title: "Восстановление пропусков в методе гамма каротажа"

author: "Churikov Nikita"

Abstract

Метод гамма каротажа (GR) является одним из обязательных методов каротажа при проведении геофизического исследования скважин. Однако, во время проведения замеров, датчик, по той или иной причине, может перестать записывать наблюдения. По этой причине, необходимо будет производить измерения в скважине заново, что ведет к дополнительным денежным затратам. Мы предлагаем альтернативу, в виде заполнения таких пропусков при помощи методов машинного обучения для достаточно большого количества пропусков: производится обучение модели на присутствующих данных в скважине, и по ним, модель пытается синтезировать пропущенные участки скважины.

Introduction

Геофизическое исследование скважин является важной частью поиска полезных ископаемых в земле. Процесс исследования скважин является долгим, трудоемким и требует серьезных денежных вложений. По этой причине, когда происходит сбой в записи информации о скважине, исследователи, как правило, стараются придумать способ избежать повторных замеров. Типовым вариантом является линейная интерполяция для заполнения пробелов небольшого размера (до 5 точек), а для пропусков большего размера требуются более сложные методы или производят повторный замер данных, что ведет к дополнительным расходам.

Также геофизик, при интерпретации кривой, мог сам в нечаянно добавить пропуски небольшого размера. Визуально, такие пропуски будут незаметны, но численно -- это будут пропущенные значения.

В конце концов, при автоматической обработке данных и применении временных рядов, необходимо придумывать способы обработки пропусков, поскольку алгоритмы не способны с ними работать. Самый простой способ -- отбросить все значения, которые являются пропусками, однако это довольно грубый способ, и мы теряем полезную информацию, используя его. Потому, существуют различные способы заполнения пропущенных значений в данных [1]. Но в большинстве своем, они рассчитаны на то,

~~чтобы не привлекать дополнительных данных~~ [2].

Постановка задачи

Есть дана некоторая скважина, в которой были использованы некоторые методы $g={\gamma_1 \dots \gamma_n} \in \Gamma$, где Γ -- все возможные методы каротажа. При этом в некотором методе каротажа γ_j пропущено некоторое количество значений, количество которых ≤ length(well)/2, а в остальных методах γ_j методы каротажа присутствуют без пропусков, и обозначим их как **X**, а метод, имеющий пропущенные значения как *y*

В таком случае, данная задача сводится к задаче обучения с учителем [3], где *y* -- целевая переменная, а **X** -- наблюдаемые значения.

Данные

В целях исследования, были использованы данные, которые В датасете присутствует 53 скважины. В этой работе были рассмотрены замеры по 5 скважинам с глбиной больше 12 футов. Они имеют префикс **TRADING BAY UNIT D-** и они имеют следующие названия: **TRADING BAY UNIT D-05, D-06, D-09, D-12, D-13**.

Пропущенные значения в данных являются нормальным явлением. Даже в данных, которые лежат в открытом доступе они есть. Во всем используемом датасете имеется ~800 пропусков до 6 значений.

В общем случае существует достаточно большое количество методов каротажа [5] и для конкретных задач связанных с геофизикой необходимо использовать свои методы. В рассматриваемых данных присутствуют следующие методы: GR, RHOB, SP, ILD, DT

В качестве целевой переменной *y* будем рассматривать **GR**, а в качестве наблюдаемых значений **X** все остальные методы каротажа.

На Рис. 2 видно, что большинство каротажных кривых подобны нормальному распределению, однако ILD имеет тяжелый правый хвост, а SP имеет значения слева и справа относительно нуля и несимметрична.

RHOB также имеет одобие на тяжелый левый хвост, однако там достаточно много значений, потому нельзя эту часть назвать хвостом. скорее это поведение кривой в целом.

Также наша целевая переменная -- GR подобна нормальному распределению, а потому можно предположить, что методы регрессии будут его хорошо восстанавливать.

