

ЧАСТНЫЕ АСПЕКТЫ РАЗВИТИЯ ЦИФРОВОЙ МИНЕРАЛОГИИ: СОВРЕМЕННОЕ СОСТОЯНИЕ ПРОБЛЕМЫ

А. С. Василюк

Белорусский государственный университет, факультет географии и геоинформатики,
пр. Независимости 4, 220030 Минск, Республика Беларусь; andrew.vasiluck@gmail.com

Рассмотрены различные аспекты истории развития цифровой минералогии.

Ключевые слова: минерал; статистическая обработка данных.

Various aspects of the history of the development of digital mineralogy are considered.

Key words: mineral; static data processing.

Минералогия – одна из самых древних дисциплин геологического цикла, а потребность в использовании объектов её исследования – минералов – появилась ещё во времена существования австралопитеков. В кенийском археологическом памятнике Ломекви 3, на западном берегу оз. Рудольф, недалеко от места находки кениантропа (один из видов австралопитека), были найдены самые древние в мире каменные орудия ломеквийской культуры возрастом 3,3 млн лет, среди которых есть кремнистые минералы-яшмоиды [1].

Более широкое применение минералы нашли позже во времена медного (самородная медь), бронзового (касситерит) и железного (гематит, магнетит, лимонит) веков. Из этих минералов добывали различные металлы, которые использовались для изготовления орудий труда, посуды, благоустройства жилья. Они нашли применения и в искусстве, и в построении валютной системы, тем самым обусловив переход от натурального хозяйства к товарно-денежным отношениям (медные, серебряные, золотые монеты). В значительной степени именно драгоценные минералы определили эпоху «великих географических открытий», а поиск «философского камня» и вовсе дал начало дисциплине алхимии, которая со временем переродилась в настоящую науку о химических веществах. Таким образом происходил процесс постепенного накопления знаний. Однако, хранить, передавать и анализировать такие данные было довольно затруднительно ввиду несовершенства технологий. Написанное на бумаге утрачивалось во время стихийных бедствий, от которых на тот момент не было никакой защиты, а войны и грабежи с поджогами библиотек, зданий и даже целых городов были нередкими явлениями. Иными словами, бумажные варианты книг и папирусов нельзя было рассматривать как надёжный источник накопления и передачи знаний.

Первые носители информации, решавшие проблемы устойчивости к природным и антропогенным катастрофам, появились в 1804 г. Это были металлические перфокарты, но поскольку одна такая карта могла вместить в себя не более 80 символов и была при этом достаточно тяжёлой, вскоре её сменила перфолента, способная передавать в сотни раз больше информации. Однако и здесь были свои недостатки: перфоленты тоже были тяжёлыми, что ограничивало их перемещение на большие расстояния и не позволяло увеличивать длину, записывая больше информации.

Следующий скачок в технологиях произошел через 100 лет после создания перфолент. В 1952 г. была создана гибкая магнитная лента, которая долгое время применялась в катушках и компакт-кассетах. После появились магнитные диски, за ними дискеты, жёсткие диски, компактные диски (CD), USB флеш-накопитель и SSD. Весь путь от перфокарт до современных накопителей информации занял 215 лет, но быстрая скорость передачи информации на большие расстояния появилась только 30 лет назад и была приурочена к введению первой

свободной и многопользовательской сети, к которой было подключено около 7 500 пользователей.

Сейчас же доступ в интернет, портативные накопительные устройства и электронные вычислительные машины многократно увеличили как эффективность исследований в области минералогии, так и общую базу знаний.

Благодаря цифровизации и доступности информации, практически любой участник сети интернет может воспользоваться скомпилированными и агрегированными данными других участников, а также выкладывать в общий доступ свои данные.

В настоящей статье сделан анализ минералогических данных, предоставленных проектом RRUFF и Международной Минералогической Ассоциацией (ММА/IMA) в 2014 г. [2].

