ГЕОФИЗИЧЕСКИЕ МЕТОДЫ ИЗУЧЕНИЯ НЕДР

УДК 550.8.052 DOI:10.7242/echo.2022.3.5

ВОССТАНОВЛЕНИЕ ДАННЫХ МАГНИТНЫХ ВАРИАЦИЙ ГИБРИДНЫМ АЛГОРИТМОМ МАШИННОГО ОБУЧЕНИЯ

П.Н. Новикова
Горный институт УрО РАН, г. Пермь

Аннотация: В статье рассматривается применение методов машинного обучения для восстановления и прогнозирования данных магнитных вариаций на примере наблюдений обсерватории в г. Магадан. За основу берется классическое разложение временных рядов, которыми являются магнитные вариации, на 4 составляющие: тренд, сезонная, циклическая и остаточная компоненты. Каждая компонента отдельно моделируется при помощи линейных и нелинейных алгоритмов регрессионных деревьев решений. На этой основе строится гибридный алгоритм, способный восстанавливать пропущенные данные, а также прогнозировать магнитные вариации на заданный промежуток времени. Показано применение автоматического алгоритма, убирающего выбросы данных. Приведен пример восстановления отсутствующих данных на отрезке времени пять суток.

Ключевые слова: магнитные вариации, временные ряды, методы машинного обучения, прогнозирование, восстановление данных, дерево решений.

Наблюдение за изменчивостью магнитного поля является важной задачей, которая находит применение для решения широкого круга научных и производственных задач. Значительным моментом является соблюдение непрерывности регистрации вариаций магнитного поля Земли, т.к. пропуски в данных могут привести к невосполнимой потере информации о геофизических событиях или техногенных катастрофах. Особенно это актуально при регистрации магнитных вариаций наземными обсерваториями и спутниками [2].

Существенный объем данных, накопленный с 60-х годов прошлого столетия, позволяет всесторонне изучать поведение магнитных вариаций в разных регионах, автоматически регистрировать аномальные события, такие как магнитные бури, анализировать периодические компоненты вариаций. Предлагается использование методов машинного обучения для прогнозирования временных рядов, которыми являются магнитные вариации.

Исходные данные

Геомагнитные вариации были взяты из открытого ресурса сайта Мирового центра данных по солнечно-земной физике, г. Москва[1]. Для исследований выбраны данные годовых абсолютных измерений среднечасовых значений полного вектора магнитной индукции F, зарегистрированные протонными магнитометрами в геомагнитной обсерватории г. Магадан за 2014 г.

Гибридная модель

Для восстановления отсутствующих значений и прогнозирования будущих магнитных вариаций была экспериментально подобрана гибридная модель обучения «с учителем», основанная на классическом разложении данных временных рядов на отдельные компоненты: трендовую, сезонную, циклическую и остаточную. Такая модель подтверждена многолетними наблюдениями за переменным магнитным

[1] Материалы сайта http://www.wdcb.ru/ Мирового центра данных по солнечно-земной физике, г. Москва

полем Земли, обусловленным воздействием солнечной активности, процессами ионосферы, различных космических излучений и др. [2, 3]. Магнитные вариации состоят из периодических и непериодических, спокойных и возмущенных компонент. К спокойным периодическим компонентам относятся солнечно-суточные вариации (наиболее выраженная компонента), 27-дневные, сезонные и годовые вариации. Возмущенные вариации могут быть как короткопериодными (с периодом менее суток), непериодическими, так и резкими апериодическими изменениями – магнитными бурями [2, 3].

Под трендовой компонентой понимается основная тенденция изменения данных в течение длительного времени. Сезонная компонента магнитных вариаций включает несколько периодичных изменений [2, 3]: значимыми будут суточная, месячная, сезонная компоненты. Под циклической составляющей будем понимать несистематические, но повторяющиеся изменения данных, не имеющих четкой периодичности. Остаточная часть будет описывать некоторую случайную компоненту данных.

В гибридном подходе каждая составляющая магнитных вариаций моделируется отдельно, с последовательным вычитанием каждой компоненты из исходных данных. Для обучения использовались как весь набор данных, только прошлые значения или прошлые и будущие значения полного вектора магнитной индукции. Для моделирования временных рядов были выбраны различные модификации прогнозных регрессионных деревьев решений.

