

УДК 519.254:550.822(476)

А. Н. Маевская¹, Н. Н. Шешко², М. А. Богдасаров³¹магистр геогр. наук, аспирант каф. географии и природопользования
Брестского государственного университета имени А. С. Пушкина²канд. техн. наук, доц., нач. научно-исследовательской части
Брестского государственного технического университета³д-р геол.-минерал. наук, проф., член-кор. НАН Беларуси,
зав. каф. географии и природопользования
Брестского государственного университета имени А. С. Пушкинаe-mail: ¹maevskaya.anna@inbox.ru**АЛГОРИТМ ОБРАБОТКИ ДАННЫХ ГЕОЛОГИЧЕСКИХ ИЗЫСКАНИЙ
С ПРИМЕНЕНИЕМ ГИС-ТЕХНОЛОГИЙ (НА ПРИМЕРЕ МАТЕРИАЛОВ
БУРОВОЙ ИЗУЧЕННОСТИ ТЕРРИТОРИИ БРЕСТСКОЙ ОБЛАСТИ)**

Рассмотрена проблема «больших данных», подходы к их классификации, а также основные методы, применяемые при их обработке. Приведен анализ наиболее распространенных способов предварительного статистического анализа пространственных данных. На примере информации, полученной в результате геологических изысканий, проведенных на территории Брестской области, разработан алгоритм, позволяющий с применением геоинформационных технологий осуществлять обработку данных геологического бурения. Представленный алгоритм включает в себя несколько последовательных этапов и учитывает существующие подходы к анализу пространственных данных. Для автоматизации процессов обработки информации создан набор инструментов «processing of geological data».

MAYEVSKAYA A. N., SHESKO N. N., BOGDASAROV M. A.**ALGORITHM FOR PROCESSING DATA OF GEOLOGICAL SURVEYS USING GIS
TECHNOLOGIES (ON THE EXAMPLE OF THE MATERIALS OF DRILLING STUDY
OF BREST REGION TERRITORY)**

The problem of «big data», the approaches to their classification, as well as the main methods used in their processing are considered. The analysis of the most common methods of preliminary statistical analysis of spatial data is presented. On the example of information obtained as a result of geological surveys carried out on the territory of the Brest region, an algorithm has been developed that allows, using geoinformation technologies, to process geological drilling data. The presented algorithm includes several sequential stages and takes into account existing approaches to the analysis of spatial data. A set of tools «processing of geological data» has been created to automate information processing processes.

Введение

В настоящее время в различных областях человеческой деятельности отмечается активный рост объемов накапливаемой информации и увеличение скорости ее производства. В связи с этим высокую актуальность приобретает проблема «больших данных». Зачастую понятие big data, или «большие данные» (калька с англ.), употребляют, когда речь идет о данных больших объемов или о технологиях обработки и использования, методах поиска необходимой информации, которые применяются при работе с ними [1; 2]. Традиционно «большие данные» разделяют на две категории: 1) данные, получаемые в результате проведения научных наблюдений и экспериментов; 2) данные, получаемые из социальной сферы (социальные сети, интернет, экономика) [3].

Некоторые авторы отмечают, что значительная часть «больших данных» относится к типу геопространственных. Такие данные могут быть представлены тремя формами: растровые данные (геоизображения), векторные данные (точки, линии, многоугольники), графические данные [4].

Технологии, применяемые при обработке «больших данных», призваны выполнять три основные операции: 1) обработка больших по сравнению со «стандартными» сценариями объемов данных; 2) умение работать с быстро поступающими данными в больших объемах; 3) умение работать как со структурированными, так и неструктурированными данными параллельно в разных аспектах [5].

Работа с большими потоками информации осуществляется по принципу «Volume» – объем; «Velocity» – быстродействие обработки; «Variety» – разнообразие сведений, хранящихся в массиве. Сегодня к этим трем принципам присоединяется четвертый – «Value», что обозначает ценность информации. Следовательно, информация должна нести теоритическую и практическую значимость, что, в свою очередь, будет оправдывать затраты на ее хранение и обработку.