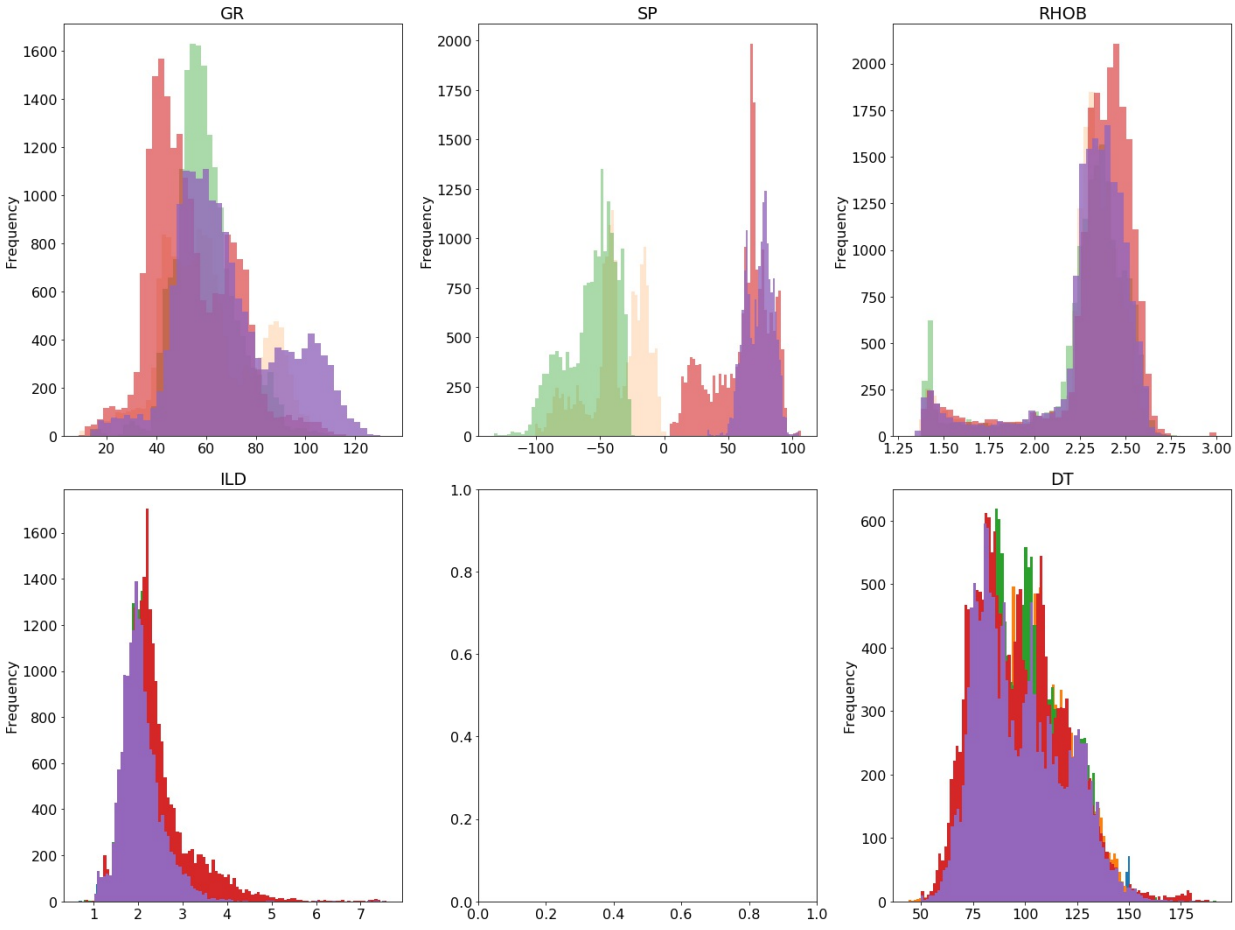


Рис.2 Распределения значений ~~без предобработки~~

Методология

Представленные данные имеют очень разный масштаб. Так, например, метод SP имеет значения от -100 до 100, а DT в промежутке от 1.25 до 3. Также данные являются шумными, поскольку они получены с датчиков. Шум и масштаб данных подводят нас к тому, что данные необходимо некоторым образом подготовить, чтобы они были полезны.

