**Постановка задачи**

**Геофизическая постановка задачи**

Газовая скважина исследована аппаратурой мультиметодного многозондового нейтронного каротажа (ММНК) и по методике ММНК-Кг определены  значений коэффициента газонасыщенности *Кг* в разных пластах разреза. После этого с целью проверки корректности всей технологии проводятся испытания скважины на приток, в которых на качественном уровне измеряется состав фактически добываемой продукции газоводяной смеси. Это означает, что продукция классифицируется на небольшое число градаций  3--5 разбиений шкалы *Кг* на эквидистантные широкие интервалы протяженностью по *Кг*= [*Кг*] / , где [*Кг*] – максимально возможный диапазон изменения *Кг* в исследуемых геолого-промысловых условиях. Например, наиболее часто используемыми интервалами изменения *Кг* для типовых [*Кг*] = [0,1],  = 4 являются по терминологии газовиков: <<вода (0--0.25), вода+газ (0.25--0.5), газ+вода (0.5--0.75), газ (0.75--1)>>. Затем проверяются доли , % правильных попаданий предсказанных ММНК-Кг численных значений *Кг* в соответствующие им широкие интервалы *Кг*. Если большинство этих долей  > 80-90%, то технология признается корректной, т.к. получила качественное подтверждение по результатам испытаний, считающихся одним из наиболее прямых и убедительных способов тестирования методик в скважинной геофизике. Представляется очевидным, что при достаточно большом числе  определений *Кг* и, разумеется, при условии выполнения достаточно жесткого критерия подтверждаемости по  любая разумно введенная оценка фактической средней погрешности определения *Кг* должна дать величину, существенно меньшую широких интервалов разбиения *Кг*. Другими словами, это означает, что технологию ММНК можно будет переквалифицировать из качественной по способу ее подтверждения в количественную по фактически достигаемому уровню погрешности определения *Кг*.

Поэтому задачами работы явились следующие:

• численное обоснование этого утверждения на основе теории вероятностей с разработкой алгоритма и программы расчета фактической средней погрешности определения *Кг* по всем имеющимся данным определений и подтверждений в исследуемой скважине;

• численное изучение поведения погрешности в зависимости от варьируемых параметров , , [*Кг*],  и характера вероятностного распределения найденных значений *Кг* на [*Кг*], (*Кг*) - от равномерного до гауссового с большой дисперсией. Диапазоны изменения варьируемых параметров:

- 

- 

- [*Кг*] = [0.5, 1], [0.25, 1], [0, 1]

- 

• выдача практических рекомендаций по выбору единственного управляемого параметра  в зависимости от априори задаваемых геофизиками и газовиками параметров  и [*Кг*], а также от фактически получившихся характеристик – распределения (*Кг*) и значений  в результате сопоставления определений и подтверждений *Кг*.

**Математическая формулировка**

Цель -- предсказать погрешность  выдаваемых нашей программой значений  искомого параметра .

Область изменения  [*Кг*] разбита на интервалы , , для каждого из которых есть  экспериментов. Считается, что искомая погрешность  может меняться от интервала к интервалу, но постоянная внутри интервала (т.е. точнее писать ). Фиксируем  и работаем в выбранном интервале, поэтому далее индекс интервала  опущен.

Есть  экспериментов, про которые известно, что в каждом из них истинное значение  попало в интервал, т.е.  . На каждый из этих экспериментов у нас есть результат работы нашей программы . Предполагается, что истинное значение  распределено по Гауссу со средним  и некой дисперсией , т.е.



где  - вероятность того, что  верно, а



-- Гауссово распределение.

Ищем зависимость



такую, что вероятность реализации описанной выше ситуации (т.е. что все истинные значения попали в интервал) , т.е.



Из сторонних соображений считается известным минимально возможная погрешность , т.е. если метод выдает , то считаем .

**Предлагаемое решение**

**Идея и приближения**

Задав , можно посчитать вероятность реализации ситуации, описанной в постановке -- попадания всех истинных значений параметра , распределенных по Гауссу каждый около своего , в интервал . Далее предположение - эта вероятность равна нашей целевой вероятность . Не очевидно, почему это должно выполняться точно (скорее всего это не выполняется), но для оценки предложено использовать такую модель.