В целом для обработки больших массивов информации применяются такие техники и методы, как машинное обучение, имитационное моделирование, прогнозная аналитика, искусственные нейронные сети, пространственный и статистический анализ, визуализация аналитических данных и др.

Постановка проблемы

В результате непрерывного роста скорости накопления «больших данных» возрастает актуальность разработки алгоритмов автоматизированной обработки таких наборов информации применительно к решению конкретных задач, возникающих в различных научных направлениях.

Целью работы является создание алгоритма автоматизированной обработки сведений, полученных по результатам геологических изысканий, проводимых в разное время специалистами РУП «Белгеология». Имеющиеся данные отражают информацию о строении четвертичных отложений территории Брестской области и включают в себя около 5 000 точек.

Поскольку рассматриваемые в нашем исследовании данные относятся к типу геопространственных, качественная их обработка невозможна без применения специализированных программных ГИС-продуктов. В данном случае был использован настольный пакет ArcGIS 10.5.

Важно подчеркнуть, что обработка имеющегося набора данных, а также выполнение на их основе моделирования геологических поверхностей с применением традиционных способов обработки пространственной информации (т. е. вручную) была бы практически невыполнимой задачей. Это обусловлено тем, что картографирование геологических подземных структур – сложный процесс. Для построения только одной структурной геологической карты, вне зависимости от варианта ее создания (метод треугольников, метод профилей, метод схождения и др.), необходимо выполнить ряд трудоемких шагов. Во-первых, нанести имеющиеся структурные точки на топографическую основу с указанием абсолютных отметок залегания пласта. Во-вторых, провести интерполирование, качество которого в значительной степени зависит от навыков специалиста, а также от ряда других субъективных факторов. В-третьих, на основе интерполированной поверхности построить изолинии. Все вышеперечисленное – далеко не полный список действий, которые необходимо произвести для построения картографической модели вручную. Кроме того, на каждом этапе ручной отрисовки могут допускаться ошибки, на исправление которых уходит много времени. Еще одним минусом традиционной обработки пространственных данных является то, что удаление точек с аномалиями зачастую производится «на глаз», без учета основных принципов пространственной статистики, или вовсе не осуществляется. Это, в свою очередь, может приводить к значительным искажениям картографируемого явления [6–8].

Результаты и их обсуждение

В работе представлен алгоритм, предусматривающий последовательность обработки данных, отражающих результаты геологических изысканий, выполненных на территории Брестской области в программном продукте ArcGIS 10.5 (рисунок 1). Рассмотрим более подробно этапы реализации алгоритма.



Рисунок 1. – Алгоритм, предусматривающий последовательность обработки данных, отражающих результаты геологических изысканий

I. Подготовительный этап. В рамках данного этапа с целью сбора данных, отражающих основные характеристики, получаемые в ходе геологического бурения, было выполнено проведение инвентаризационных работ. На основе собранных материалов в программной среде Microsoft Access была создана полная инвентаризационная

база данных геологических скважин. Разработанная база включает следующие сведения: номер скважины, координаты расположения (широта и долгота), альтитуда устья, забой, год бурения, глубина залегания слоев, сведения о стратиграфии и литологии, сведения о районе расположения скважины и локальной структуре, к которой она приурочена. Наличие географических координат каждой скважины позволяет внедрять спроектированную базу в ГИС-оболочки с функцией автоматического создания точечного слоя.

Реализованная база данных соответствует требованиям, предъявляемым к созданию баз данных ГИС, и является: 1) полной, достаточно подробной для предполагаемого создания картографического продукта, включает все необходимые сведения для осуществления анализа или математико-картографического моделирования исследуемого объекта; 2) позиционно точной, абсолютно совместимой с другими данными, которые могут добавляться в нее; 3) достоверной, правильно отражающей характер явлений; 4) легко обновляемой.

II. Этап проверки корректности исходных данных. В рамках данного этапа был осуществлен поиск «выбросов» в данных, т. е. опорных точек, которые сильно выбиваются из последовательности, не вписываются в модель по какой-либо причине.

Для обнаружения аномальных значений в имеющемся наборе данных рассматривалось применение нескольких методов, встречающихся в научной литературе по статистике и геостатистике [9–16].

