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. ~~Однако для малого числа пропусков, будет показано, что достаточно использовать линейную интерполяцию.~~

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. Возможным вариантом может быть линейная интерполяция для заполнения пробелов небольшого размера (до 5 точек), а для больших кусков данных -- перенос кусков кривых из других скважин или повторный замер.

Также геофизик, при интерпретации кривой, мог сам в .las добавить пропуски небольшого размера, которые, при просмотре всей кривой практически не заметны, но в las файлах такие участки будут отмечены как NaN-ы.

В конце концов, при применении методов машинного обучения, необходимо придумывать способы обработки пропусков, поскольку алгоритмы не способны с ними работать. Самый простой способ -- отбросить все значения, которые являются NaN-ами, однако это довольно грубый способ, и мы теряем полезную информацию, используя его. Потому, существуют различные способы заполнения пропущенных

значений в данных [https://pandas.pydata.org/pandas-docs/stable/missing_data.html]. Но в большинстве своем, они рассчитаны на то, чтобы не привлекать дополнительных данных.

~~Специфика каротажных данных в том, что определенные комбинации методов используются для поиска определенных полезных ископаемых и исследования состояния скважины, потому невозможно сказать заранее, какие методы будут использованы, поскольку они зависят как от выбора эксперта, так и заказчика [мб тут какая-то ссылка].~~

Постановка задачи

Пусть дана некоторая скважина, в которой были использованы некоторые методы $g={\gamma_1 \dots \gamma_n} \in \Gamma$, где Γ -- всевозможные 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 [ссылка], где *y* -- целевая переменная, а **X** -- наблюдаемые значения.

Данные

В целях исследования, были использованы данные, которые **ВЫКЛАДЫВАЕТ КТО-ТО В АЛЯСКЕ, НАДО ПОПРАВИТЬ**[http://doa.alaska.gov/ogc/DigLog/diglogindex.html#]. В датасете присутствует 53 скважины. В этой работе был рассмотрен один куст с префиксом **TRADING BAY UNIT D-**, поскольку в нем присутствуют глубокие скважины (~12 км) с дельтой замеров в 50 сантиметров. В результате мы будем использовать в данной работе 5 скважин: **TRADING BAY UNIT D-05, D-06, D-09, D-12, D-13**.

Для понимания того, насколько необходимо восстанавливать данные, посмотрим, какого размера пропуски встречаются в датасете. На Рис. 1 мы приводим распределение реальных пропусков в данных.