Предобработка

Для предобработки были использованы следующие методы:

1. Standart Scaler ко всем данным;

    i. Поскольку мы передаем данные нейронной сети, то необходимо, чтобы данные были в одном масштабе. Лучшим стандартизатором показал себя Standart Scaler [можно ссылку на sklearn]
2. \log(X^{ILD} + 1) к методу каротажа ILD
    i. Такое преобразование является типичной практикой для избавления от тяжелых хвостов.
3. (Применяется к SP) Прямое преобразование фурье -> Отбрасывание медленных трендов (Мы брали частоту 5) -> Обратное преобразование фурье.
    i. Этот момент является чуть более интересным. Идея в том, что SP с ростом глубины начинает смещаться вниз. При этом происходит только смещение. Для того, чтобы избавиться от этого тренда, можно применить преобразование преобразование фурье, а затем обратное. В результате, значения сместятся к "более менее" общему среднему.

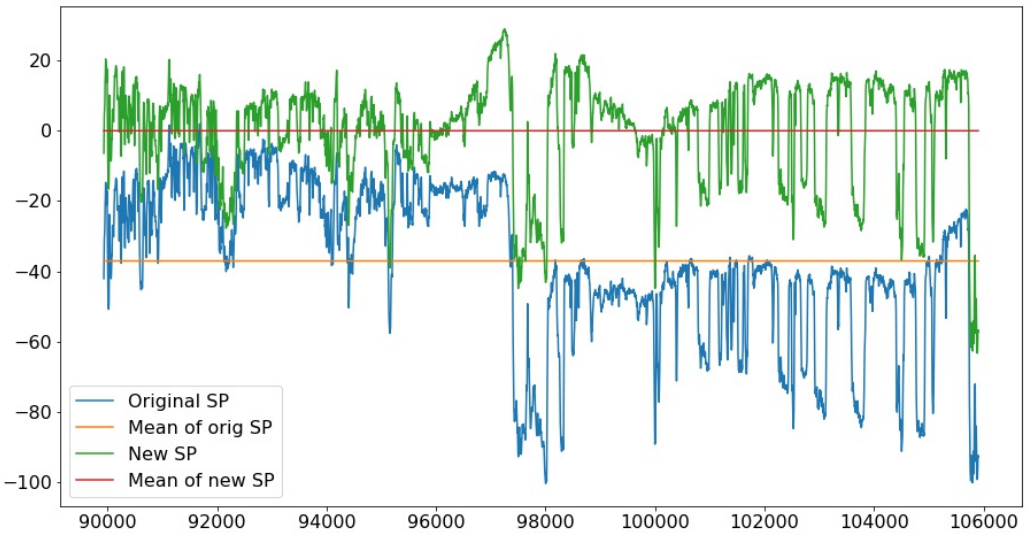


Рис. 3 До и после преобразования Фурье. По оси X -- Индексы значений (не играют роли). По оси Y -- значния SP

Генерация новых признаков

Поскольку признаков немного, то есть потребность в добавлении дополнительной информации. Существуют различные способы генерации новых признаков [6].

В нашем исследовании мы использовали несколько способов генерации новых фичей с помощью Quantile Smooth и Window generator:

Quantile Smooth -- идея метода в том, что мы берем квантили некоторого порядка не от

всего набора данных, а от скальзящих окон. В результате, мы получаем некоторый новый признак, который является квантилем порядка n от окна размером k признака \gamma_j. Такое преобразование можно воспринимать, как сглаживание данных. А каротажные кривые являются довольно шумными, потому локальные шумы могут быть проблемой.

