КРОСС-ВАЛИДАЦИОННАЯ ОЦЕНКА ПАРАМЕТРОВ РЕГРЕССИИ ПРИ ИНТЕРПРЕТАЦИИ СЕЙСМИЧЕСКИХ ДАННЫХ

CROSS-VALIDATION ESTIMATION OF REGRESSION IN SEISMIC INTERPRETATION

УДК 550.832
© Р.А. Суханов, А.В. Екименко, 2017

Р.А. Суханов, А.В. Екименко
Научно-Технический Центр «Газпром нефти» (ООО «Газпромнефть НТЦ»)

Электронный адрес: sukhanov.ra@gazpromneft-ntc.ru, ekimenko.av@gazpromneft-ntc.ru

Ключевые слова: кросс-валидация, перекрестная проверка, оценка устойчивости модели

R.A. Sukhanov, A.V. Ekimenko Gazpromneft NTC LLC, RF, Saint-Petersburg

Regression analysis, as the most simple and understandable way to predict parameters, has received a wide coverage, including in geosciences. The error of this method can be estimated by such parameters as correlation coefficient and standard error. Cross-validation (cross-validation) is one of the ways to assess the stability of a model in which part of the input does not participate in the analysis, but it is uses for evaluation. Usually cross-validation is not used in regression analysis, however, using these approaches in aggregate, it is possible to estimate the correlation coefficient and the standard error for each of the implementations, as well as evaluate the contribution of input data to a particular well. It is possible to identify problem points in the initial data at the stage of mapping.

Keywords: cross-validation, evaluation of the stability of the model

ВВЕДЕНИЕ

Регрессионный анализ как наиболее простой и понятный способ прогнозирования тех или иных параметров получил широкое распространение, в том числе в областях геонауки (сейсморазведка, петрофизика, геология).
В рамках комплексной интерпретации сейсмических данных построение прогнозных моделей на основе регрессии используется при структурной интерпретации и прогнозе фильтрационно-емкостных свойств (ФЕС) пластов. В первом случае анализируется регрессия время – глубина, и при установлении значимой связи карты изохрон пересчитываются в глубины с использованием оцененного уравнения. Во втором случае анализируются, как правило, одна или несколько карт динамических параметров (амплитуды, частоты и др.) и их связь с пористостью, эффективными толщинами, проницаемостью. Таким образом, построив регрессионные модели по результатам измерений в точках скважин, можно прогнозировать ФЕС пластов в межскважинном пространстве по материалам сейсморазведки.
Построение прогнозных моделей всегда сопровождается оценками погрешности прогноза. Погрешность регрессионных моделей можно оценить по коэффициенту корреляции и величине стандартной ошибки. Целесообразно в рамках анализа данных приводить оба этих параметра вместе, поскольку их использование в отдельности не может полностью охарактеризовать прогнозную точность модели. Например, при изучении взаимосвязи времен и глубин часто фиксируются высокие коэффициенты корреляции, которые свидетельствуют о применимости регрессии для трансформации время – глубина, вместе с тем высокая стандартная ошибка может сделать использование такой модели неэффективным.
Одним из способов оценки устойчивости модели является кросс-валидация (перекрестная проверка). Методика заключается в том, что часть входных данных не принимает участия в построении прогнозной модели, а используется для оценки погрешности прогноза.
В данной работе предлагается выполнять оценку коэффициента корреляции и величины стандартной ошибки при исключении части информации. Исключение из выборки разного числа скважин позволяет оценить вклад входных данных той или иной скважины, а также идентифицировать проблемные значения в исходном наборе данных. Для решения описанных задач был разработан инструмент, который является идентификатором таких проблем, пересмотр которых снизил бы неопределенность при прогнозе тех или иных свойств пласта.

ОПИСАНИЕ И РЕАЛИЗАЦИЯ ПЕРЕКРЕСТНОЙ ПРОВЕРКИ

Существует несколько подходов к оценке устойчивости модели на основе кросс-валидации, различающиеся способом разделения всей выборки на обучающую и контрольную. В рамках данной работы кратко описаны две разновидности таких способов: Leave One Out (LOO), при котором исключается из расчетов по одному значению, и Leave P Out (LPO), при котором исключается несколько значений [P>1]. Эффективность и информативность первого и второго способов зависит от числа сформированных перекрестных выборок. Общая формула из области комбинаторики, позволяющая описать конечное число итераций, имеет следующий вид:

$$C_n^k = \frac{n!}{k!(n-k)!},$$

где n, k – число скважин соответственно общее и в комбинации.

Подход LOO [1–3] является быстрым и считается классическим способом оценки, например, точности структурных построений для исследования погрешности интерполяции. Для оценки параметров регрессии при отключении нескольких скважин одновременно данный способ не подходит, необходимо использовать метод LPO.

Подход LPO [4] является более длительным способом оценки устойчивости модели и во многом зависит от числа элементов (скважин), участвующих в анализе, и числа элементов (скважин), задействованных для тестирования/анализа. Сбор статистических данных на каждом этапе кросс-валидации требует много времени и немалого объема информации, сравнимого с BigData. Достоинством способа LPO является возможность оценить параметрь регрессии для любой комбинации.