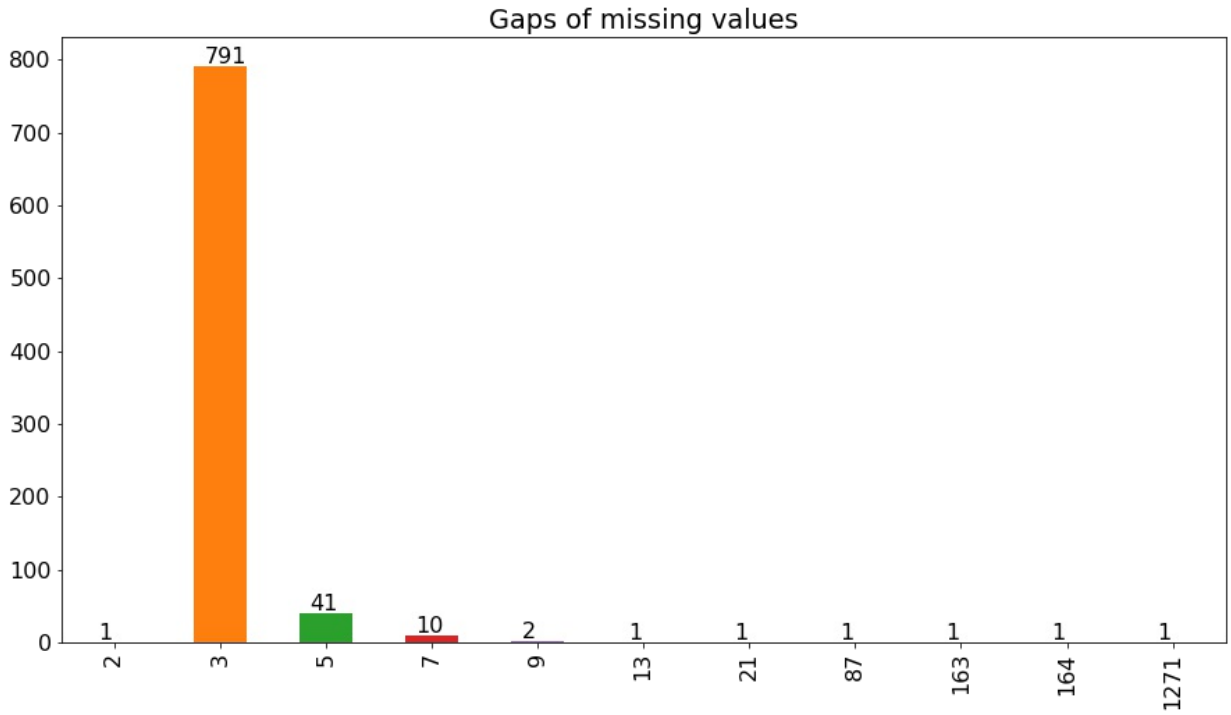

Рис. 1 Распределение размеров пропусков в данных

На Рис. 1 видно, что в большинстве случаев присутствуют пропуски от 2 до 5 значений, однако бывают случаи пропуска 164 и даже больше 1000 значений. Мы симулируем пропуски до 200 значений, поскольку в случае 1000 пропущенных значений имеет смысл повторить измерения скважины.

~~Методы каротажа присутствующие в данных~~

В общем случае существует достаточно большое количество методов каротажа [ссылка на википедию] и для конкретных задач связанных с геофизикой необходимо использовать свои методы. В рассматриваемых данных были использованы следующие методы:

- GR -- гамма-каротаж. Очень простой и распространённый метод, измеряющий только естественное гамма-излучение от пород, окружающих скважину. Существует его чуть более усложнённый вариант — спектрометрический гамма-каротаж (СГК или ГК-С), который позволяет различить попавшие в детектор геофизического зонда гамма-кванты по их энергии. По этому параметру можно точнее судить о характере слагающих толщу пород.
- RHOB -- гамма-гамма каротаж. Геофизический зонд облучает породу гамма-излучением, в результате которого порода становится радиоактивной и в ответ тоже излучает гамма-кванты. Именно эти кванты и регистрируются зондом.
- SP [~~https://en.wikipedia.org/wiki/Spontaneous_potential_logging~~]~~--~~

- ILD [http://petrowiki.org/Induction_logging] --
- DT [https://en.wikipedia.org/wiki/Sonic_logging] --

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

~~Анализ данных~~

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.