1. *Использование инструментов исследовательского (научного) анализа (ESDA), представленных в модуле ArcGIS Geostatistical Analyst.* В частности, для поиска точек с аномальными значениями применяется три основных инструмента: построение гистограмм, построение облака вариограммы/ковариации, построение карт Вороного [17]. Однако применение данного набора инструментов в условиях имеющихся данных не позволило получить корректных результатов. Так, например, инструмент «построение облака вариограммы/ковариации» по умолчанию работает с наборами данных, в которые входит максимум 500 точек данных, что гораздо меньше, чем в рассматриваемом в данной работе наборе данных.

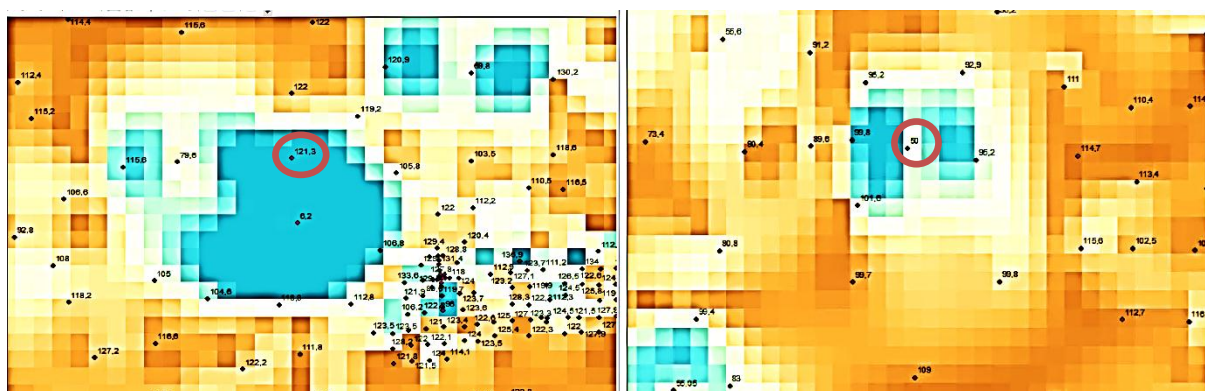
При работе с инструментом «гистограмма» отмечается наличие участков, на которых в качестве «выбросов» отображаются точки с корректными по сравнению с окружающим набором данных значениями. Кроме того, при использовании данного метода невозможно выполнение последующего одновременного автоматического удаления точек с «выбросами». Построение карт Вороного на основе кластерного метода и метода энтропии также не позволило выявить весь набор точек с выпадающими значениями.

2. *Поиск выбросов на основе вычисления среднеквадратичного отклонения.* Для реализации данного способа было выполнено построение раstra среднеквадратичного отклонения с применением метода «скользящего окна» (*Spatial Analyst – Neighborhood – Focal statistic*) с заданием разных типов (круг, кольцо, клин, прямоугольник), а также параметров окрестности.

Применение разных параметров окрестности влияет на величину среднеквадратичного отклонения, но зоны с высоким среднеквадратичным отклонением при этом остаются неизменными. Таким образом, в дальнейшем рассматривался вариант проведения классификации полученного раstra и удаление ячеек с высокими показателями дисперсии. Однако после проведения более детального визуального анализа полученных поверхностей, можно было заметить, что на некоторых участках в число ячеек, отображающих данные с «выбросами», попадают значения, не требующие удаления (рисунок 2а), либо точка с экстремумом попадает в пределы участка раstra с нормальными показателями данного параметра (рисунок 2б).

1. Поиск «выбросов» в данных с применением инструмента «кригинг» (kriging). В качестве еще одного варианта для поиска аутлайеров было рассмотрено проведение интерполяции способом кригинга, который представляет собой метод, основанный на статистических характеристиках входных данных, таких как среднее значение и дисперсия [18; 19].

Идея применения данного метода для поиска «выбросов» заключается в следующем: при построении модели интерполированной поверхности данным инструментом точки с аномальными значениями не учитываются в построении. Таким образом, разница между исходным значением в данной точке «выброса» и проинтерполированным значением по методу кригинга будет очень сильно отличаться. Таким образом, данные полученные по ошибкам интерполяции с применением данного метода могут быть использованы для поиска «выпадающих» значений.