Модель машинного обучения

Деревья решений – это способ представления правил в иерархической, последовательной структуре, где каждому объекту соответствует единственный узел, дающий решение [5]. Процесс построения дерева начинается с определения корневого (начального) узла и выбирается на основе всего набора данных. Каждая точка ветвления в дереве называется узлом и представляет набор данных, который содержит некоторые или все записи исходного набора данных. Конечные неразделенные узлы называются листьями. Структура дерева представляет собой «листья» и «ветки». На рёбрах («ветках») дерева решения записаны признаки, от которых зависит целевая функция, в «листьях» записаны значения целевой функции, а в остальных узлах – признаки, по которым различаются случаи. Чтобы классифицировать новый случай, надо спуститься по дереву до листа и выдать соответствующее значение. Дерево решений, в котором у каждого узла может быть только два ответвления, называется бинарным.

Регрессионное дерево решений – это автоматический метод машинного обучения, который аппроксимирует входную функцию, представленную количественной переменной, вычисляя предсказанное среднее значение для каждого узла в дереве.

Разберем работу простого алгоритма регрессионных деревьев на примере анализа времени подъема в зависимости от возраста человека[2] (рис. 1). Корень дерева оценивается по всему набору данных и представляет среднее значение всех наблюдений (рис. 1, А). Далее будем классифицировать данные по возрасту (рис. 1, Б): серая линия делит датасет на две группы, в каждой из которых вычисляется свое среднее значение. Следующий уровень бинарного дерева повторно разбивает предыдущие две группы (рис. 1, В). Данная процедура выполняется рекурсивно для каждого полученного подмножества до тех пор, пока не будут достигнуты критерии остановки: не перестанет улучшаться сумма квадратов или число строк, соответствующее этому узлу, не

[2] Пример взят с сайта https://end-to-end-machine-learning.teachable.com/

станет слишком маленьким. Процесс также будет остановлен, если число узлов в дереве решений станет слишком большим. В данном примере структура дерева решений позволяет сортировать людей разных возрастов на соответствующие им ячейки и делать оценки времени пробуждения.

Реализация алгоритма моделирования магнитных вариаций проводилась в Jupyter notebook на языке программирования Python с использованием основных библиотек pandas, numpy и scikit-learn [1, 4].

В данной статье показано применение гибридного машинного обучения для восстановления данных внутри набора годовых вариаций, отсутствующих по различным техническим причинам.