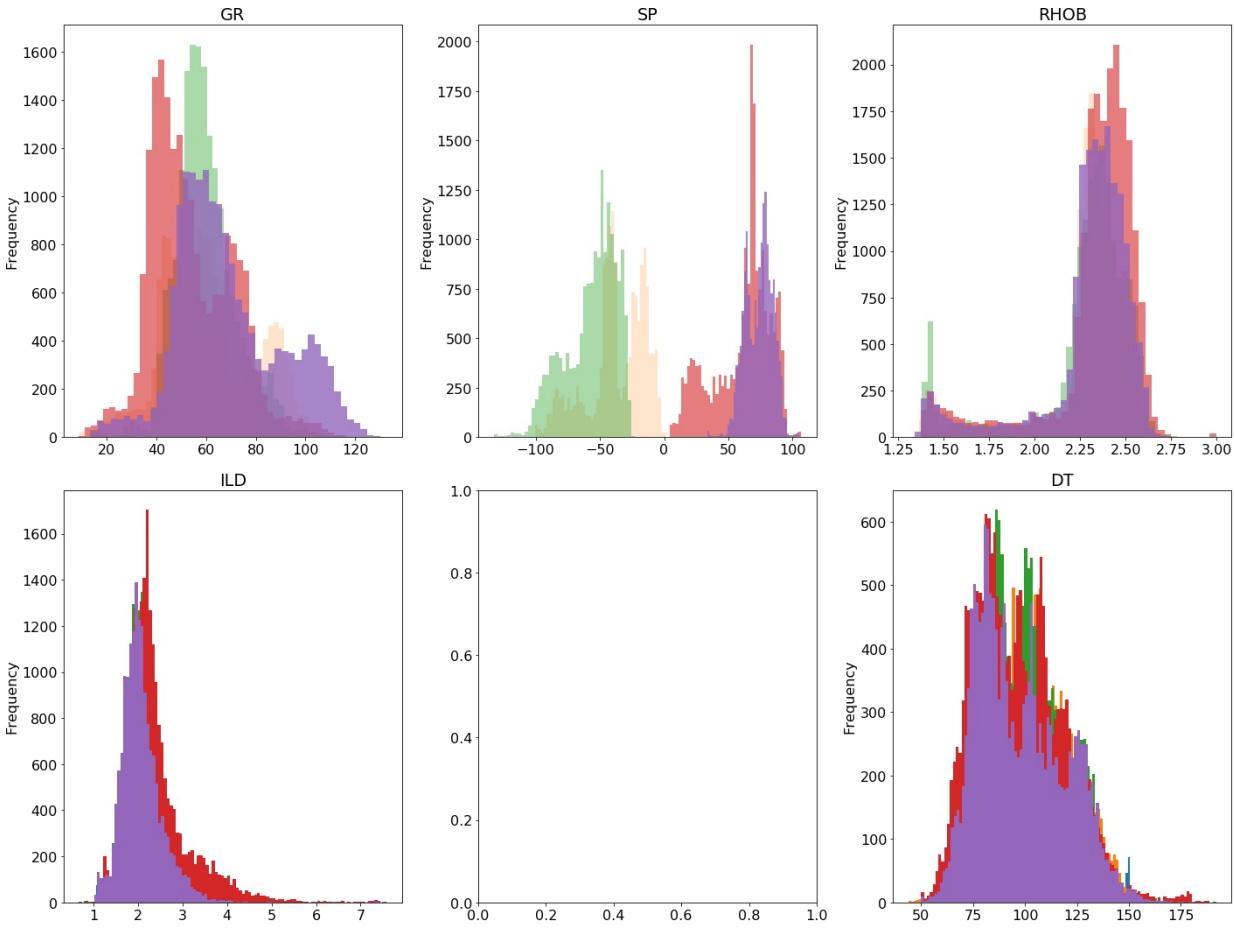

Рис.2 Распределения значений без предобработки

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.

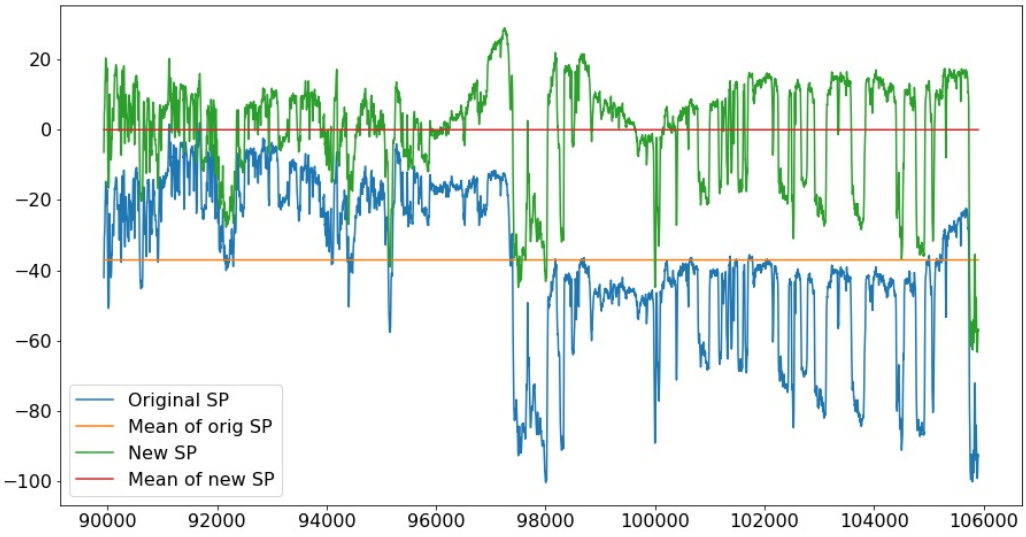

Рис. 3 До и после преобразования Фурье

Генерация новых признаков

Поскольку признаков немного, то возможность и есть потребность в добавлении дополнительной информации. Существуют различные способы генерации новых

признаков [можно сослаться на статью с контестом и sklearn с полиномиальными фичами].

В нашем исследовании мы использовали несколько способов генерации новых фичей:

Quantile Smooth -- идея метода в том, что мы берем квантили некоторого порядка не от всего набора данных, а от некоторых "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.

