РУБРИКА: МАТЕРИАЛЫ КОНФЕРЕНЦИИ

Выявление некондиционных данных на основе модернизированного анализа методом кросс-валидации

К. А. Булдаков, А. О. Шадрин (ООО «ЛУКОЙЛ-Инжиниринг»)

Точность геологических моделей во многом зависит от качества скважинных данных, которые часто являются некондиционными из-за инструментальной погрешности или неправильных замеров. В статье проанализированы возможные методы обнаружения некондиционных данных на основании геостатистических приемов и методов машинного обучения, разработан алгоритм определения выбросов. На практическом примере работы алгоритма приведены результаты анализа, сделаны выводы о характеристиках скважин, влияющих на их некондиционность.

Ключевые слова: Python, программирование, исходные данные, скважина, моделирование, точность построений, статистика, геостатистика, машинное обучение.

Identification of invalid data based on modernized cross-validation analysis

K. A. Buldakov, A. O. Shadrin (LUKOIL Engineering LLC)

The reliability of geological models largely depends on the quality of well data, which is often incorrect due to instrument error or incorrect measurements. The article analyzes possible methods for detecting incorrect data based on geostatistical techniques and machine learning methods, and develops an algorithm for determining outliers. Using a practical example of the algorithm in action, the results of the analysis are presented and conclusions are drawn about the characteristics of wells that affect their inaccuracy.

Keywords: Python, programming, source data, well, modeling, accuracy of constructions, statistics, geostatistics, machine learning.

Введение

На этапе моделирования месторождений важную роль играют исходные данные, от качества которых зависит точность построенной модели. Одними из основных выступают скважинные данные, которые зачастую содержат некондиционные замеры, обусловленные инструментальной погрешностью или ошибками при интерпретации. Модель с такими данными имеет искаженный результат, некондиционные значения обнаруживаются только после ее построения и влекут за собой последовательное их удаление или исправление в выборке. Это требует многократного перестроения модели и занимает много времени.

Данная проблема освещена многими работами, среди которых [1], рассматривающая, как зависит достоверность структурных построений от точности инклинометрии скважин, а также предлагающая методику по расчету поправок в структурные построения с учетом инструментальных и субъективных факторов. В работе [2] рассматривался вопрос о важности качественных исходных данных для построения точной геологической модели, влияющей на все дальнейшие этапы разработки месторождения. Есть и другие работы, затрагивающие вопросы на анализируемую тему.

Существующие методы обнаружения некондиционных данных (выбросов) обладают определенными недостатками для анализа пространственных данных. Традиционные статистические методы не учитывают пространственную корреляцию и локальную изменчивость параметра, что в условиях неоднородного распределения скважин приводит к ложным выводам. Геостатистические подходы учитывают пространственную структуру, но могут быть чувствительны к наличию самих выбросов, искажающих расчет вариограмм и интерполяцию [4, 5, 6].

Таким образом, данная проблема является актуальной, особенно для месторождений Западной Сибири, поэтому мы считаем целесообразным применить современные методы статистического анализа и машинного обучения для решения задачи идентификации некондиционных скважинных данных для повышения точности геологических моделей.

Целью работы является разработка методики обработки первичных скважинных данных.

Для достижения цели поставлены следующие задачи:

1. Проанализировать существующие методы выявления выбросов, выбрать наилучшие решения для рассматриваемой задачи.

2. Сформулировать основные параметры исходных данных для использования в разрабатываемом алгоритме.

3. Выбрать оптимальный язык программирования для реализации алгоритма, ознакомиться с библиотеками и возможностями языка.

4. Разработать алгоритм, сочетающий статистический анализ и пространственные характеристики скважин.

5. Интегрировать в алгоритм метод интерполяции с целью проверки достоверности рассчитанных и фактических данных.

6. Визуализировать результаты работы алгоритма для удобства интерпретации.

Методология

В ходе анализа существующей литературы были сформулированы основные проблемы, которые возникают при подготовке исходных скважинных данных для построения модели, и возможные пути их решения.

Основной сложностью анализируемых параметров является неравномерное распределение скважин по площади, зависящее от геологического строения объекта. В большинстве случаев скважины разбуриваются в наиболее перспективных местах, а не по правильной геометрической сетке, вследствие чего требуется:

1. Необходимость структурирования исходных данных по координатам X, Y — использование алгоритмов KD-Tree, разбивающие 2D-плоскость на определенные области.

2. Определение кластеров скважин в пространстве с помощью полигонов Вороного — геометрически обоснованный подход определения соседей скважин путем разбиения плоскости на полигоны [3].