В настоящее время кросс-валидация включена в ряд программных продуктов, но работает только в режиме LOO, где модель задается не уравнением регрессии, а интерполяцией в межскважинном пространстве. Для применения LOO и LPO при регрессионном анализе авторами было разработано программное решение, позволяющее решить поставленные задачи.

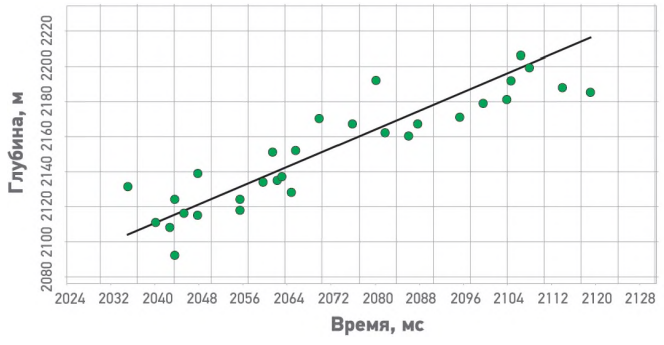

Рис. 1. Зависимость время – глубина

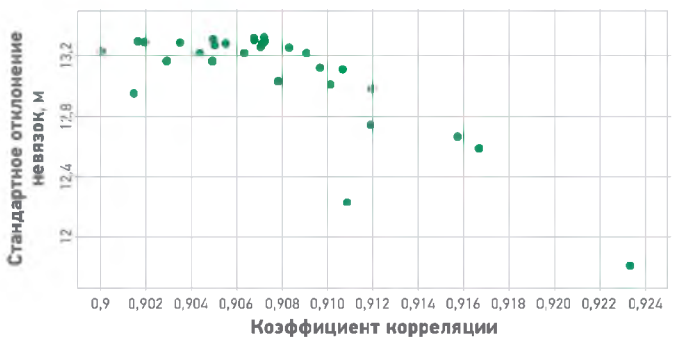

Рис. 2. Результаты работы алгоритма кросс-валидации LOO

ПРИМЕНЕНИЕ КРОСС-ВАЛИДАЦИИ

Использование кросс-валидации при регрессионном анализе позволяет оценить вклад каждой скважины, участвующей в построении модели. Имеется синтетический набор входных данных для построения структурного плана. В исходные данные было внесено до 2,5 % шума. Зависимость время – глубина приведена на **рис. 1**. Данная зависимость характеризуется не только достаточно высоким коэффициентом корреляции (90,7 %), но и высоким стандартным отклонением невязок. Невязки варьируются от –26,88 до 32 м, стандартное отклонение составляет 13,10 м (**рис. 2**).

Исходя из **рис. 2** можно сделать вывод о возможности повышения коэффициента корреляции от 90,7 до 92,3 % и снижения стандартного отклонения невязок с 13,10 до 11,80 м. Отключение одной из скважин не всегда дает положительный результат, иногда параметры регрессии могут изменяться в обратном направлении. В данном случае при использовании двух параметров (коэффициентом корреляции и стандартным отклонением) можно отметить повышение «коррелируемости» данных и снижение стандартного отклонения.

Кросс-валидация методом LPO позволяет оценить параметры регрессии для одновременно отключаемых скважин. Таким образом, нет необходимости итерационно задействовать метод LOO для идентификации 1-й, 2-й, n-й скважины и пересмотра их, можно задать нужное число скважин на отключение, алгоритм сам перебер ет все возможные комбинации и рассчитает для каждой итерации коэффициент корреляции и стандартное отклонение. После этого нужно только выбрать группу скважин для анализа входных данных, связанных с этими параметрами.

По результатам применения метода LPO (**рис. 3**) можно выделить максимально позитив-

ные значения параметров регрессии при отключении групп скважин. Так, при отключении группы из трех скважин коэффициент корреляции составляет 94,4 %, стандартное отклонение – 10,31 м; группы из двух скважин – соответственно 93,3 % и 11,12 м.

Методом LPO можно выявить комбинацию скважин, на которую стоит обратить дополнительное внимание. Способ достаточно ресурсоемкий, требует много времени для расчета.

Для оптимизации расчетов был реализован метод Bootstrap, основанный на многократной генерации псевдослучайных выборок отключаемых скважин методом Монте-Карло. Способ Bootstrap работает быстрее и дает более полную оценку модели, так как число отключаемых скважин варьируется от 1 до n–3, где n – общее число скважин.

По результатам применения метода Bootstrap (**рис. 4**) можно глобально оценить устойчивость модели. Из рис. 4 видно, что имеются варианты, которые стремятся к высокому, близкому к 100 % коэффициенту корреляции и низкому стандартному отклонению. Отметим, что данный метод работает с псевдослучайными выборками, где наилучшие параметры регрессии могут быть обусловлены низкой выборкой входных данных.