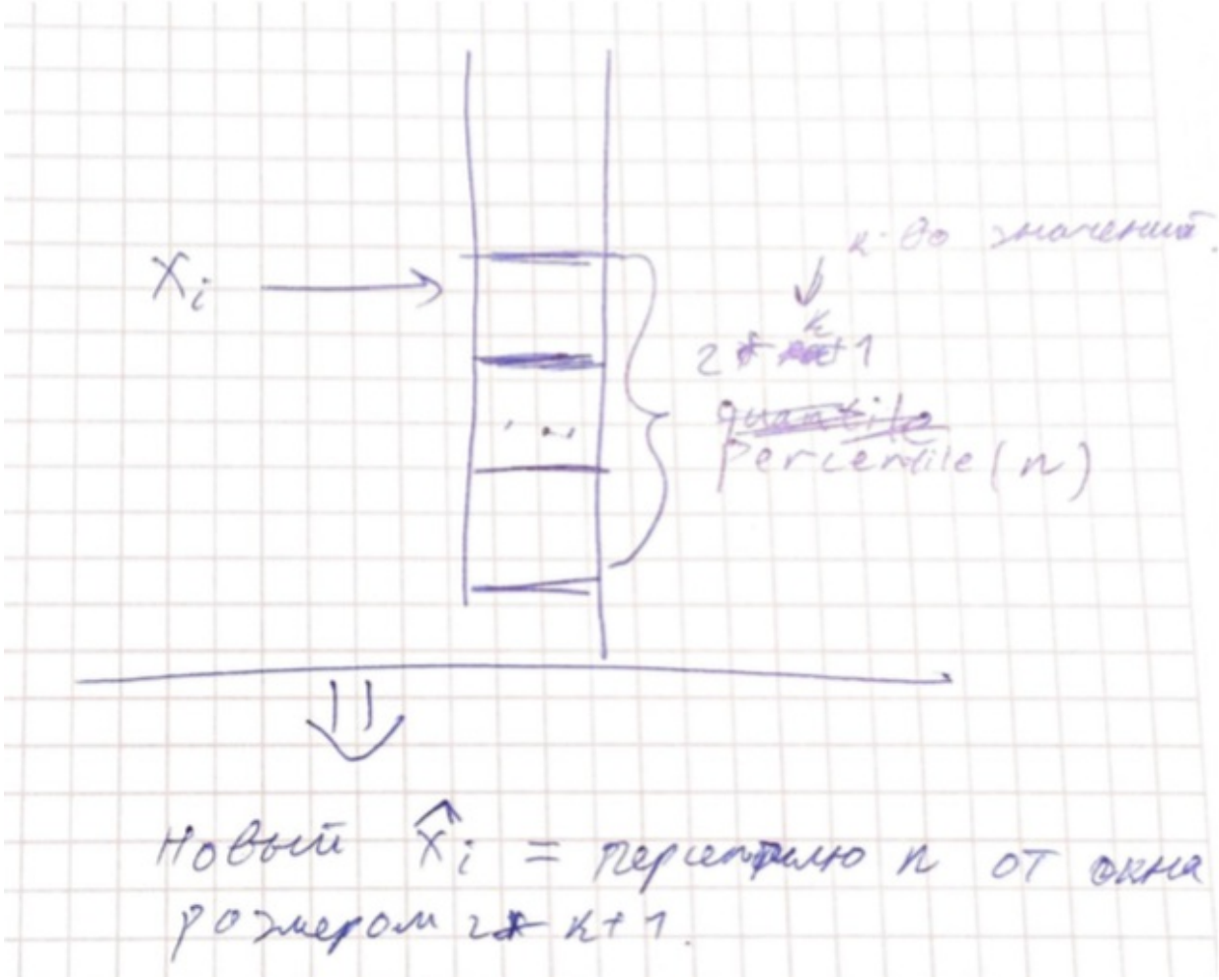


Рис. 4 Квантильное сглаживание

Window generator -- Также точка в каротажной кривой является своего рода конволюцией. Потому, чтобы развернуть (deconvolve) её, мы предлагаем взять n значений кривой до некоторой точки x_i и после и создать из них 2*n новых признаков. В результате в одном наблюдении будет содержаться информация о данных до наблюдения и после. Такое преобразование можно воспринимать как, своего рода, рекуррентный слой в нейронной сети.