3. Учет неравномерности плотности сетки скважин — алгоритм должен адаптироваться к плотности данных через определенные стратегии для каждой области.

Приемы машинного обучения направлены на точечный анализ определенного признака для небольших наборов данных, при работе с большим массивом значений со значительной неоднородностью точек в пространстве данные алгоритмы теряют анализируемый параметр, что влечет за собой ложные заключения о рассматриваемом признаке. Методы геостатистики показывают корректные результаты при условии, что они используют точную модель вариограммы, которую можно получить только при исключении всех выбросов из выборки. Статистические методы определения аномальности основаны на отклонении среднего значения на три и более стандартных отклонения,

что зачастую дает ложные результаты, так как краевые скважины могут иметь отличные от основной выборки значения, более того, для правильного анализа характеристик также требуется очищенная выборка.

Основываясь на вышесказанном, приходим к следующим решениям:

4. Методы машинного обучения используют небольшие массивы данных, вследствие чего они неприменимы для месторождений Западной Сибири — использование метода кросс-валидации для оценки ошибки предсказания как критерия оценки аномальности скважин [7].

5. Неустойчивость геостатистических методов к наличию выбросов в выборке — использование методов интерполяции для построения собственной модели и анализа расхождений между фактическими и рассчитанными значениями [6].

6. Константные критерии выделения выбросов влекут ошибочные выводы — необходимы динамические пороги определения выбросов для каждой области.

7. Результаты алгоритма должны быть наглядными и понятными — вывод рассчитанных параметров в таблице Excel, а также визуализация итогов работы алгоритма.

Учитывая предложенные пути решения выделенных проблем, на языке Python написан алгоритм для выявления точек-выбросов — «Сигма_view**»**, который имеет структуру, представленную на рисунке 1.

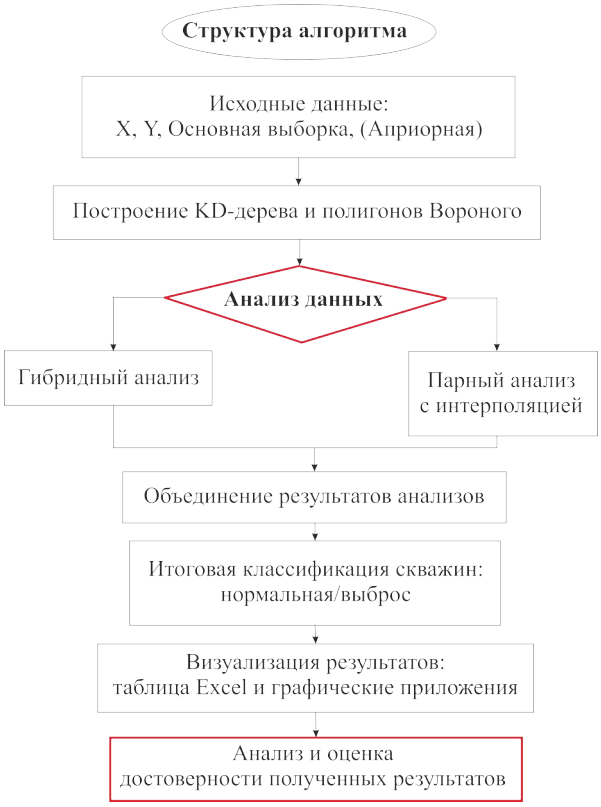

Рисунок 1. *Блок-схема разработанного алгоритма*

1. В качестве исходных данных для работы алгоритма будут использоваться таблицы РИГИС, содержащие координаты скважин (X, Y), значения анализируемого параметра («Основная выборка») и, опционально, параметр (t_0), который может выступать в качестве объясняющей переменной для анализа.

2. После загрузки данных автоматически строится k-мерное дерево KD-Tree для разделения выборки скважин в пространстве по координатам X, Y, а также полигоны Вороного для геометрического разбиения площади на блоки с целью определения кластеров скважин.

3. Модуль анализа скважинных данных состоит из двух принципиально новых алгоритмов.

Алгоритмы являются независимыми и предназначены для оценки аномальности точки на основе ее локального окружения.

Первый — гибридный, он предназначен для оценки аномальности точки на основе ее локального окружения. Его ключевой особенностью является адаптивность к плотности данных и локальной изменчивости точек. На рисунке 2 приведен принцип работы алгоритма.