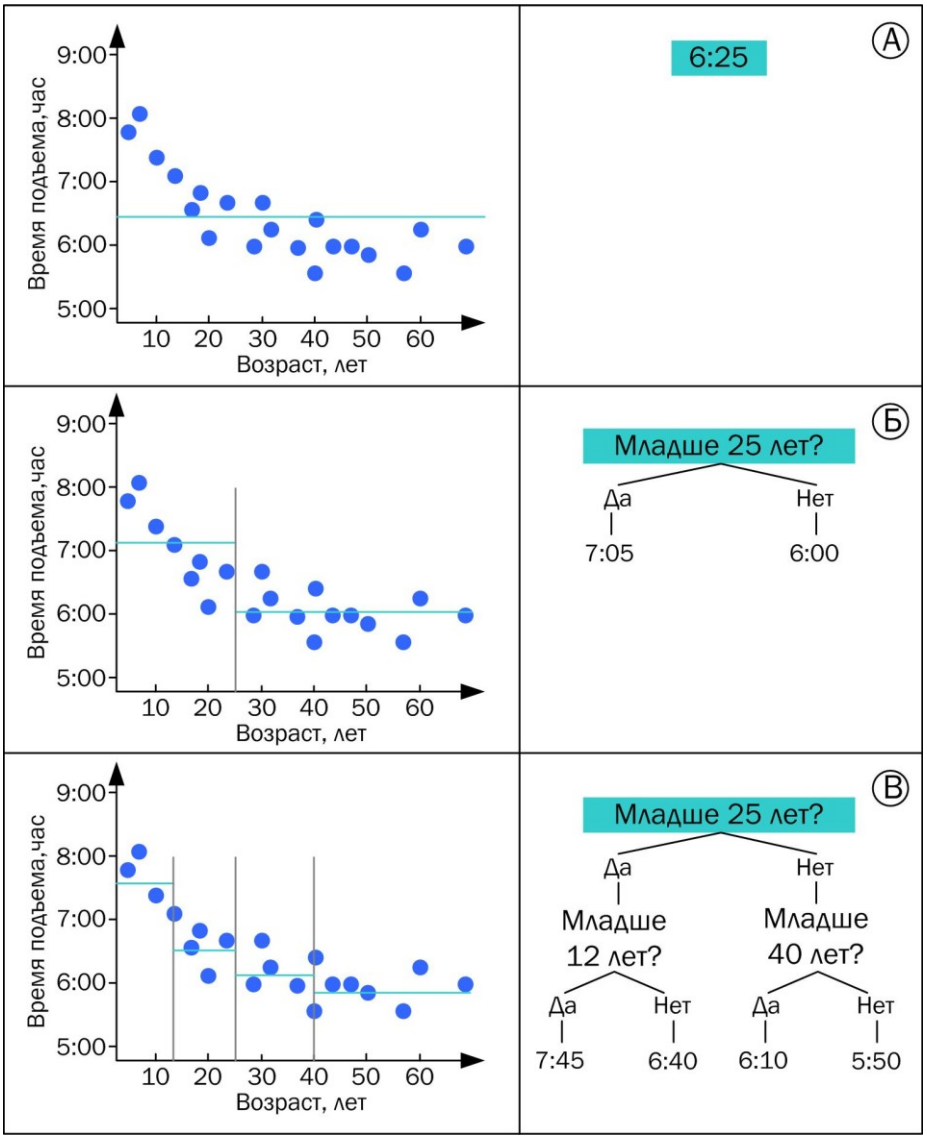

Рис. 1. Пример работы алгоритма регрессионного дерева решений

Отсутствующие значения

Представленный набор исходных данных содержит отсутствующие значения, которые необходимо заменить на значимые, т.к. большинство методов машинного обучения чувствительны к пропускам данных.

Для замены отсутствующих значений были применены две стратегии. Если отсутствующих значений в данных наблюдается не более трех подряд, то для замены данных

использовался метод линейной интерполяции. В противном случае пропуски данных заполнялись постоянным значением – средним значением ближайшего суточного интервала данных (рис. 2).

Выбросы

Выбросом может считаться небольшая часть данных, отличающаяся от основных данных по выбранным критериям. Часто в качестве таких критериев выступают низкая плотность распределения области данных и недостаточное количество соседних данных, находящихся на близком расстоянии [6].

Строго говоря, выбросов в данных вариаций в том понимании, в котором принято говорить об ошибочных данных, практически нет. Однако, есть небольшая часть данных, несмотря на хорошую согласованность с поведением всей кривой наблюдения значительно превышающих доверительный интервал значений по амплитуде (рис. 2, 1). Учитывая близкий к линейному тренд магнитных вариаций, доверительный интервал в исходных данных является динамическим. Таким образом, стоит задача «выравнивания» данных по амплитуде так, чтобы не изменить структуру исходных данных. При этом статистические методы, использующие в качестве показателей фиксированный порог удаления значений (например, через квартили данных), работают недостаточно хорошо, «обрезая» часть полезного сигнала.

Для быстрого удаления выбросов был применен метод машинного обучения без учителя библиотеки scikit-learn, основанный на принципе Монте-Карло – изоляционное дерево (Isolation Forest Tree) [6, 7]. Данный алгоритм при помощи ансамбля бинарных деревьев решений сегментирует пространство признаков случайным образом, при этом отсекая изолированные точки от нормальных кластеризованных данных. Нормальность определяется через среднее арифметическое глубин листьев, в которые попадает то или иное значение. Результат итерационно усредняется по нескольким запускам стохастического алгоритма. Алгоритм распознает аномалии различных видов: как изолированной точки с низкой локальной плотностью, так и кластеры аномалий малых размеров.