Поясним разумность данного выбора. Будем брать пробные  и смотреть как от этого зависит ожидаемое поведение истинных значений  относительно наших точек . Для примера возьмем весь интервал  и предположим что у нас имеются 5 точек, для которых наша программа выдала ответы 0.22, 0.3, 0.35, 0.4, 0.43. Если предположить, что погрешность наший предсказаний , то плотность вероятности для каждого из 5 истинных значений будет выглядет так:

./pics/peaks\_med

Разным цветам отвечают разные эксперименты. Для каждого эксперимента: проколотый круг на оси Х -- наше предсказание ответа, купол -- распределение плотности вероятности того, что истинное значение ответа примет значение , в зависимости от .

На Рис. видно, что для всех точек кроме синей почти вся кривая (вероятность = площадь под кривой) находится в исследуемом интервале. Это значит, что при  для всех точек кроме синей вероятность того, что интинное значение параметра попадет в интервал, равна почти 100%. Синяя же точка находится на расстоянии  (в данном случае 0.02), что значит, что вероятность того, что истинное значение параметра в синем эксперименте попадет в интервал  будет . Попадания истинных значений в интервал -- события независимые, поэтому вероятность реализации картины в целом будет произведением вероятностей попадания каждого значения в интервал по отдельности. В нашем случае все вероятности кроме синей , поэтому общая вероятность  синяя вероятность  84%. Это значит, что если бы погрешность нашей программы была 0.02, то вероятность случатся тому что случилось на рассматриваемых 5 экспериментах в совокупности была бы 84 %. Поняв это, можно решить обратную задачу: сказать, что мы верим эксперименту на скажем 95%, и найти такое , при котором вероятность его реализации как раз будет 95%. Понятно, что задача решаема -- если в предыдущем примере мы возьмем , то распределение вероятностей для истинных значений будет

./pics/peaks\_sharp\_med

Обозначения аналогичны Рис. . Вероятность реализации эксперимента >99%, что эквивалентно практически полному доверию эксперименту.

Если же все наши экспериментальные точки лежат ближе к центру исследуемой области, то оценка на погрешность выходит грубой. Это понятно из следующего примера. Сдвинем точку 0.22 из предыдущего примера в точку 0.28. График для  будет выглядить как

./pics/peaks\_smooth

Обозначения аналогичны Рис. . Вероятность реализации эксперимента 92.5%.

Т.е. при налиции точки 0.22, близкой к левой границе иследуемого интервала 0.2, вероятность реализации эксперимента уже при  была 84%, что говорило о том, что в реальности скорее всего погрешность была меньше. Здесь же даже при  вероятность все еще >90%. Это на самом деле логично, т.к. если все наши точки у центра интервала, то единственный известный нам факт (на котором и строится вся оценка опгрешности) о том, что все истинные значения попали в интервал, поволяет отбросить только самые большие значения погрешностей.

Заинтересовавшийся читатель может найти весь инходный код проекта и мои контакты для вопросов и предложений https://github.com/PolyachenkoYA/errorEstimationIOGTздесь. В частности по ссылке лежит программа для оценки ошибок по описываемому здесь методу.

Теперь приведем аналитическое выражение описанной идеи:

**Аналитика**

Вероятность попадания -ой истинной точки в интервал



Введем функцию (известную как функция ошибок):



Попадание каждого истинного значения в интервал - независимое событие, поэтому вероятность реализации нашей совокупности экспериментов



Для нахождения желаемого  решаем уравнение на  при заданном .



Уравнения явно не решается аналитически. Но несложно показать, что функция  монотонна, а интервал изменения  известен и невелик, откуда следует, что уравнени легко решается численно даже самыми простейшими методами вроде деления отрезка пополам. В примерах использован алгоритм, реализованный в функции  в Matlab и описанный в

**Продолжение примера аналитикой**

Продолжим использовать 5 точек из раздела 2.1. В разделе разделе 2.1 был описан алгоритм как мы задавшись определенным  можем оценить вероятность , с которой при этом  реализовались бы имеющиеся у нас экспериментальные данные. Сделав так для мноих различных , можно для каждого из них получить свое значение  (Синяя кривая на Рис. ).

./pics/vis

Зависимость  вероятности реализации эксперимента при данной погрешности программы. Красная линия -- принятая минимально возможная погрешность , Синяя вертикаль -- найденная оценка, синяя горизонталь -- наш выбор , зеленый пунктир - минимальное расстояние точек до границы.

Видно, что наличие множества (> 1) точек позволяет улучшить оценку с очевидного значения минимального расстояния до границы -- синяя линяя линия левее зеленой, т.е. оценка по предлагаемому методу лучше чем наивная оценка сверху на глаз.