В данном датасете имеется 5 671 запись и 14 характеристик:

- Mineral Name = название минерала (Гранат, вивианит, электрум и т. д. Серебро, золото и другие самородные элементы – также названия минералов);
- RRUFF Chemistry (plain) = RRUFF формула (Стойберит = $Cu_2+5O_2(V_5+O_4)_2$ – вариация формулы $Cu^{2+}_5O_2(V^{5+}O_4)_2$);
- IMA Chemistry (plain) = формула Международной Минералогической Ассоциации (Стойберит = $Cu_5O_2(VO_4)_2$ – вариация формулы $Cu_5O_2(VO_4)_2$);
- Chemistry Elements = химические элементы в составе минерала (Стойберит = Cu, V, O, Серебро = Ag);
- IMA Number = уникальный IMA номер (если существует) (Стойберит = IMA1979-016);
- RRUFF IDs = уникальный RRUFF идентификатор (если существует, может быть несколько) (Серебро = R070416, R070463, R070754);
- Country of Type Locality = страна (включая значения 'unknown' и '?');
- Year First Published = первое задокументированное описание минерала;
- IMA Status = официальный статус IMA (Approved, Grandfathered, Pending publication);
- Structural Groupname = структурная группа минерала (Платарсит = Пирит);
- Fleischers Groupname = структурная группа минерала по Fleischers Glossary (2008);
- Status Notes = публикация, где впервые был описан минерал;
- Crystal Systems = вид сингонии (моноклинная, кубическая, гексагональная и т. д.);
- Oldest Known Age (Ma) = самый древний известный образец (в миллионах лет).

Его анализ выполнен с помощью высокоуровневого языка программирования Python (версии 3) и специальных библиотек для анализа и визуализации данных:

- Numpy – библиотека для работы с n-мерными массивами и оперирования высокоуровневыми математическими операциями.
- Pandas – надстройка над numpy, даёт удобные инструменты для работы с табличными структурами данных (DataFrame, Series).
- Matplotlib – низкоуровневая библиотека для визуализации двумерных и трёхмерных графиков.
- Seaborn – высокоуровневая библиотека, предоставляющая интерфейс для работы с графиками.

Для начала прочитаем данные и выведем первые 3 записи, чтобы посмотреть, как они выглядят (рис. 1).

Здесь видны не все столбцы, однако важно упомянуть, что среди отдельных значений будут встречаться NaN – это отсутствующие значения, которые мы не можем восстановить. В отдельных случаях, можно логически предположить почему эти данные пропущены, но в большинстве случаев это связано с техническими проблемами при создании датасета. В выведенных выше записях мы видим, что идентификатор RRUFF для минерала абелаит отсутствует, т. к. минерал был описан в 2014 г. и, вероятнее всего, ему ещё не успели присвоить какой-либо идентификатор. В данном примере мы можем логически предположить причину, по которой значение отсутствует, но заполнить его сами мы не можем.

Поскольку у всех отдельных минералов есть год первого его описания в различных научных трудах, то можно посмотреть, как менялось количество впервые описанных (открытых) минералов по годам (рис. 2). Эти данные помогут определить время зарождения самой науки минералогии и оценить темпы её развития.

Mineral Name	RRUFF Chemistry (plain)	IMA Chemistry (plain)	Chemistry Elements	IMA Number	RRUFF IDs	Country of Type Locality	Year First Published	IMA Status	Structural Groupname
0 Abellaite	NaPb2+2(CO3)2(OH)	NaPb2(CO3)2(OH)	Na Pb C O H	IMA2014-111	NaN	Spain	2014.0	Approved	
1 Abelsonite	Ni2+C31H32N4	NiC31H32N4	Ni C H N	IMA1975-013	R070007	USA	1975.0	Approved	
2 Abenakiite-(Ce)	Na26Ce3+6(SiO3)6(P5+O4)6(C4+O3)6(S4+O2)O	Na26Ce6(Si6O18)(PO4)6(CO3)6(SO2)O	Na Ce Si O P C S	IMA1991-054	NaN	Canada	1991.0	Approved	