а) пример, когда в число ячеек, отображающих данные с выбросами, попадают точки, не требующие удаления

б) пример, отражающий случай, когда точка с экстремумом попадает в ячейку с нормальными показателями

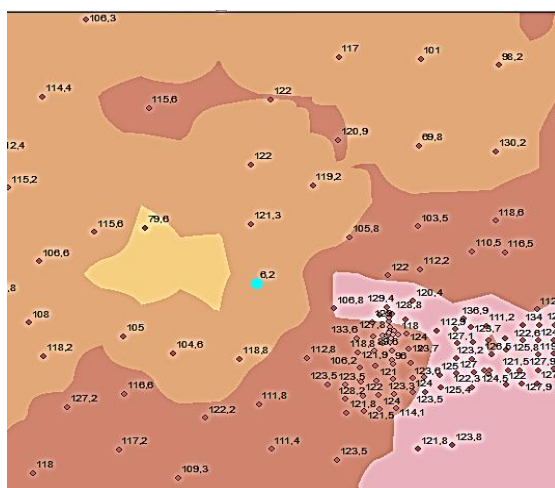
Рисунок 2 – Пример использования инструмента «Focal statistic» для поиска выпадающих значений

В данной работе использован метод ординарного кригинга, представленный в наборе инструментов Spatial Analyst. В ходе построения были использованы разные параметры создания интерполяционной модели. Наибольшее влияние среди возможных задаваемых характеристик оказал радиус поиска.

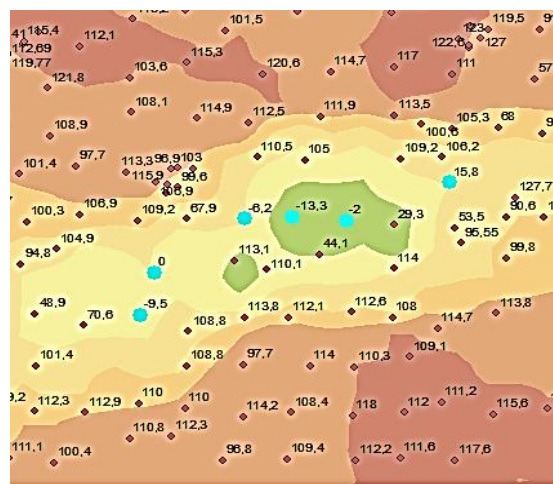
Затем полученные в ходе интерполяции результаты были добавлены в исходную таблицу атрибутов с использованием инструмента «Extract». После чего в калькуляторе поля было выполнено вычитание полученных в ходе интерполяции результатов из исходных данных, таким образом выявлены показатели ошибки между проинтерполированным значением и значением в исходной точке.

Выполнив анализ ошибок, получаемых разными параметрами кригинга, выбор остановился на мерах, задаваемых инструментом по умолчанию. Применение стандартных характеристик для имеющегося набора данных показало наиболее приемлемые результаты. В дальнейшем на основе атрибута с рассчитанной разницей были выбраны опорные точки с высокими показателями ошибки (рисунок 3).

Таким образом, именно метод ординарного кригинга проявил себя наиболее пригодным для анализа «выбросов» в нашем наборе данных (рисунок 3а). Применение этого способа не затрачивает много времени. Но стоит отметить, что при его использовании, кроме абсолютных ошибок отображаются и скважины, приводящие к локальным искажениям (рисунок 3 б).



а) пример скважины,
в которой ошибка интерполяции
превысила значение 50



б) пример, когда в ходе выборки
отображаются значения,
приводящие к локальным искажениям

**Рисунок 3. – Пример использования инструмента кригинга
для поиска «выбросов» в данных**

III. Этап моделирования граничных геологических поверхностей. Данный этап включал в себя несколько последовательных шагов и выполнялся с учетом существующих подходов к созданию геологических структурных поверхностей [20–23]:

Шаг 1. Выбор метода моделирования. В современном научном исследовании сформировалось два основных логических метода формулирования моделей – индуктивный и дедуктивный [24] – вне зависимости от того, о какой модели идет речь. Для проектирования геолого-генетической модели, представленной в данной работе, был выбран метод от общего к частному, т. е. построение от более крупных стратиграфических подразделений к наименьшим, т. к. именно построение моделей крупных структур лучше поддается процессам автоматизации.