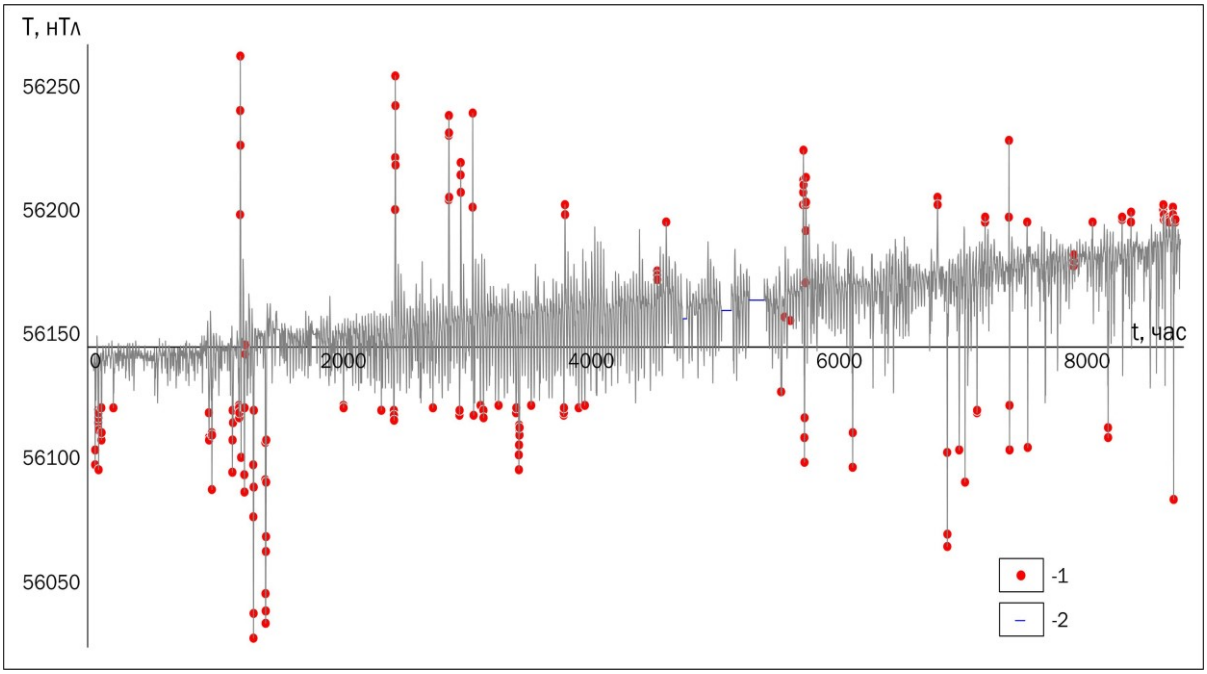

Рис. 2. Средние часовые магнитные вариации обсерватории г. Магадан за 2014 г.:
1 – выбросы, определенные методом Isolation Forest Tree, 2 – замененные отсутствующие значения

В исследуемом наборе данных достаточно исключать не более 2-5% выбросов (рис. 2), что необходимо для устойчивого определения дальнейших компонент разложения.

Тренд

Долгосрочная компонента годовых магнитных вариаций определялась через полиномиальную регрессию. Четко прослеживается тенденция к увеличению уровня вариаций. Экспериментальным путем наиболее подходящим оказался тренд 2 порядка (рис. 3).

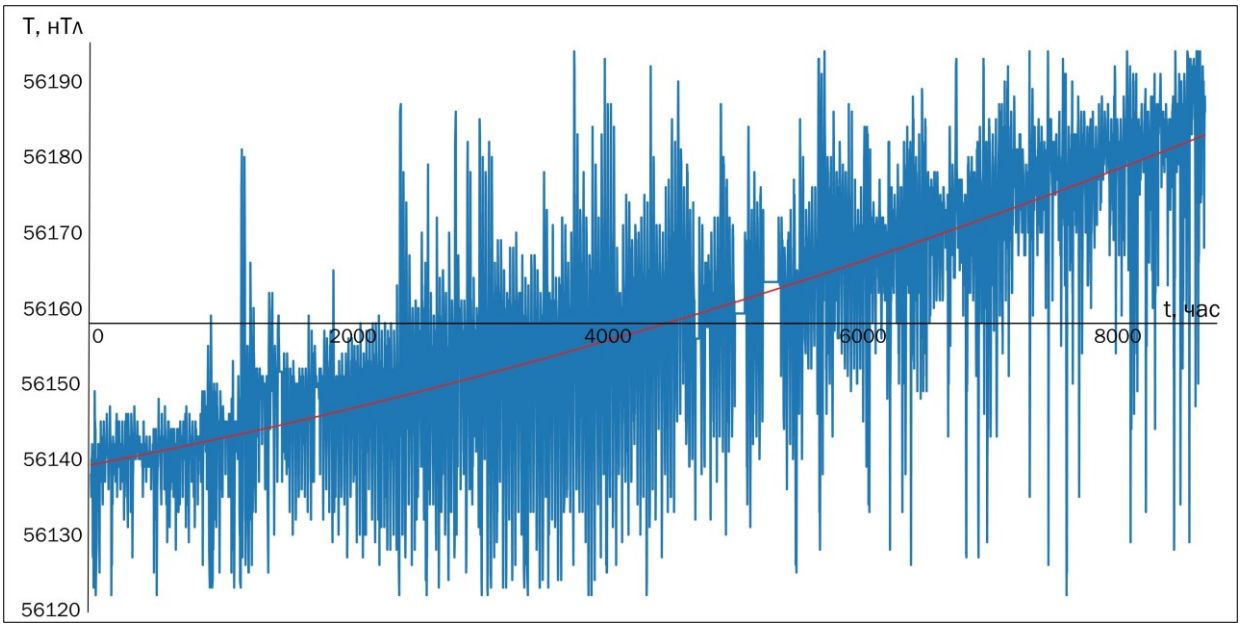

Рис. 3. «Очищенные» данные магнитных вариаций (синяя линия) и вычисленный тренд (красная линия)