Рисунок 1 – Первые 3 записи исследуемых данных

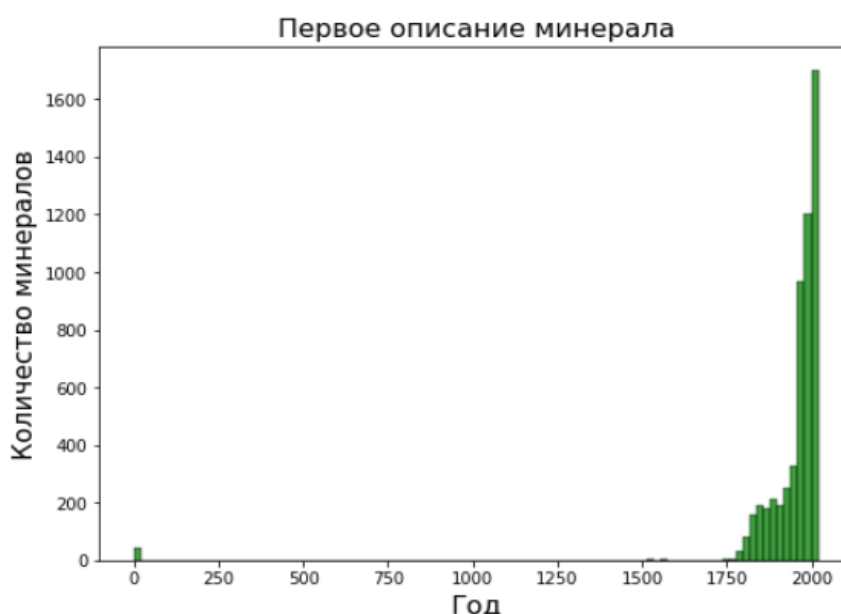


Рисунок 2 – Количество описанных минералов по годам

Видно, что активно находить и описывать новые образцы стали с 1750–1800 гг. и делали это, в основном, европейские страны, такие как Италия, Германия и Великобритания. Самые описанные минералы представлены преимущественно рудообразующими минералами (барит – Ba, рутил – Ti, крокоит – Pb и Cr, торбернит – U, а также графит и платина). Сопоставляя эти данные между собой, можно прийти к выводу, что интерес к таким минералам обусловила начавшаяся в 1760-х гг. европейская промышленная революция. Помимо прочего, есть некоторые образцы, датируемые 0 годом. Туда относятся 40 минералов, среди которых есть самородные минералы (золото, серебро) и те, которые использовались человеком ещё до нашей эры – малахит, топаз, киноварь, гипс и т. д. Скорее всего, учёным не удалось выделить самое раннее описание этих минералов, поэтому им было присвоено значение 0.

Поскольку в данных есть формула минерала и отдельно вынесены химические элементы, которые её составляют, то можно посмотреть на самые распространённые из них. Топ-5 самых распространённых элементов в порядке убывания: O, H, Si, Ca, Fe. Построим круговую диаграмму, чтобы увидеть абсолютную частоту появления химического элемента в минералах (рис. 3).