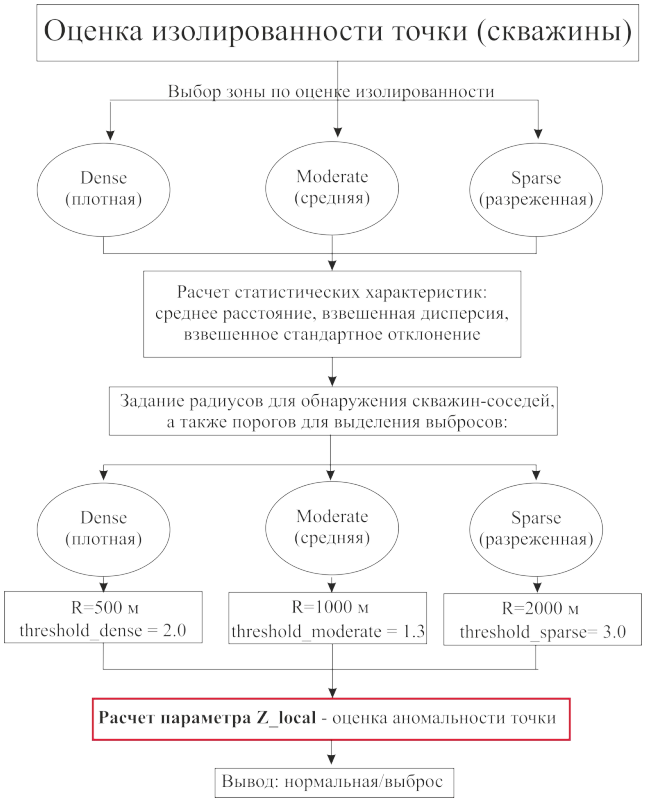

Рисунок 2. *Принцип работы гибридного анализа*

Основным этапом анализа является расчет параметра Z_{i_local}, так как в зависимости от его значения принимается решение об отнесении скважины к числу нормальных или аномальных:

$$Z_{i\_local} = \frac{v_i - \mu_w}{\sigma_w},$$

где v_i — значение в анализируемой точке i;

μ_w — взвешенный средний вес точек-соседей;

σ_w — взвешенное стандартное отклонение.

Далее полученное значение Z_{i_local} сравнивается с порогом для каждой зоны и принимается итоговое решение:

$$is_{outlier\_hybrid} = |Z_{i\_local}| > threshold_{strat},$$

где $threshold_{strat}$ — порог для соответствующей стратегии (например, $threshold_{dense}$= 2.0 и т. д.).

Второй анализ — **парный**, он представляет собой модернизированный алгоритм кросс-валидации, основанный на принципе пространственной

согласованности локального поля данных и коллективного свидетельства точек. Принцип работы парного анализа представлен на рисунке 3.

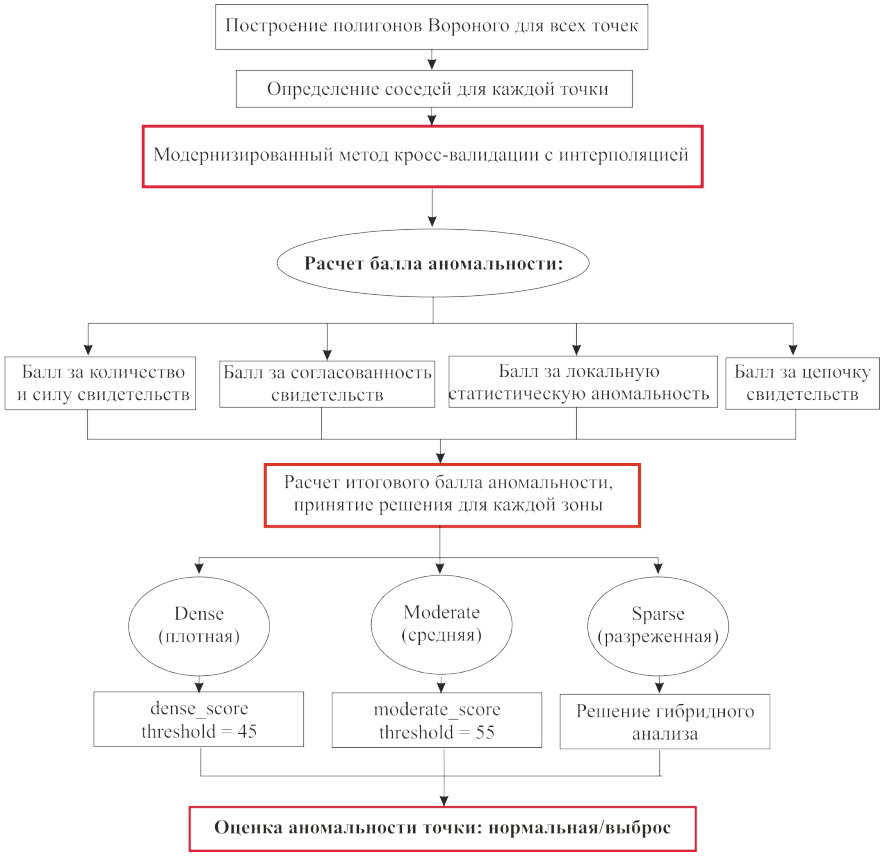

Рисунок 3. *Принцип работы парного анализа*

Ключевым этапом работы алгоритма является сбор свидетельств аномальности с помощью интерполяции методом обратных взвешенных расстояний и кросс-валидации:

1. Интерполяция без исключения — для точки-свидетеля k рассчитывается ожидаемое значение методом интерполяции по всем ее соседям, вычисляется ошибка (модуль разности между измеренным и предсказанным значением) и приводится к безразмерному виду.

2. Интерполяция с исключением анализируемой точки — из множества соседей точки k исключается точка i. Интерполяция повторяется, давая новое предсказание и новую нормированную ошибку. Рассчитывается улучшение точности интерполяции, если улучшение значимое, то фиксируется свидетельство.

3. Интерполяция с исключением пары точек — аналогичная процедура выполняется для одновременного исключения пары точек (i, j). Если улучшение при исключении пары больше, чем при исключении

любой из точек по отдельности, это фиксируется как более весомое парное свидетельство.

Система баллов в анализе позволяет избежать ложных срабатываний и повышает надежность получаемых результатов.

4. Объединение результатов гибридного и парного анализов необходимо, так как они эффективны для разных зон: парный анализ — плотные и средние, гибридный — разреженные. В качестве итоговых выбросов выбирается пересечение решений анализов.

5. Итогом анализа является визуализация, которая представлена набором графических приложений и таблицей Excel с основными статистическими параметрами. На рисунке 4 показан пример выводимых результатов.

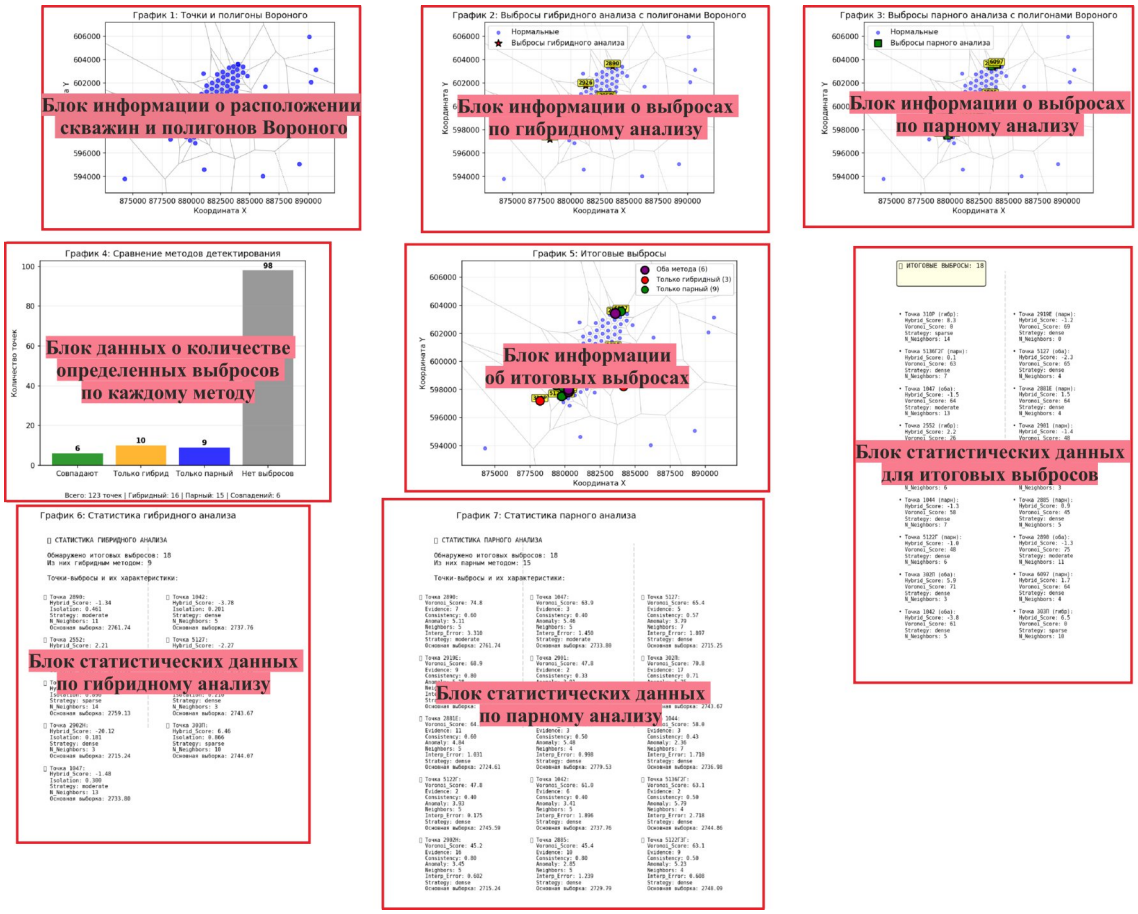

Рисунок 4. *Пример визуализации результатов*