Модели

В исследовании были использованы следующие модели:

- Линейная интерполяция [ссылки, наверно]
- Интерполяция третьего порядка [ссылки, наверно]
- Двухслойная нейронная сеть [ссылки]
- Двухслойная нейронная сеть со "сдвигом"

Линейная интерполяция

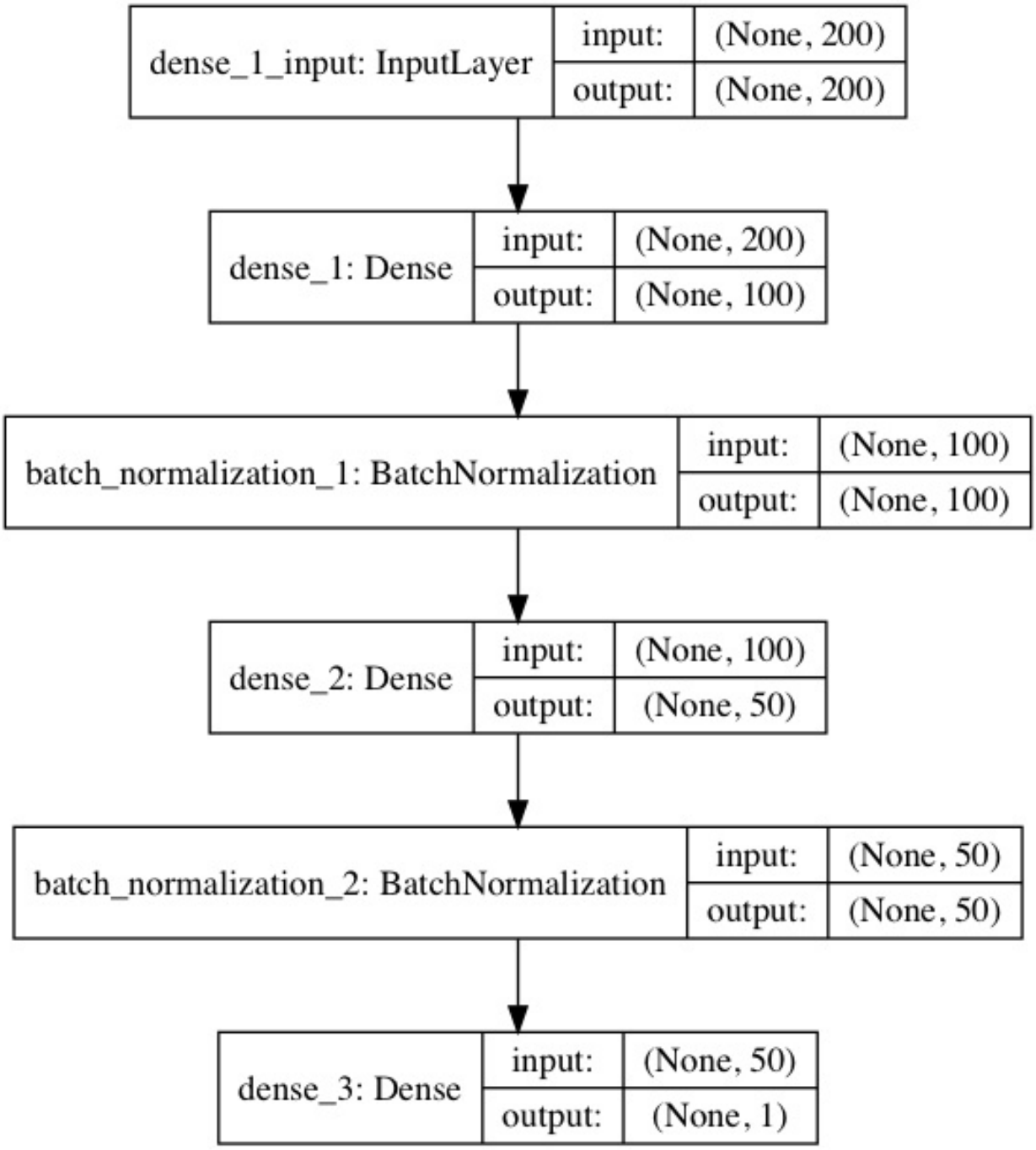
Данный метод является достаточно интуитивным при малых пропусках (до 5 пропущенных точек). Его огромное преимущество в его простоте: он не требует дополнительных данных, для работы ему требуется всего лишь начальная и конечная точка, чтобы провести прямую.