Шаг 2. Выбор инструмента интерполяции. Он осуществлялся из представленных в наборе инструментов модуля *Spatial Analyst* программного продукта ArcGIS 10.5. При выборе подходящего для обработки имеющегося набора данных инструмента учитывались следующие особенности: количество опорных точек; плотность их расположения по территории; природа данных, планируемая область применения. При этом, как отмечают некоторые авторы, наиболее важное значение среди вышеописанных параметров играет количество опорных точек. Если в основе построений лежит густая сеть опорных точек, эффективность использования ГИС-технологий для выявления и анализа геологического строения территории является высокой. В то же время при разреженной сети исходных данных применение методов автоматической интерполяции для построения моделей погребенного рельефа является невозможным [25]. В целом все виды интерполяции, представленные в данном наборе инструментов, показали корректные результаты, что обусловлено достаточно густой сетью опорных точек.

Шаг 3. Моделирование. Он состоял из нескольких подэтапов:

- 1) подготовка исходных данных, когда выполнялось объединение стратиграфических подразделений до самого высокого стратиграфического уровня; в данном случае – отдел (*Data Management Tools – Generalization – Slove by attribute*);
- 2) выборка слоев из объединенной базы данных (*Analysis Tools – Extract – Select*);
- 3) построение grid-моделей кровли и подошвы стратиграфических горизонтов.

IV. Этап верификации построенных поверхностей. Он реализовывался с учетом основных принципов анализа растровых поверхностей в ГИС [26–28]. Этот этап необходим для проверки корректности построенных интерполяционных моделей.

Первоначально в рамках реализации данного этапа был осуществлен поиск ошибок моделирования. Для обнаружения ошибок была выполнена переклассификация полученных поверхностей, изначально смоделированных с одинаковым разрешением растра (*Spatial Analyst – Reclassify – Reclassify*) по значениям ячеек кровля, подошва (например, подошва голоцена и кровля плейстоцена). После этого с использованием инструмента *Map Algebra*, представленного в наборе инструментов *Spatial Analyst*, производилось вычитание данных поверхностей (рисунок 4):

$$R_o = r_{\Pi} - r_k,$$

где R_o – растр ошибок; r_{Π} – переклассифицированный растр подошвы вышележащего слоя; r_k – переклассифицированный растр кровли нижележащего слоя.