Результаты работы алгоритма и их анализ

Разработанный алгоритм тестировался на массиве скважинных данных с более чем 10 тысячами значений для одного из месторождений Западной Сибири.

Для анализа результатов была построена структурная карта, представленная на рисунке 5.

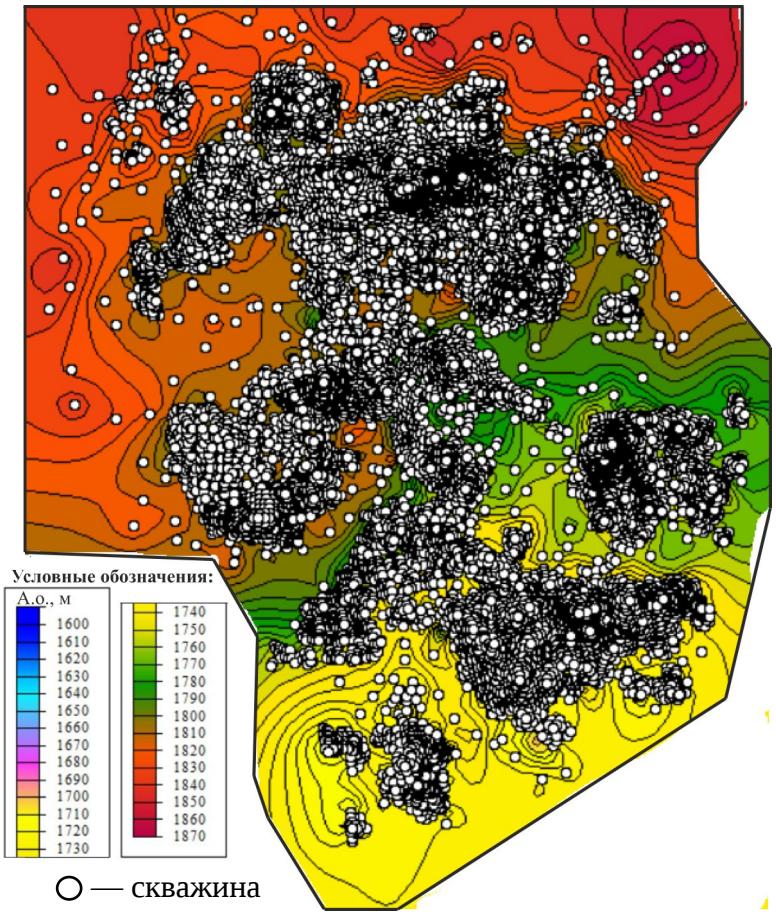

Рисунок 5. *Структурная карта, построенная по скважинным данным*

Для рассматриваемой ситуации алгоритм выдает следующие результаты:

- общая статистика: всего точек — 13557;
- выбросы гибридного анализа: 2234;
- выбросы парного анализа: 3284;
- итоговые выбросы: 1367.

При беглом анализе построенной карты явно выделяются выбросы, приведенные на рисунке 6. Выделенные скважины резко отличаются по абсолютному значению от соседних скважин на 30–70 метров, что нашим

детектором помечается как выброс. Всего на карте в районе 250–300 очевидных выбросов, подобные типы аномалий алгоритм идентифицировал верно.

Однако в категорию итоговых алгоритмом отнесено 1367 скважин, а это означает, что помимо очевидных выбросов выделены скважины, которые относятся к типу наибольшей степени опасности, так как их некондиционность не очевидна при беглом анализе, что может повлиять на то, что подобная скважина будет учтена при построениях и снизит точность нашей модели. Скважины по типу 1848Н и 1332Н также идентифицируются как аномальные, так как при незначительном расстоянии между точками разница абсолютных отметок не может составлять несколько метров (рис. 7).