Интерполяция третьего порядка

Развитие идеи с интерполяцией. В теории, строится полином третьего порядка и получается некоторая, более сложная кривая. Но, как будет показано дальше, этот метод не работает в данной задаче.

Двухслойная нейронная сеть

Нами была использована нейронная сеть следующей архитектуры.



Была выбрана такая несложная архитектура с целью быстрого обучения на данных. На данный момент, сеть обучается на GPU Quadro 4000 за 1 минуту. В сумме с препроцессингом и предсказанием пропусков получается где-то 2 минуты на обучение и восстановление данных.

Двухслойная нейронная сеть со "сдвигом"

Специфика и преимущество нашей задачи в том, что мы можем по максимуму добавлять информацию из обучающей выборки. Часто бывают, случаи, когда модель понимает поведение кривой, но не попадает в начальную и конечную точки истинной

кривой. И потому её хотелось своего рода "передвинуть" в нужное место. Линейная интерполяция обладает информацией о том, где ей начинать и где заканчивать. Почему бы нейронной сети не дать эту же информацию?

В результате, объедине линейной интерполяции и нейронной сети привело к следующей идее:

- Обучить модель на присутствующих данных
- Предсказать, что скажет модель на batch-е обучающих данных.
- Взять разницу предсказанного обучения и истинных значений и посчитать среднее отклонение
- Прибавить его к предсказанным значениям.

Evaluation

Для оценки качества использовалась mean absolute error разделенная на разницу 99%-го и 1%-го перцентилей. Такая нормировка позволяет поместить ошибку в интервал от 0 до 1 (чем меньше, тем лучше).