Определение тренда и удаление его из исходных данных является необходимым шагом для дальнейшего моделирования данных. Тренд используется в качестве базовой линии для более сложных моделей, неспособных изучать тенденции.

Сезонность

Для изучения сезонности данных строилась периодограмма Фурье с оценкой спектральной мощности по фиксированным периодам. По периодограмме наиболее ярко прослеживается суточная компонента, но также более размыто прослеживаются полугодовые, квартальные, месячные и т.д. компоненты, а также более мелкие суточные периоды.

Для модели сезонности будем считать, что компоненты линейны (также экспериментально установлено, что линейная модель наиболее удачно с физической точки зрения описывает данные).

Сезонность определялась через модель линейной регрессии LinearRegression библиотеки scikit-learn, которая позволяет на основе представленной выборки данных подбирать периодические компоненты, заданные по периодограмме, учитывая тренды 1, 2 и 3 порядков в данных. В качестве обучающих признаков модели были использованы периодические кривые функций Фурье (пары синусоидальных и косинусоидальных кривых) с частотами фиксированного сезона. Также как и для тренда, в качестве базисной линии для конструирования признаков

было выбрано полиномиальное изменение сезонной компоненты. Алгоритм линейной регрессии вычисляет веса, которые будут соответствовать сезонному компоненту в целевом ряду.

Построенная модель сезонности в основном отражает суточные колебания магнитных вариаций с несколькими характерными точками (рис. 4, Б). Амплитуда колебаний составляет от -12.5 нТл до 11.5 нТл. При этом стоит заметить, что средний уровень колебаний изменяется и отражает более длительные периодические компоненты, так на рисунке 4, А при усреднении подобранной сезонности с недельным периодом, хорошо проявляется месячная компонента.

Цикличность данных

После редуцирования трендовой и сезонной компоненты при повторном построении периодограммы понятно, что мы не смогли полностью подавить периодические компоненты. Поэтому будем считать циклической компонентой составляющую с не только непостоянным периодом, но также с нелинейным характером.

Цикличность определялась при помощи метода XGBRegressor, который представляет наиболее популярную версию алгоритма градиентного бустинга деревьев решений. Градиентный бустинг – это техника машинного обучения для задач классификации и регрессии, которая строит модель предсказания в форме ансамбля слабых предсказывающих моделей, обычно деревьев решений. Обучение ансамбля проводится последовательно, на каждой итерации вычисляются отклонения предсказаний уже обученного ансамбля на обучающей выборке. Следующая модель, которая будет добавлена в ансамбль, будет предсказывать эти отклонения. Таким образом, добавив предсказания нового дерева к предсказаниям обученного ансамбля, мы можем уменьшить среднее отклонение модели, которое является таргетом оптимизационной задачи [1, 4].

В качестве признаков обучения для модели градиентного бустинга были использованы серии запаздываний данных, основанных на автокорреляционной зависимости данных. Собственно, как и на периодограмме, просматривается периодическая корреляция данных: для часов наиболее значимыми являются ближайшие три часа, для дней – 2 ближайших дня, для недель – 2 ближайшие недели. В зависимости от величины отрезка пропусков целесообразно брать разные периоды: для пропусков, близких суточному периоду, можно брать запаздывания по часам и суткам; для пропусков более суток необходимо дополнительно рассматривать недельные запаздывания. Также в качестве признаков использовались характеристики по размаху исходных данных. Размах показывает сильную полиномиальную зависимость 2 порядка с пиковыми значениями в летнее время. Остаточная компонента размаха после вычета тренда крайне изменчива, ее собственный размах достигает 60 нТл. В качестве признаков были добавлены собственно размах, минимальное и максимальное значения за день, трендовая компонента размаха, остаточная компонента размаха.

На рисунке 4, Б можно увидеть, что циклическая компонента корректирует суточные вариации как по форме, так и по амплитуде, отражая, в основном короткопериодные колебания вариаций.