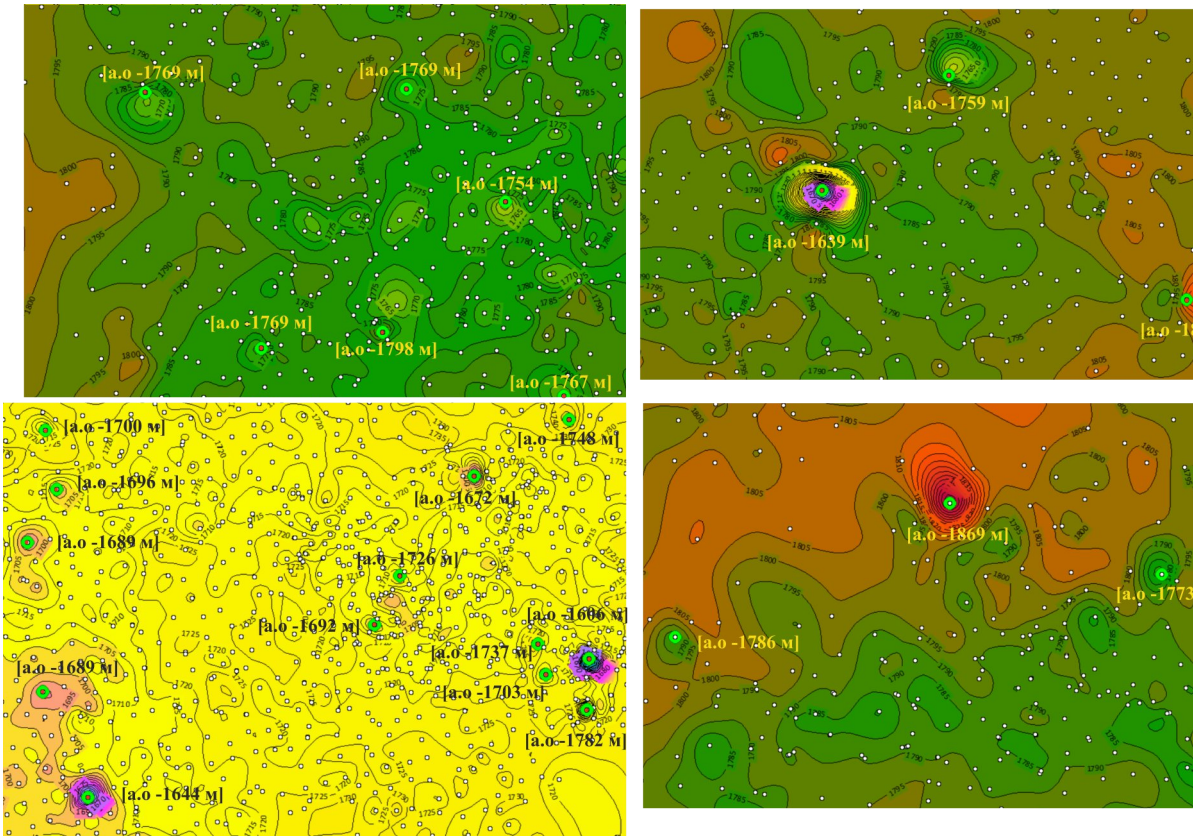

Рисунок 6. *Явные выбросы*

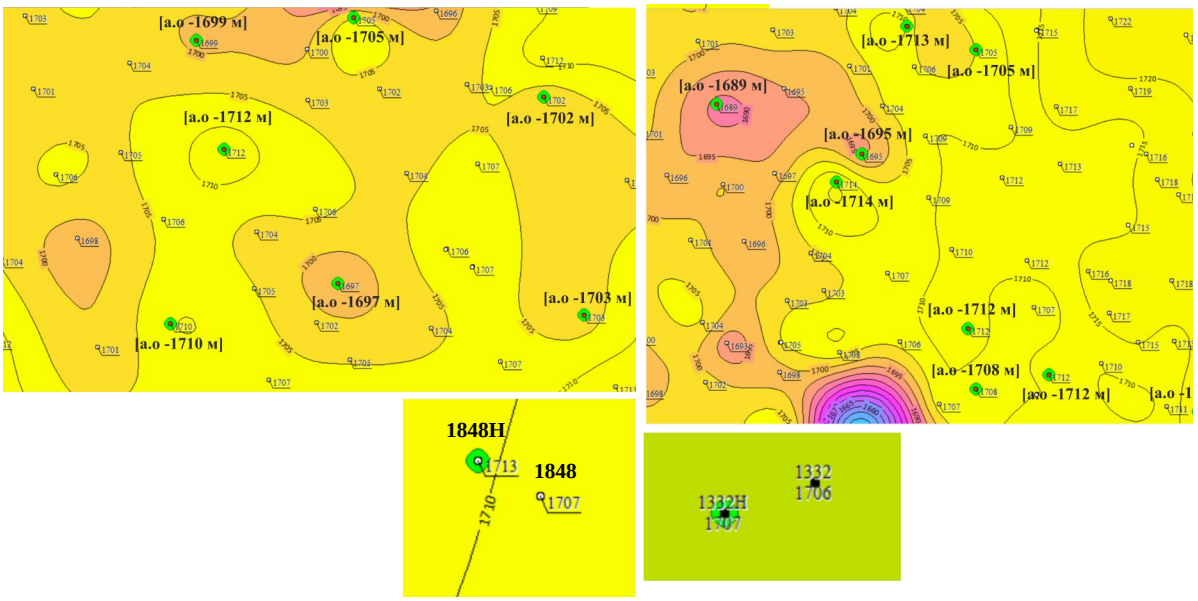

Рисунок 7. *Неочевидные выбросы*

В научной литературе существует предположение о том, что сложная конструкция скважины является потенциальным маркером некондиционных данных [1].

Проанализировав полученные статистические характеристики для всей выборки (таб. 1) и для скважин-выбросов (таб. 2), приходим к тому, что предположение не подтвердилось, так как удлинение и углы входа в пласт не имеют значимых отличий. Соответственно — большое удлинение или в целом сложная конструкция скважин не является потенциальным маркером некондиционных данных (рис. 8).

Таблица 1. *Описательная статистика для всей выборки скважинных данных*

Variable	Описательная статистика скважинных данных								
	Valid N	Mean	Median	Minimum	Maximum	Percentile 10,00000	Percentile 50,00000	Percentile 90,00000	Std.Dev.
ALTITUDE	13557	59,11	59,65	39,560	103,1	48,200	59,65	68,04	8,55