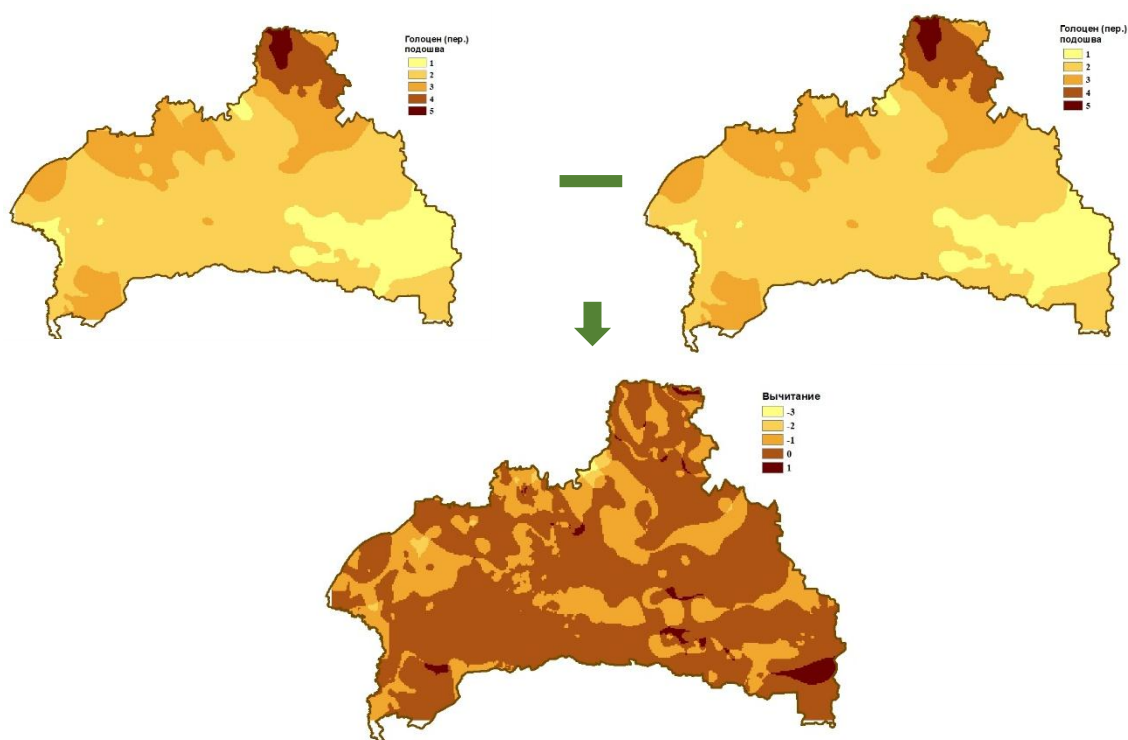


Рисунок 4. – Пример вычитания переклассифицированных поверхностей

Учитывая, что не каждая из стратиграфических единиц занимает всю территорию Брестской области, при вычитании растров в параметрах среды задавался «экстенд обработки», как у наибольшего по размерам растрового слоя. Таким образом, на месте отсутствующей поверхности при вычитании автоматически устанавливается значение «no data».

Далее растр со значением «no data» обрабатывается функцией *isNull*, в результате чего подставляются необходимые показатели из другого растра или значения, получаемые при вычитании нескольких растровых поверхностей. Например,

$$\text{Con} ("R_o" == 0), R_p, r_{\Pi} - r_{\text{кн}}),$$

где R_o – растр обработанный функцией *isNull*, R_p – результирующий растр, полученный в результате вычитания двух поверхностей; r_{Π} – растр подошвы слоя; $r_{\text{кн}}$ – растр кровли слоя, расположенного через одну стратиграфическую единицу.

Таким образом, если значение пиксела в растре R_0 равно нулю, то выражение записывает в результат (в пиксель) значение, полученное при вычитании растров подошвы и кровли, если условие не выполняется, в результат (пиксель) записывается значение, получаемое из других растров (дополнительного растра), который нужно подставить (например, кровли слоя, лежащего через один горизонт).

По результатам обработки растра формируются два варианта поверхностей:

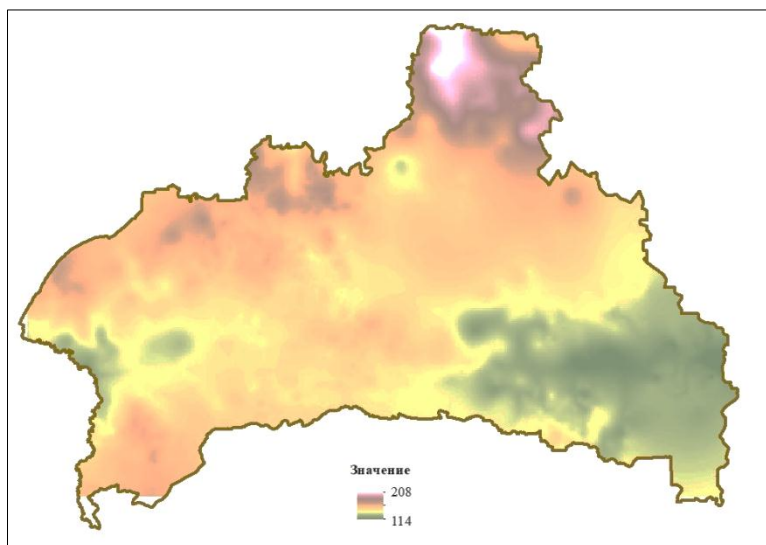
1) поверхности, где ошибки в построении отсутствуют: модель корректна, и ее дальнейшая обработка не требуется;

2) имеются ошибки в построении: поверхность смоделирована некорректно, и требуется проведение операции уравнивания.