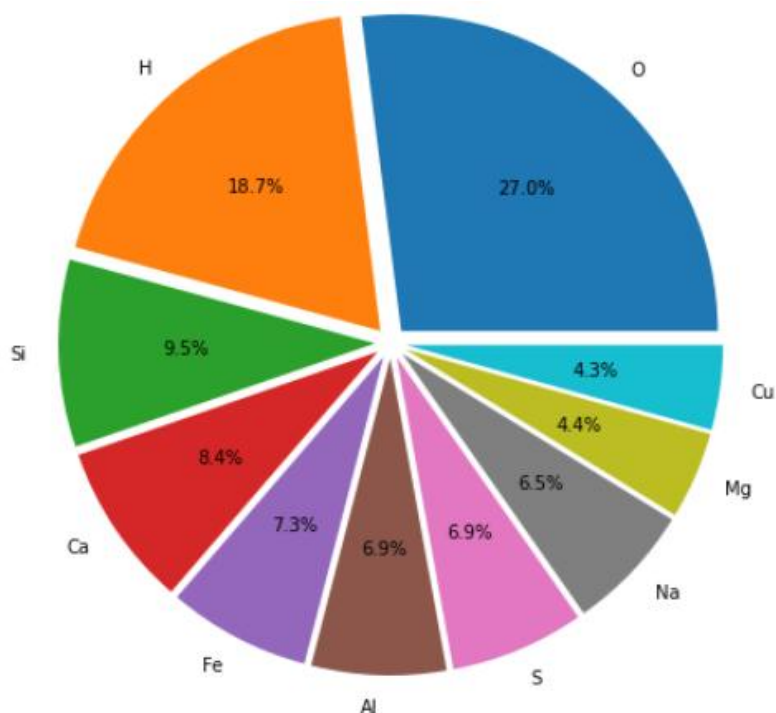


Рисунок 3 – Частота появления элемента в минералах

Стоит отметить, что эти данные не отображают массовую долю элементов в составе минералов, а лишь показывают частоту их появления в тех или иных образцах. В любом случае, такое распределение очень напоминает химический состав земной коры, и это вовсе не случайно. Земная кора преимущественно состоит из твёрдых минералов и горных пород. Это означает, что химический поэлементный состав земной коры будет схож с химическим составом «усреднённого» таким образом минерала.

Продолжая анализировать имеющиеся данные, можно посмотреть какие страны приняли самое активное участие в поиске и описании минералов (рис. 4).

Из графика видно, что США и Россия явно лидируют. На каждую из них приходится приблизительно по 800 открытых минералов. Затем с большим отрывом следуют Италия и Германия – по 350 описанных минералов. Сложно предположить причины, которые привели именно к такому распределению, скорее всего так просто сложилось исторически.

Ещё один интересный подсчёт – самые распространённые сингонии среди минералов. В характеристике по симметриям в данных указано 8 значений – 6 основных видов сингоний, аморфная структура и минералы с неизвестной симметрией (рис. 5).

Из шести основных сингоний видно, что наибольшее распространение получила моноклинная, относящаяся к низшей категории. Сюда входят 1 896 образцов. Самая редкая – аморфная структура, к которой приурочено всего лишь 17 образцов (эвансит, метастибнит, заратит и проч.). Также, среди минералов с ещё не установленной сингонией есть гекторит, вудвардит (рис. 6), ксантоксенит, ферримолибдит.

Помимо прочего, в датасете есть данные по возрасту самого древнего найденного образца по каждому из минералов (рис. 7).

Самый большой возраст – 4,7 млрд лет, имеют 37 минералов, среди которых графит, алмаз, хромит, диопсид, пентландит и некоторые другие. К самым молодым, открытым относительно недавно, относятся 124 минерала, и большая часть из них синтезированы (везерилит, виджимулсалит, ацетамид). Средний возраст всех образцов составляет 1,25 млрд лет, медианный возраст – 600 млн лет.