ELONG	13533	163,62	129,80	0,100	1119,6	22,000	129,80	343,80	142,71
ANGLE	13541	22,15	15,45	0,080	92,0	2,817	15,45	69,61	22,19

Таблица 2. *Описательная статистика для скважин-выбросов*

Variable	Описательная статистика для скважин-выбросов								
	Valid N	Mean	Median	Minimum	Maximum	Percentile 10,00000	Percentile 50,00000	Percentile 90,00000	Std.Dev.
ALTITUDE	1367	58,44	58,38	39,560	103,0	48,190	58,38	67,15	7,86
ELONG	1363	148,36	109,80	0,100	1119,6	20,100	109,80	325,10	141,70
ANGLE	1366	17,60	11,00	0,102	88,2	2,250	11,00	38,27	20,27

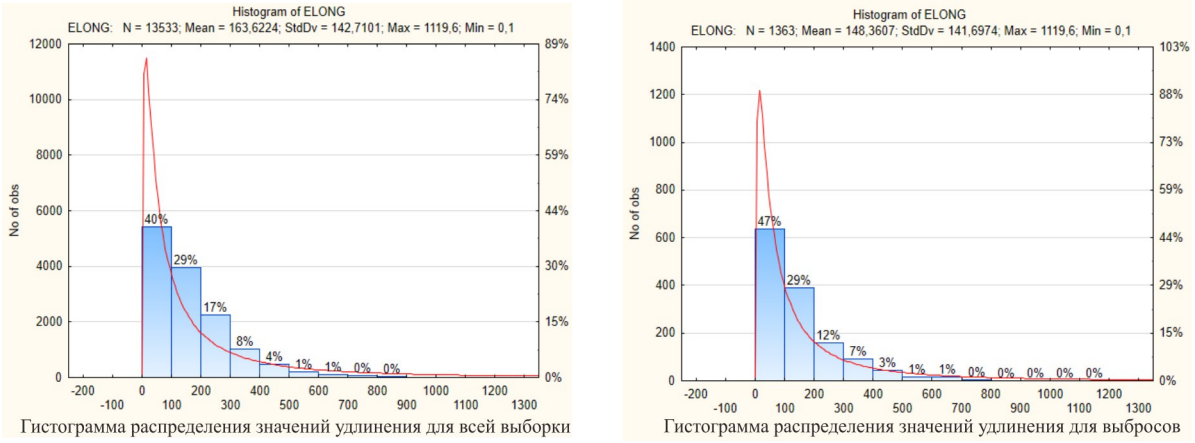

Рисунок 8. *Сопоставление гистограмм распределения значений удлинения*

Заключение

Таким образом, точность геологических моделей напрямую зависит от качества скважинных данных, которые зачастую содержат некондиционные замеры. Модели, построенные с учетом подобных значений, имеют неверные результаты и являются менее достоверными для дальнейших решений.