**Результат применения**

**Типичные значения**

Можно исследовать, как оценка погрешности зависит от количества имеющихся экспериментальных данных в <<усредненном>> случае, когда ответы нашей программы расположены в интервале на равных промежутках:

./pics/uni\_line

Равномерное расположение 10 пробных точек в интервале .

Сгенерировав таким равномерным образом несколько наборов <<предсказаний>> нашей программы, можно получить зависимость получаемой оценки погрешности от количетва имеющихся экспериментальных точек:

./pics/linear

Зависимость  оценки погрешности программы от количества экспериментальных точек при их равномерном распределении в интервале.

Можно делать так для разных наборов параметров, описанных в геофизической постановке задачи. Приведем некоторые полученные таким образом зависимости для типичных значений параметров:

./pics/fig\_AB figureЗависимость  оценки погрешности программы от количества экспериментальных точек. Различные интервалы допустимых *Кг*. *Кг*

|  |  |  |
| --- | --- | --- |
| *Кг* | 0.05 | 0.02 |
| [0; 1.0] | 3 | 7 |
| [0.25; 1.0] | 2 | 5 |
| [0.5; 1.0] | 1 | 3 |

Минимальные значения N, необходимые для достижения данного  при данном *Кг*.

./pics/fig\_P figureЗависимость  оценки погрешности программы от количества экспериментальных точек. Различные степени доверия эксперименту. 

|  |  |  |
| --- | --- | --- |
|  | 0.05 | 0.02 |
| 80% | 2 | 7 |
| 90% | 2 | 5 |
| 97% | 1 | 4 |

Минимальные значения N, необходимые для достижения данного  при данном .

./pics/fig\_M figureЗависимость  оценки погрешности программы от количества экспериментальных точек. Различные разбиения типичного интервала. 

|  |  |  |
| --- | --- | --- |
|  | 0.05 | 0.02 |
| 3 | 3 | 7 |
| 4 | 2 | 5 |
| 5 | 1 | 4 |

Минимальные значения N, необходимые для достижения данного  при данном .

Может быть также интересен наиболее сложный случай:

, , *Кг* = [0; 1.0]

При таких параметрах для достижения  нужно , а для  нужно . Т.е. при наличии 12 точек в интервале можно с уверенность говорить о подтверждении хорошей точности метода в данной области *Кг*.

**Теоретический анализ**

На глаз зависимость на Рис. близка к , что ожидаемо, т.к. погрешность в основном определяется минимальным расстояние до границы, которое при выбранной расстановке точек убывает как .

Можно проверить отклонения от закона  -- рис.().

./pics/log

, логарифмический масштаб, попытка линеаризации

Видно, что наклон с правда близок к , но небольшие отклонения от линейности есть.

**Описание и рекомендации к программе**

Для удобства применения изложенного метода создана программа, позволяющее получить оценку погрешности для любого заданного набора экспериментов.

Окно программы выглядит так:

./pics/whole\_window\_nums

Окно программы. 1,2 -- границы анализируемого интервала. 3 -- априорная минимально допустимая погрешность. 4 -- степень доверия эксперименту. 5 -- имеющиеся результаты работы программы. 6 -- сделать расчет с введенными данными.

Поясним, что изображено на рисунках:

**Исследуемый интервал, ввод имеющихся данных**

./pics/whole\_window\_interval

Значения полей 1 и 2 -- значения  и , т.е. границ исследуемного интервала.

./pics/whole\_window\_interval2

Таблица 5 -- имеющиеся ответы программы.

**Основной рисунок**

./pics/whole\_window\_34

Поле 3 -- априорное значение минимально допустимой погрешности. Оно задается по соображениям пользователя исходя из сторонней информации. Поле 4 -- степень доверия эксперименту. Если мы уверены в эксперименте на 95% -- пишем 0.95. Основная задача программы состоит в использовании этой величины -- по точкам 5 и интервалу 1,2 программа строит голубую кривую и находит тот Х (т.е. тот ), при котором кривая принимает значение в поле 4 (т.е. вероятность реализации эксперимента совпадает с нашими ожиданиями).

./pics/whole\_window\_numsLeg

Описание легенды. 1 -- зависимость вероятности реализации имеющейся совокупонсти экспериментов от предполагаемой погрешности предсказаний. 2 -- априорно заданная минимальная погрешность. 3 -- минимальное расстояние от заданного множества точек до блжайшей границы интервала. 4 -- степень доверия эксперименту. 5 -- ответ программы, получанная оценка погрешности предсказаний.

[heading=bibintoc]