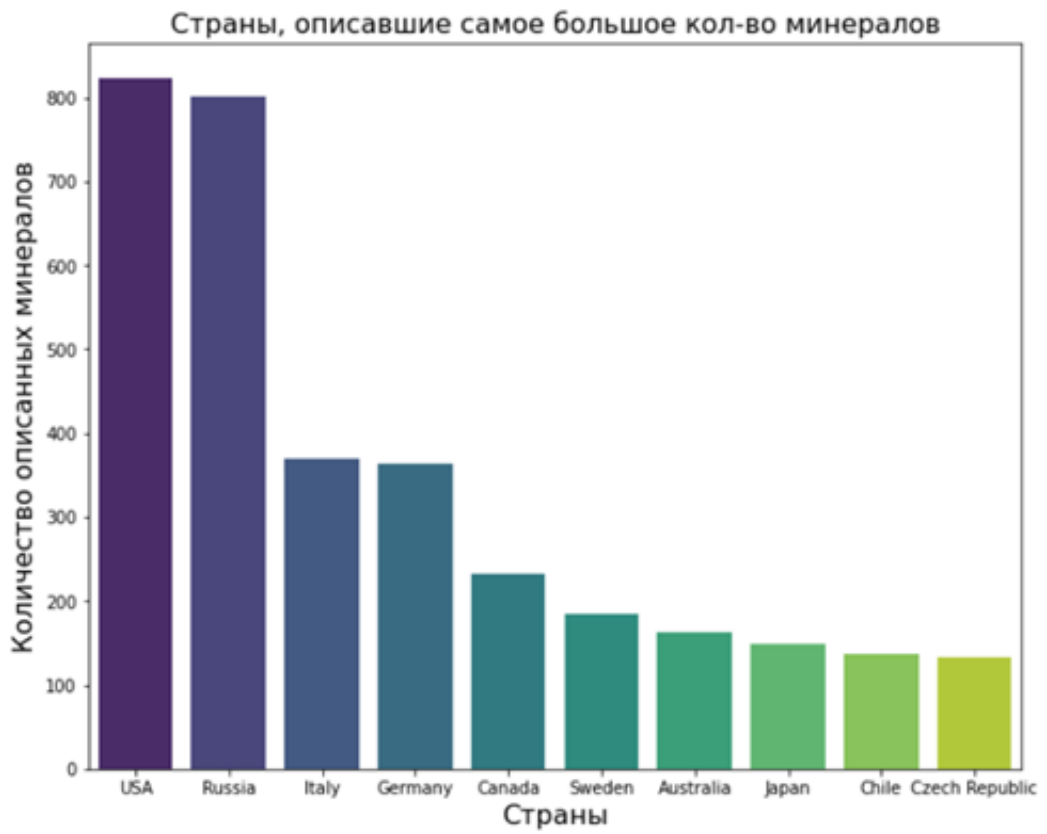


Рисунок 4 – Топ-10 стран, описавших самое большое количество минералов

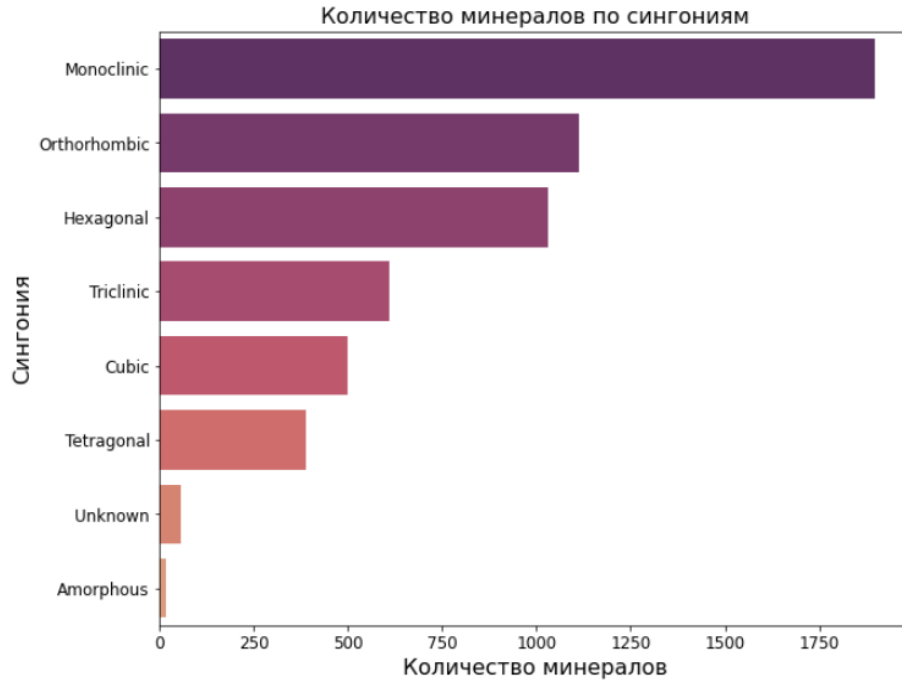


Рисунок 5 – Количество минералов по сингониям

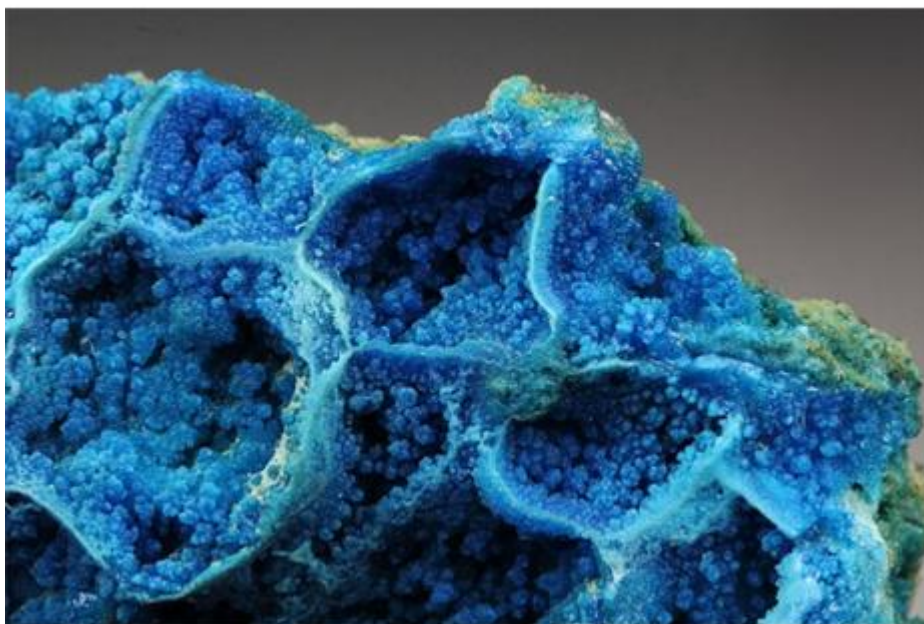


Рисунок 6 – Минерал вудвардит

Количество минералов по абсолютному возрасту старейшего найденного образца

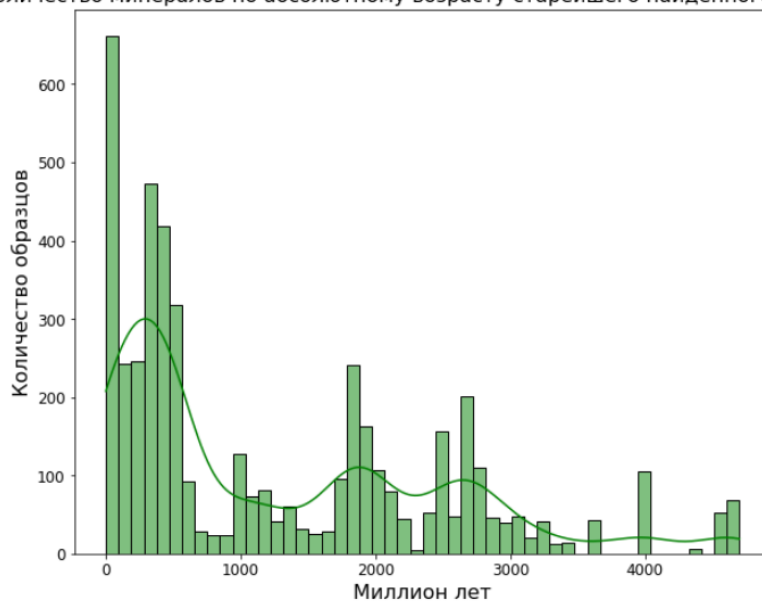


Рисунок 7 – Абсолютный возраст самых древних образцов минералов

Таким образом, цифровые технологии предоставляют широкий набор инструментов для анализа как числовых (год открытия, возраст минерала), так и категориальных (страна, сингония) данных. Уже обычный статистический анализ уже позволяет рассказать многое об истории минералогии, но этого обычно недостаточно. Чаще всего гораздо важнее предсказывать конкретные значения уже имеющихся тенденций. Например, можно проследить как менялось количество открытых минералов по годам (рис. 8).