Остаточные аномалии

Остаточная компонента изменяется в довольно широких пределах в несколько нТл по амплитуде, несмотря на медианное и среднее значения, близкие к нулю. Случайная компонента моделировалась также градиентным бустингом, основанным на использовании ансамбля деревьев решений.

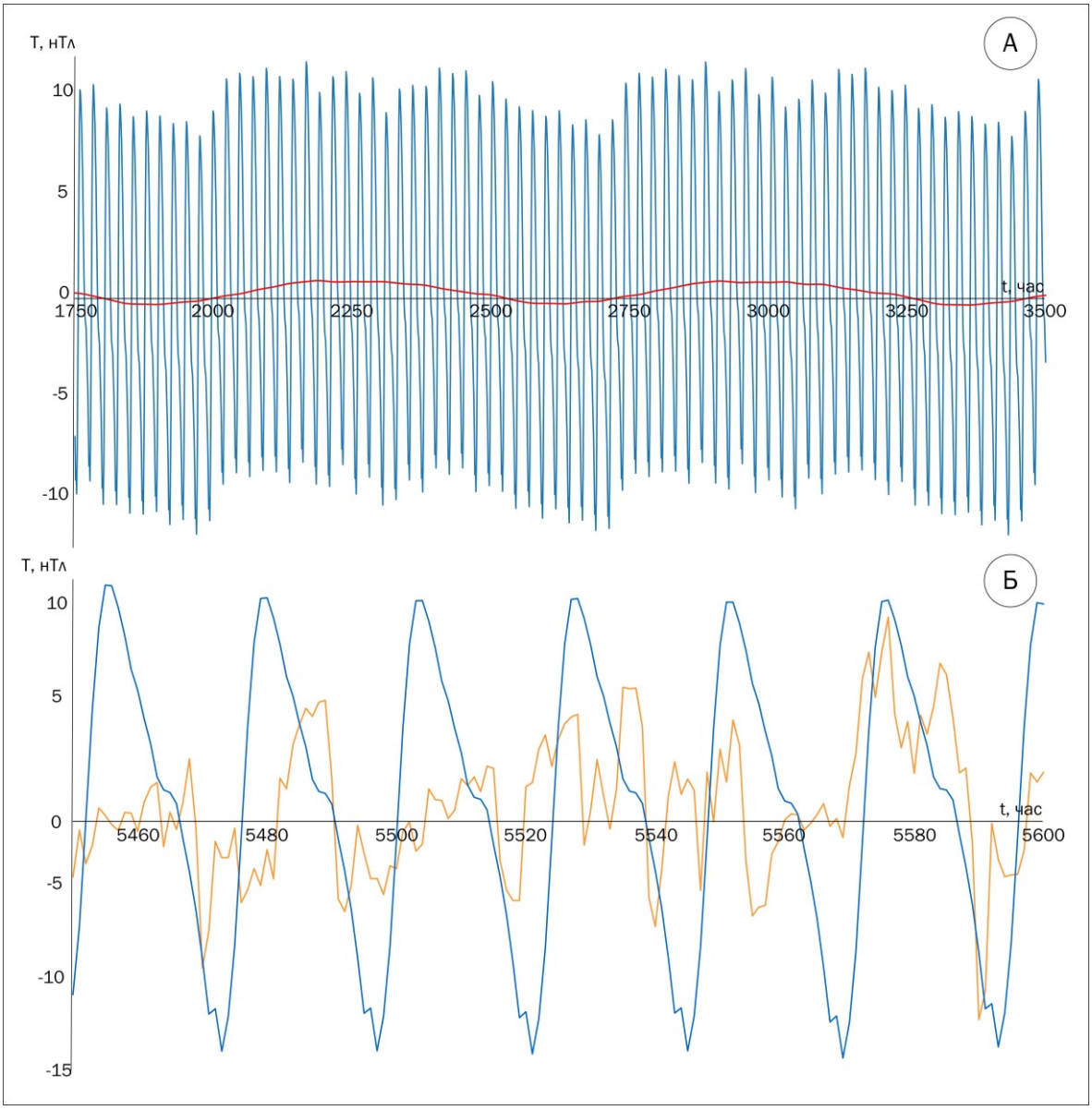

Рис. 4. Моделирование сезонной и циклической компонент годовых магнитных вариаций: синяя линия – сезонная компонента, красная линия – среднее значение сезонной компоненты по недельному периоду, оранжевая линия – циклическая компонента

Восстановление отсутствующих значений

В представленном наборе данных присутствует несколько существенных пропусков, в основном в пределах суток. Но есть и более длительные отрезки без наблюдений, на одном из них показана работа гибридного алгоритма (рис. 5).