В настоящее время нет алгоритма, который бы учитывал локальную статистику и пространственные характеристики скважин. Поэтому разработка нового алгоритма выделения выбросов является актуальной задачей.

Разработанный алгоритм основан на современных методах статистического анализа и машинного обучения. Он является эффективным инструментом при подготовке скважинных данных, так как обеспечивают комплексный учет всех возможных факторов аномальности скважин и предоставляет достоверные и интерпретируемые результаты.

Тестирование алгоритма для выборки из более чем 10 тысяч скважин показало, что он способен обнаруживать выбросы разного характера.

Опровергнуто предположение о том, что удлинение или сложная конструкция скважин является потенциальным маркером некондиционных данных.

Данный алгоритм может быть интегрирован в программное обеспечение для моделирования, например в Isoline GIS или Golden Surfer, с целью выделения аномальных значений на этапе загрузки скважинных данных в модель или предложения выбора их игнорирования при построении структурных поверхностей.

Следующим этапом доработки алгоритма является разработка автоматической настройки параметров алгоритма на основе машинного обучения в зависимости от плотности сетки скважин и геологического строения объекта, а также полноценная интеграция сейсмических данных в анализ для повышения достоверности результатов.

Список литературы

1. Щергина Е. А. Практика оценки инклинометрии скважин в моделировании нефтегазовых объектов / Е. А. Щергина, А. Б. Сметанин, В. Г. Щергин, А. С. Мартынов // Геология, геофизика и разработка нефтяных и газовых месторождений. — 2022. — № 12(372). — С. 31–41. — https://doi.org/10.33285/2413-5011-2022-12(3712)-31-41.

2. Карамурзаева А. Б. Анализ достоверности скважинных данных, заложенных в геологическую модель месторождений Бузачинского свода / Карамурзаева А. Б. // Науки о Земле и смежные экологические науки. — 2022. — С. 95–99. — https://doi.org/10.56525/FYEE6657.

3. Забоева А. А. Декластеризация исходных данных при построении и контроле качества трехмерных геологических моделей / А. А. Забоева, А. С. Предеин, И. С. Никитин. // Нефть и газ. — 2011. — № 3. — С. 15–21.

4. Нехороших Д. С. Стохастическое моделирование пространственно-распределенных данных по окружающей среде / Нехороших Д. С., Демьянов В. В., Каневский М. Ф., Чернов С. Ю., Савельева Е. А. — Москва: Институт проблем безопасного развития атомной энергетики РАН, 2000.

5. Демьянов В. В. Геостатистика: теория и практика / В. В. Демьянов, Е. А. Савельева. — М.: Наука, 2010.

6. Шарапов И. П. Применение математической статистики в геологии / Шарапов И. П. — М.: Недра, 1971.

7. Кросс-валидация. — Текст: электронный. — URL: https://education.yandex.ru/handbook/ml/article/kross-validaciya.

References

1. Shchergina E. A. The practice of evaluating well inclinometry in the modeling of oil and gas facilities / E. A. Shchergina, A. B. Smetanin, V. G. Shchergin, A. S. Martynov // Geology, geophysics, and development of oil and gas fields. — 2022. — No. 12(372). — P. 31–41. — https://doi.org/10.33285/2413-5011-2022-12(3712)-31-41 (in Russ.).

2. Karamurzaeva A. B. Analysis of the reliability of well data used in the geological model of the Buzachinsky arch deposits / Karamurzaeva A. B. // Earth Sciences and Related Environmental Sciences. — 2022. — P. 95–99. — https://doi.org/10.56525/FYEE6657 (in Russ.).

3. Zaboeva A. A. Declustering of source data in the construction and quality control of three-dimensional geological models / A. A. Zaboeva, A. S. Predein, I. S. Nikitin. // Oil and Gas. — 2011. — No. 3. — Pp. 15–21 (in Russ.).

4. Nekhoroshikh D. S. Stochastic modeling of spatially distributed environmental data / Nekhoroshikh D. S., Demyanov V. V., Kanevsky M. F., Chernov S. Yu., Savelyeva E. A. —Moscow: Institute for Problems of Safe Development of Nuclear Energy, Russian Academy of Sciences. — 2000 (in Russ.).

5. Demyanov V. V. Geostatistics: Theory and Practice / V. V. Demyanov, E. A. Savelyeva. — Moscow: Nauka, 2010 (in Russ.).

6. Sharapov I. P. Application of mathematical statistics in geology / Sharapov I. P. — Moscow: Nedra, 1971 (in Russ.).

7. Cross-validation [Electronic resource]. — URL: https://education.yandex.ru/handbook/ml/article/kross-validaciya (in Russ.).