На данной диаграмме наблюдается чёткий экспоненциальный рост количества открытых и синтезированных по годам минералов. Мы можем аппроксимировать эти данные с помощью полиномиальной кривой, решив задачу нелинейной регрессии, и предсказать значения на 2015, 2016, 2017 и все последующие года.



Рисунок 8 – Диаграмма рассеяния количества открытых минералов по годам



Рисунок 9 – Диаграмма рассеяния количества открытых минералов по годам (рис. 8) с аппроксимирующей линией

Видно, что линия (рис. 9) хорошо усредняет исторические данные, и поэтому можно опираться на её прогнозы относительно будущего. Так, количество открытых минералов 2015 г. по прогнозам модели будет в районе 95, что оказалось недалеко от истины.

Статистические методы цифрового анализа данных в минералогии находят своё применение в описании рудных месторождений, когда надо оценить имеющиеся разведанные объёма месторождения, процент полезного элемента/соединения в метре кубическом руды, количество и качество примесей. Цифровой анализ также используют в лабораториях для изу-

чения образцов, особенно при наличии некоторой статистически значимой выборки, когда нужно проверить какую-то гипотезу относительно геологического объекта. Цифровая минералогия отвечает за создание и поддержание различных баз данных, с метаданными о тех или иных объектах. По мере развития технологий такой способ изучения дисциплины становится доступнее для более широкого круга лиц.

Предиктивный цифровой анализ, как затронутый в данной статье метод полиномиальной нелинейной регрессии, является более сложным инструментом, также известным под термином ML (machine learning/машинное обучение). Сюда относится целый ряд предсказательных задач, включающий в себя задачи регрессии и классификации. Регрессия применяется для предсказания континуальных (непрерывных) числовых переменных, в то время как классификация нужна исключительно для предсказания каких-либо классов (бинарная классификация – 2 класса, многоклассовая классификация – 3 и более классов). Регрессия применяется для предсказания времени выработки месторождения, если нам известны запасы и динамика объёмов выработки по временным периодам, для предсказания массовых долей элементов в минерале при определённых термобарических условиях и т. д. Классификация – это метод машинного обучения для предсказания минерала по набору каких-то признаков (твёрдость, прозрачность, цвет и др.), для определения того, выгодно ли месторождения для дальнейшей выработки (бинарная классификация да/нет).

Библиографические ссылки

1. *Белявский Н. А.* Стрoение земной коры континентов по геолого-геофизическим данным. М.: Недра, 1981.
2. IMA Database of Mineral Properties [Электрон. ресурс]. URL: <https://rruff.info/ima/> (дата обращения: 1.09.2021).

УДК 549

НЕЙРОСЕТЬ ДЛЯ ОПРЕДЕЛЕНИЯ МИНЕРАЛОВ

А. С. Василюк

Белорусский государственный университет, факультет географии и геоинформатики,
пр. Независимости 4, 220030 Минск, Республика Беларусь; andrew.vasiluck@gmail.com

Рассмотрены различные аспекты создания нейросети для определения минералов.

Ключевые слова: минерал; нейросеть.

Various aspects of creating a neural network for determining minerals are considered.

Key words: mineral; neural network.

Существуют продвинутые методы цифровой минералогии, решающие сложные задачи определения минерала по его фотографии на основе нейросетевого подхода.

Нейросети – достаточно распространённые алгоритмы для решения ряда сложных задач. Ввиду большого разнообразия программных архитектур и вычислительных компонент они являются очень гибким инструментом. Предсказание временных рядов, классификация изображений, детекция предметов на видео, генерация картинок, распознавание речи – всё это возможно благодаря нейросетям.

В самом общем случае нейросеть представляет собой некоторую математическую модель, построенную на взаимодействии программного представления биологических нейронов (рис. 1) [1].