Был выбран отрезок времени с отсутствующими значениями на протяжении 5 суток, наблюдаемый в августе. Предположим, что наблюдения отсутствуют исключительно по техническим причинам, не связанными с магнитными бурями. Тогда мы можем оценить среднестатистическое поведение магнитных вариаций за этот период, используя предложенный гибридный алгоритм.

Трендовая компонента была оценена по всему набору данных, что оправдано внутренним расположением отсутствующих данных. Сезонная компонента, как более устойчивая, прогнозировалась только по прошлым данным. Циклическая изменчивая компонента рассчитывалась как по прошлым данным, так и по будущим. Функции запаздывания выбирались с длительностью в сутки и неделю.

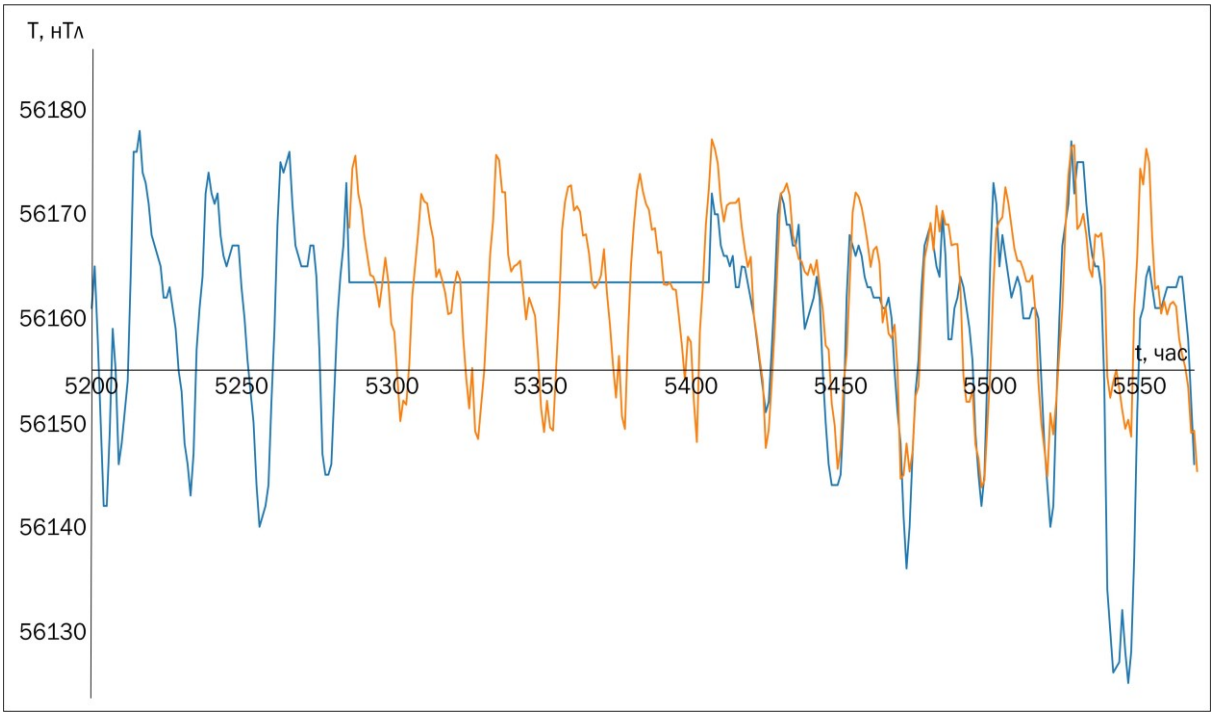

Рис. 5. Восстановление отсутствующих значений внутри годового набора данных магнитных вариаций гибридным методом машинного обучения: синяя линия – исходные вариации, оранжевая линия – смоделированные вариации

На рисунке 5 показаны результаты моделирования отсутствующих значений. Для оценки точности моделирования были также рассчитаны прогнозные значения на неделю вперед. Среднеквадратическая погрешность составила 8 нТл, что составляет примерно 10% от максимальной амплитуды данных.

Выводы

Моделирование магнитных вариаций, состоящих из многих линейных и нелинейных компонент и являющихся существенно изменчивыми, особенно в высоких широтах, является нетривиальной задачей, которую изучают с применением детерминистических (например, спектральный анализ) и статистических методов [2]. Применение методов машинного обучения позволит быстро анализировать большой объем данных (например, минутные вариации), прогнозировать и восстанавливать магнитные вариации, быстро определять аномальные значения, характерные для магнитных бурь.