Для уравнивания (усреднения) поверхностей с ошибками было рассмотрено два варианта, которые позволяют получить корректные результаты: во-первых, использование инструмента *Spatial Analyst – Lokal – Cell Statistics*, который позволяет вычислять статистику на основе значений ячеек из нескольких растров; во вторых, с использованием инструмента *Con (Conditional)*:

$$Con(("r_n" = "r_k"), "r_n" + "r_k" / 2).$$

Следовательно, если значения ячеек растра кровли будут отличаться от значений ячеек растра подошвы, то в результирующий растр будет записано среднее значение ячеек обоих растров (рисунок 5).



**Рисунок 5. – Поверхность,
полученная в результате уравнивания**

На завершающем этапе для автоматизации процессов обработки данных геологического бурения по разработанному алгоритму был создан набор инструментов «*processing of geological data*» (POGD), включающий себя ряд разработанных авторских моделей.

Заключение

В связи с увеличением объемов накапливаемой информации в различных сферах науки возникает необходимость разработки алгоритмов автоматизированной обработки информации. Учитывая, что значительная часть накапливаемых данных относится

к типу геопространственных, разработка методических аспектов их обработки тесным образом связана с применением ГИС-технологий.

В работе показан механизм создания алгоритма обработки данных, получаемых в результате геологических изысканий с возможностью автоматизации процесса обработки в программной среде ArcGIS 10.5. Разработанный алгоритм включает в себя несколько последовательных взаимосвязанных этапов и позволяет выполнять предварительный анализ пространственных данных с целью обнаружения «выбросов» в них, что в конечном итоге дает возможность построить более качественную картографическую модель за счет разработанного и реализованного алгоритма уравнивания поверхностей, позволяющего минимизировать ошибки, возникающие в ходе моделирования.

Представленный в работе алгоритм успешно апробирован применительно к набору данных буровой изученности территории Брестской области и может быть использован в ходе предварительной обработки материалов геологического бурения для других территорий.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Przulj, N. Network analytics in the age of big data / N. Przulj, N. Malod-Dognin // *Science*. – 2016. – Vol. 353, is. 6295. – P. 123–124.
2. Корнев, М. С. История понятия большие данные (big data): словари, научная и деловая периодика / М. С. Корнев // *Вестн. Рос. гос. гуманитар. ун-та. Сер. Литературоведение, языкознание, культурология*. – 2018. – № 4. – С. 81–85.
3. Significance and Challenges of Big Data Research / H. Jin [et al.] // *Big Data Research*. – 2015. – № 2. – P. 59–64.
4. Lee, J. Geospatial Big Data: Challenges and Opportunities / J. Lee, M. Kang // *Big Data Research*. – 2015. – № 2. – P. 74–81.
5. Как большие данные стали одной из самых интересных задач IT-индустрии [Электронный ресурс]. – Режим доступа: <https://postnauka.ru/specials/big-data>. – Дата доступа: 03.03.2020.
6. Структурная геология и геологическое картирование : учеб. пособие / С. А. Коваль [и др.]. – Воронеж : Воронеж. гос. ун-т, 2005. – 36 с.
7. Минова Н. П. Построение структурных карт : метод. указания и задания / Н. П. Минова. – Ухта : УГТУ, 2010. – 28 с.
8. Меритт, М. Моделирование подземных структур в ArcGIS / М. Меритт // *ArcReview*. – 2017. – № 2 (81). – С. 5–6.
9. Геостатистический анализ данных в экологии и природопользовании (с применением пакета R) / А. А. Савельев [и др.]. – Казань : Казан. федер. ун-т, 2012. – 120 с.
10. How to Deal with Outliers in Your Data [Electronic resource]. – Mode of access: <https://conversionxl.com>. – Date of access: 21.06.2019.
11. Barnett, V. Outliers in statistical data (Probability & Mathematical Statistics) / V. Barnett, T. Lewis. – New York : Willey Press, 1978. – 188 p.
12. Додонов, Ю. С. Устойчивые меры центральной тенденции: взвешивание как возможная альтернатива усечению данных при анализе времен ответов / Ю. С. Додонов, Ю. А. Додонова // *Психол. исслед.* – 2011. – № 5 (19). – С. 1–14.
13. Демьянов, В. В. Геоestatистика: теория и практика / В. В. Демьянов, Е. А. Савельева, Р. В. Арутюнян. – М. : Ин-т проблем безопасного развития атомной энергетики РАН, 2010. – 327 с.
14. Порохов, Г. С. Математические методы моделирования в геологии / Г. С. Порохов. – СПб : С.-Петербург. гос. горный ин-т, 2006. – 223 с.