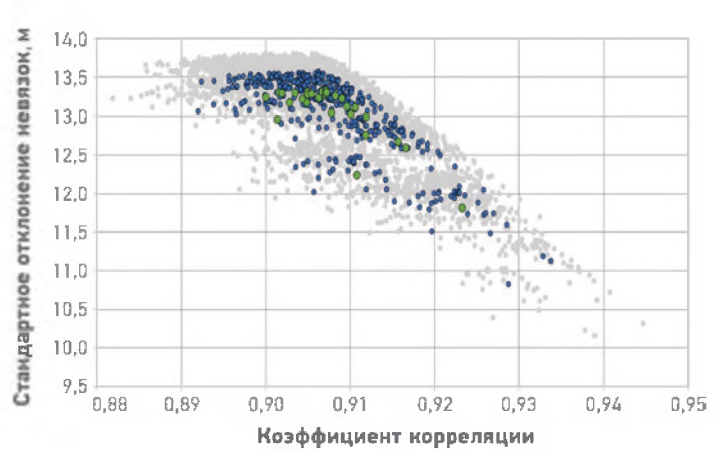

Рис. 3. Результаты работы алгоритма кросс-валидации (метод LOO+LPO)

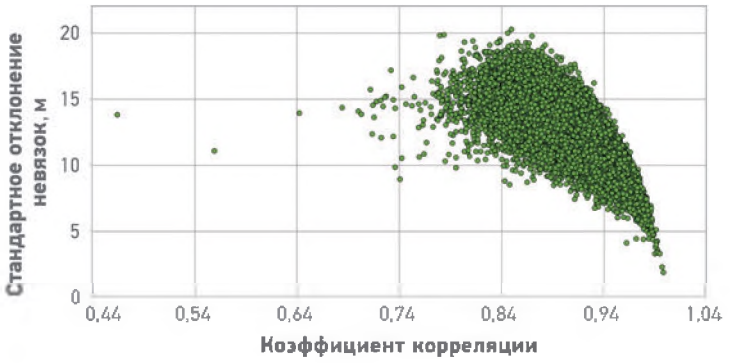

Рис. 4. Результат работы алгоритма кросс-валидации (метод Bootstrap)

ЗАКЛЮЧЕНИЕ

Для каждого прогноза необходим анализ неопределенностей. Описанный в статье инструмент позволяет снизить неопределенность, связанную с погрешностью в исходных данных для карт прогноза. Примером таких неопределенностей может служить неустойчивая фаза при корреляции отражающих горизонтов по данным сейсморазведки, интерполяция в области 2D сейсморазведки либо неуверенное выделение границ пласта по каротажным кривым. С применением алгоритма кросс-валидации процесс самоэкспертирования становится быстрее. Разработанный инструмент является лишь идентификатором проблемных значений параметров, пересмотр которых может существенно повлиять на целостность модели.

Кросс-валидация позволяет опеределить «вес» каждой скважины. В процессе кросс-валидационного анализа можно встретить такую зависимость, при которой вся модель «держится» на одной скважине, ее отключение в несколько раз снижают параметры регрессии.

Метод кросс-валидации тестировался как на верхнеюрских пластах, так и на нижнемеловых отложениях месторождений «Мегионнефтегаза». В настоящее время работа по построению карт с применением вышеописанного подхода ведется для наклонно направленных границ (ачимовские отложения), где корреляция по каротажным диаграммам и данным сейсморазведки не всегда однозначна.

Список литературы

1. *Allen D.M.* The relationship between variable selection and data augmentation and a method for prediction // Technometrics. - 1974. – V. 16. – P. 125–127.
2. *Geisser S.* The predictive sample reuse method with applications // Journal of the American Statistical Association. - 1975. – V. 70. – P. 320–328.
3. *Stone M.* Cross-validatory choice and assessment of statistical predictions // Journal of the Royal Statistical Society. – Ser. B. - 1974. – V. 36. – P. 111–147.
4. *Shao J.* Linear model selection by cross-validation // Journal of the American Statistical Association. – 1993. – V. 88(422). – P. 486–494.
5. *Arlot S., Celisse A.* A survey of cross-validation procedures for model selection // Statistics Surveys. – 2010. - V. 4. - DOI:10.1214/09-SS054.

References

1. Allen D.M., *The relationship between variable selection and data augmentation and a method for prediction*, Technometrics, 1974, V. 16, pp. 125–127.
2. Geisser S., *The predictive sample reuse method with applications*, J. Amer. Statist. Assoc., 1975, V. 70, pp. 320–328.
3. Stone M., *Cross-validatory choice and assessment of statistical predictions*, J. Roy. Statist. Soc. Ser. B, 1974, V. 36, pp. 111–147.
4. Shao J., *Linear model selection by cross-validation*, J. Amer. Statist. Assoc., 1993, V. 88(422), pp. 486–494.
5. Arlot S., Celisse A., *A survey of cross-validation procedures for model selection*, Statistics Surveys, 2010, V. 4, DOI:10.1214/09-SS054.