(9): 826–829.

3. Taranchuk V.B. 2006. Postroenie tsifrovyykh geologo-ekologicheskikh modeley v sisteme GBD-e [Construction of digital eco-geological models in the GBD-e system]. Geoekologiya Belarusi, Minsk, Pravo i ekonomika: 72–96.

4. Taranchuk V.B. 2014. O sozdanii interaktivnykh obrazovatel'nykh resursov s ispol'zovaniem tekhnologiy Wolfram [About creating interactive educational resources using Wolfram technologies]. Informatizatsiya obrazovaniya. 1(73): 78–89.

5. Taranchuk V.B. 2015a. Vvedenie v yazyk Wolfram [Introduction to the Wolfram language]. Minsk, BGU, 51.

6. Taranchuk V.B. 2015b. Osnovy programmirovaniya na yazyke Wolfram [Bases of programming in the Wolfram language]. Minsk, BGU, 49.

7. Taranchuk V.B. 2015c. Wolfram Mathematica opportunities and tools for development of intelligent tutoring systems. Belgorod State University Scientific Bulletin. Ser. Economics. Information technologies. 1(198), t. 33: 102–110.

8. Taranchuk V.B. 2017. Integrated programming complex of the tester of geological models. Belgorod State University Scientific Bulletin. Ser. Economics. Information technologies. 43(16): 148–159.

9. Taranchuk V.B. 2018. Metody i instrumentarii ochenki tochnosti komp'yuternykh geologicheskikh modelei [Methods and tools for assessing the accuracy of computer geological models]. Vestnik BrGTU, 2: 64–69.

10. Shaitura S.V. 2015. Intel'ktual'nyi analiz geodannykh [GEODATA mining]. Perspektivy nauki i obrazovaniya. 6 (18): 24–30.

11. NetTrain. Available at: <https://reference.wolfram.com/language/ref/NetTrain.html>.

12. SURFER. Available at: <http://www.goldensoftware.com/products/surfer>.

13. Taranchuk V., Kulinkovich V. 2015a. On the preparation and distribution of interactive graphics applications using Mathematica. Computer Algebra Systems in Teaching and Research. Vol. V. University of Natural Sciences and Humanities. Poland: 380–387.

14. Taranchuk V., Kulinkovich V. 2015b. On programming interactive graphic applications in Mathematica system. Computer Algebra Systems in Teaching and Research. Vol. V. University of Natural Sciences and Humanities. Poland: 388–395.

15. Taranchuk V.B. 2017. The integrated computer complex of an estimation and adapting of digital geological models. Studia i Materiały. 2 (14): 73–86.

16. Taranchuk V.B. 2018. New computer technologies, analysis and interpretation of geodata. MATEC Web of Conferences IPICSE-2018. 251, 04059. VI International Scientific Conference “Integration, Partnership and Innovation in Construction Science and Education” (IPICSE-2018): 1–8.

17. Taranchuk V.B. 2019. Examples of the use of artificial neural networks in the analysis of geodata. Open Semantic Technologies for Intelligent Systems: Research Papers Collection. Issue 3: 225–230.

18. Tsvetkov V.Ya., Lobanov A.A. 2014. Big Data as Information Barrier. European Researcher. 78(7): 1237–1242.

19. Wolfram mathematica. Wolfram computation meets knowledge. Available at: <http://www.wolfram.com/mathematica>.

Ссылка для цитирования статьи

Reference to article

Таранчук В.Б. 2019. Методы и примеры интеллектуальной обработки данных для геологических моделей. Научные ведомости Белгородского государственного университета. Серия: Экономика. Информатика. 46 (3): 511–522. DOI 10.18413/2411-3808-2019-46-3-511-522.

Taranchuk V.B. 2019. Methods and examples of intelligent data processing for geological models. Belgorod State University Scientific Bulletin. Economics. Information technologies. 46 (3): 511–522 (in Russian). DOI 10.18413/2411-3808-2019-46-3-511-522.