15. Дэвис, Дж. С. Статистический анализ данных в геологии : пер. с англ. / Дж. С. Дэвис ; под ред. Д. А. Родионова. – М. : Недра, 1990. – 319 с.
16. Corizzo, R. Anomaly Detection and Repair for Accurate Predictions in Geodistributed Big Data / R. Corizzo, M. Ceci, N. Japkowicz // Science. – 2019. – № 16. – P. 18–35.
17. ArcGIS 9. Geostatistical Analyst : Manual User. – USA : ESRI Press, 2001. – 285 p.
18. Элементарное введение в геостатистику (проблемы окружающей среды и природных ресурсов) / М. Ф. Каневский [и др.]. – М. : ВИНТИ, 1999. – 132 с.
19. Kriging – описание алгоритма [Электронный ресурс]. – Режим доступа: <http://petroportal.ru>. – Дата доступа: 21.06.2019.
20. Крошинский, В. А. Геологическое картирование северного участка Минской возвышенности на основе ГИС-технологий / В. А. Крошинский // Современные проблемы геохимии, геологии и поисков месторождений полезных ископаемых : сб. материалов Междунар. науч. конф., посвящ. 100-летию со дня рождения акад. К. И. Лукашева, Минск, 23–25 мая 2017 г. : в 2 ч. / БГУ ; редкол.: О. В. Лукашев (отв. ред.) [и др.]. – Минск, 2017. – Ч. 1 : Геология и полезные ископаемые четвертичная геология, инженерная геология. – С. 36–38.
21. Курлович, Д. М. Использование ГИС-технологий для разработки баз геоданных и информационных проектов месторождений бурых углей и горючих сланцев Республики Беларусь / Д. М. Курлович // Международный конгресс по информатике: информационные системы и технологии : сб. материалов междунар. науч. конгр., Минск, 31 окт. – 3 нояб. 2011 г. / БГУ ; редкол.: С. В. Абламейко (отв. ред.) [и др.]. – Минск, 2011. – С. 188–193.
22. Ханжиян, Е. Геоинформационная система и база геоданных на основе карт «Атласа геологического строения и нефтегазоносности юга России» / Е. Ханжиян // ArcReview. – 2005. – № 1 (32). – С. 2–3.
23. Spatial analysis and evaluation of a coal deposit by coupling AHP & GIS techniques / N. Paraskevis [et al.] // International Journal of Mining Science and Technology. – 2019. – № 29. – P. 943–953.
24. Кошкарев, А. В. Геоинформатика / А. В. Кошкарев, В. С. Тикунов. – М. : Картгеоцентр, 1993. – 213 с.
25. Миронов, О. К. Геоинформационные технологии для составления крупномасштабных геологических карт г. Москвы / О. К. Миронов // Геоэкология. Инженерная геология. Гидрогеология. Геокриология. – 2011. – № 3. – С. 198–214.
26. Митчелл, Э. Руководство по ГИС-анализу / Э. Митчелл. – М., 2000. – 175 с.
27. Tomlin, D. C. Geographic Information Systems and Cartographic Modeling / D. C. Tomlin. – Portland : Book News. – 1990. – 249 p.
28. Кукарцев, В. В. Аппроксимация поверхности растровых карт в геоинформационной системе / В. В. Кукарцев, О. А. Антамошкин // Вестн. Сибир. гос. аэрокосм. ун-та. – 2012. – №3 (43). – С. 29–32.