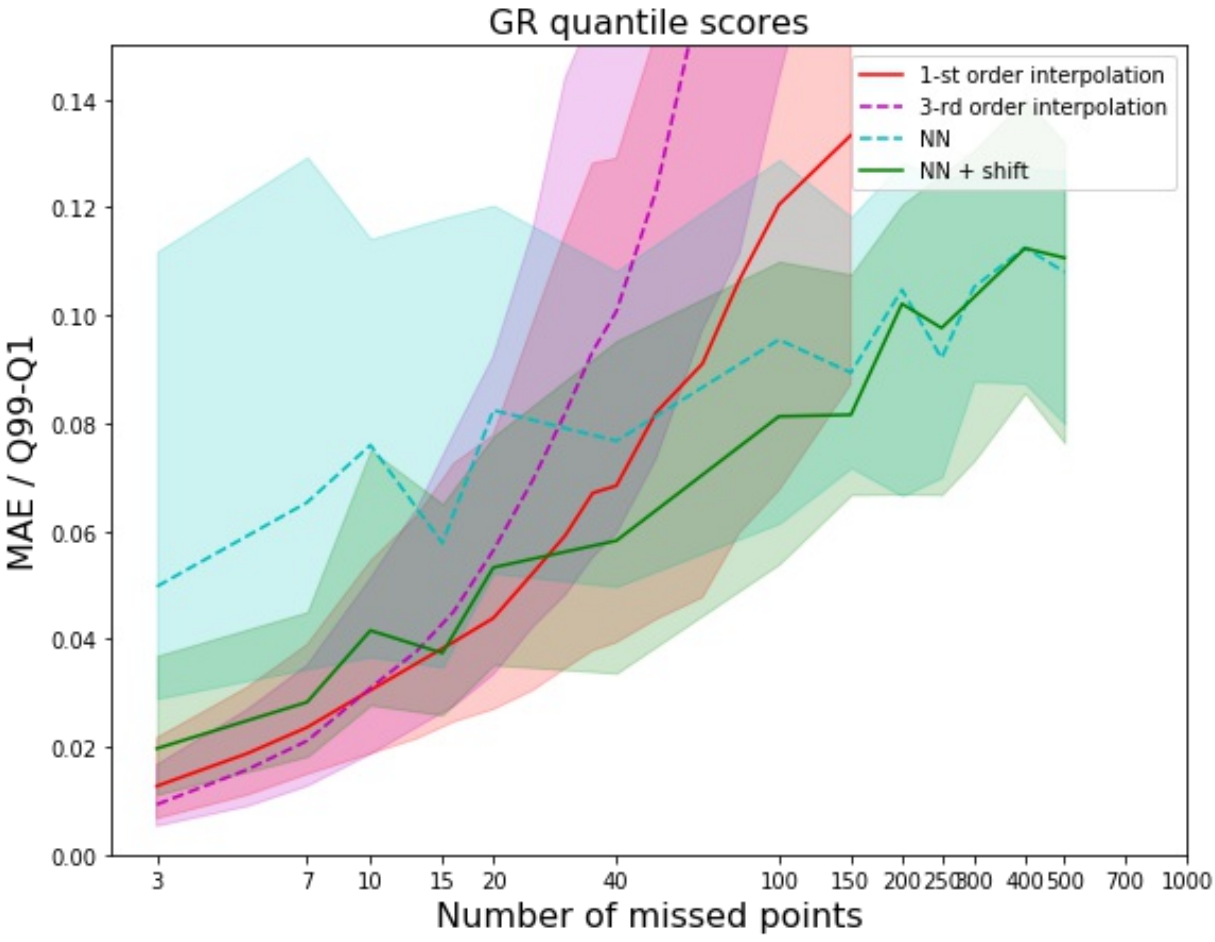


Рис. 5 Квантильное сглаживание

На Рис. 5 видно, что для достаточно небольшого размера пропусков, хорошо работает линейная интерполяция. Однако с увеличением размера, появляется потребность в использовании более умных средств.

Список литературы

[1] Working with missing data // pandas.pydata.org URL: https://pandas.pydata.org/pandas-docs/stable/missing_data.html (дата обращения: 17.04.2018).
[2] Caruso, C. and Quarta, F., 1998, 'Interpolation methods comparison', Comput. Math. Appl. 35, 109–126.
[3] Digital Well Log Files // URL: http://doa.alaska.gov/ogc/DigLog/diglogindex.html (дата обращения: 17.04.2018).
[4] Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning.
[5] Society of Professional Well Log Analysts (1975). Glossary of terms & expressions used in well logging. Houston, Texas: SPWLA. p. 74 p.
[6] Bestagini, P. A Machine Learning Approach to Facies Classifica- tion Using Well Logs / Paolo Bestagini, Vincenzo Lipari, Ste- fano Tubaro // SEG Technical Program Expandend Abstracts 2017. — 2017. — P. 2137–2142. — http://library.seg.org/doi/abs/10.1190/segam2017-17729805.1