Представленный пример демонстрирует возможности в изучении как полностью всего набора данных годовых вариаций, так и их компонент, а также прогнозировании отсутствующих значений данных. Показан автоматический метод поиска аномальных значений Isolation Tree. В перспективе дальнейших исследований необходимо изучить и дополнить признаки, влияющие на изменчивость магнитных вариаций, для лучшего прогнозирования данных.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Бринк Х., Ричардс Д., Феверолф М. Машинное обучение. – Спб.: Питер, 2017. – 336 с.
2. Гвишиани А.Д., Лукьянова Р.Ю., Соловьёв А.А. Геомагнетизм: от ядра Земли до Солнца. – М.: РАН, 2019. – 185 с.: ил.
3. Магниторазведка: справ. геофизика / под ред. В.Е. Никитского, Ю.С. Глебовского. – 2-е изд., перераб. и доп. – М.: Недра, 1990. – 469 с.

4. Python и машинное обучение: машинное и глубокое обучение с использованием Python, scikit-learn и TensorFlow 2: пер. с англ. – 3-е изд. – СПб.: ООО «Диалектика», 2020. – 848 с.
5. Breiman L., Friedman J.H., Olshen R.A., Stone C.J. Classification and regression trees. – Monterey, 1984. –
6. Hawkins D.M. Identification of Outliers. – Berlin: Springer, 1980. – 198 p.
7. Liu F.T., Ting K.M., Zhou Z-H. Isolation forest // Proceedings – IEEE International conference on Data Mining, ICDM. – 2008. – P. 413-422. – DOI: 10.1109/ICDM.2008.17.

УДК 550.(832.4 + 834.05) DOI:10.7242/echo.2022.3.6

ИНТЕГРАЦИЯ ДАННЫХ ГИС И 3D-СЕЙСМОРАЗВЕДКИ ДЛЯ ПРОГНОЗА ФЛЮИДОНАСЫЩЕНИЯ ПОРОДНОГО МАССИВА

А.Д. Тезиков, А.Б. Трапезникова
Горный институт УрО РАН, г. Пермь

Аннотация: В настоящее время в задачи сейсморазведки ставится не только прогноз наличия потенциальных пород коллекторов, но и прогноз флюидонасыщения. Для реализации этого необходимы качественные зависимости между петрофизическими параметрами и сейсмическими атрибутами. Формирование таких зависимостей основано на петроупругом моделировании скважин. Атрибутный анализ, основанный на связях с петрофизическими параметрами, повышает достоверность прогноза расположения и насыщения потенциальных резервуаров по сейсмическим данным.

В данной статье представлены результаты применения динамического анализа по атрибутам «средняя частота» и «энергия» на примере одной из нефтеперспективных структур, выделенной по данным кинематической интерпретации 3D сейсморазведки в бассейне Таранаки, Новая Зеландия. По итогам исследования выполнен прогноз потенциальных зон флюидонасыщения пригодных для разработки.

Ключевые слова: атрибутный анализ, сейсморазведка, сейсмическая интерпретация, геофизические исследования скважин, петроупругое моделирование.

Введение

На сегодняшний день геофизические методы поиска и разведки нефтегазовых месторождений, такие как сейсморазведка, не теряют своей актуальности и находят применение по всему миру. Сейсморазведка является основным и наиболее точным геофизическим методом поиска месторождений нефти и газа, предоставляющим наиболее полный спектр информации об изучаемой геологической среде. С ее помощью можно как получить структурные изображения геологических границ, так и исследовать литологические, петрофизические и физико-емкостные свойства горных пород.

В данной работе представлен результат работы авторов в рамках образовательного проекта SEG «EVOLVE 2021». «EVOLVE» – ежегодно проводимый обществом SEG проект, объединяющий геофизиков со всего мира для совместного решения актуальных геологических, геофизических и инженерных задач.

Авторы статьи с коллективом других исследователей в рамках проекта «EVOLVE» выполняли исследования по поиску новых перспективных нефтегазоносных структур в бассейне Таранаки (Новая Зеландия) на основе данных сейсморазведки, геофизических исследований скважин и широкого спектра геологической информации.

Объектом исследования являлась турбидитовая формация Тангароа на нефтегазовом месторождении Кора в бассейне Таранаки. Таранакский бассейн представляет собой осадочный бассейн мелового и третичного периода, расположенный вдоль западной границы Северного острова в Новой Зеландии (рис. 1).