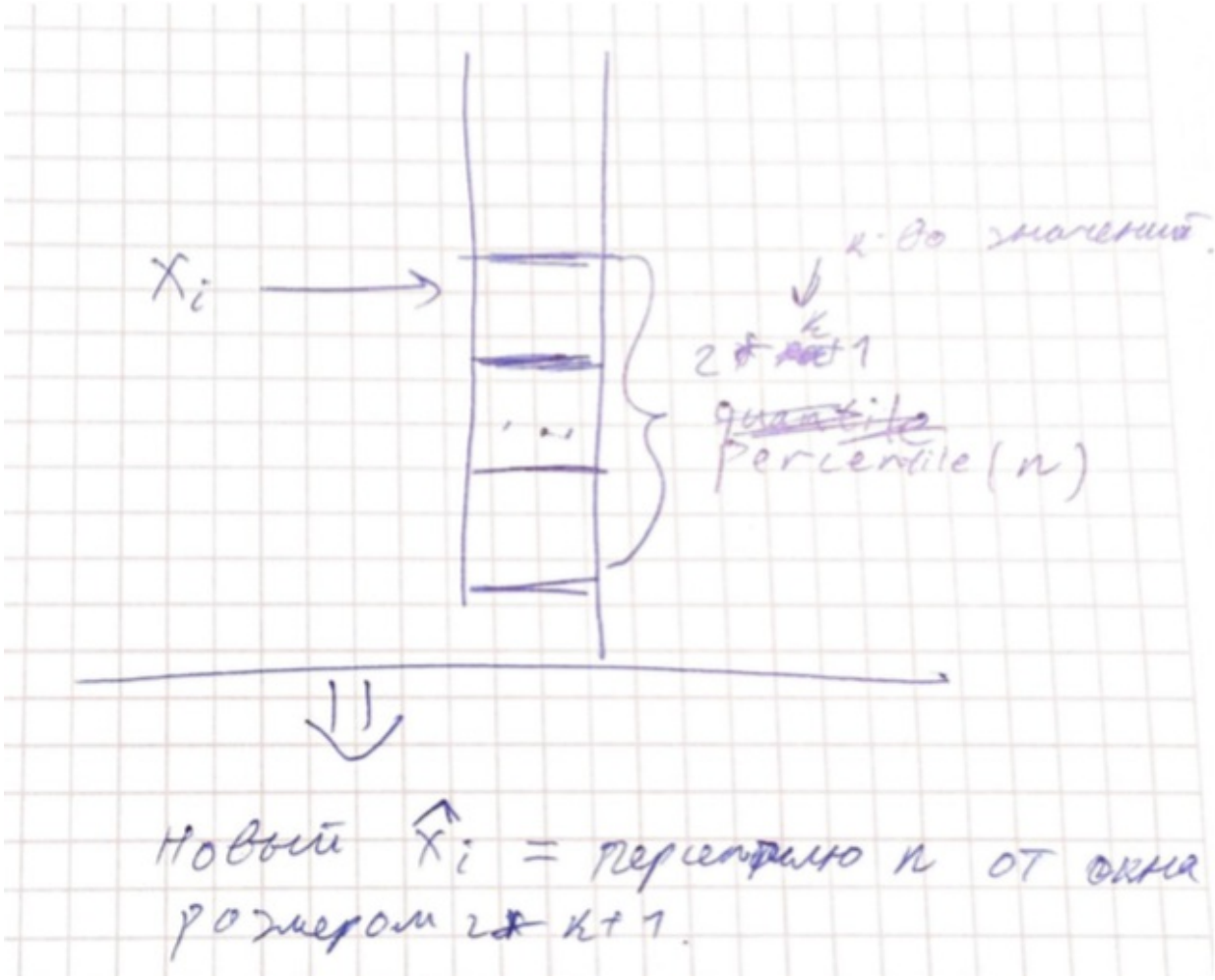

Рис. 4 Квантильное сглаживание

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.

Layer (type)	Output Shape	N_Params
dense_1 (Dense)	(None, 100)	17100
batch_normalization_1 (Batch	(None, 100)	400
dense_2 (Dense)	(None, 50)	5050
batch_normalization_2 (Batch	(None, 50)	200
dense_3 (Dense)	(None, 1)	51

Total params: 22,801

Trainable params: 22,501
Non-trainable params: 300

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 будем использовать mean absolute error разделенную на разницу 99%-го и 1%-го перцентилей. Такая нормировка позволяет поместить ошибку в интервал от 0 до 1 (чем меньше, тем лучше).

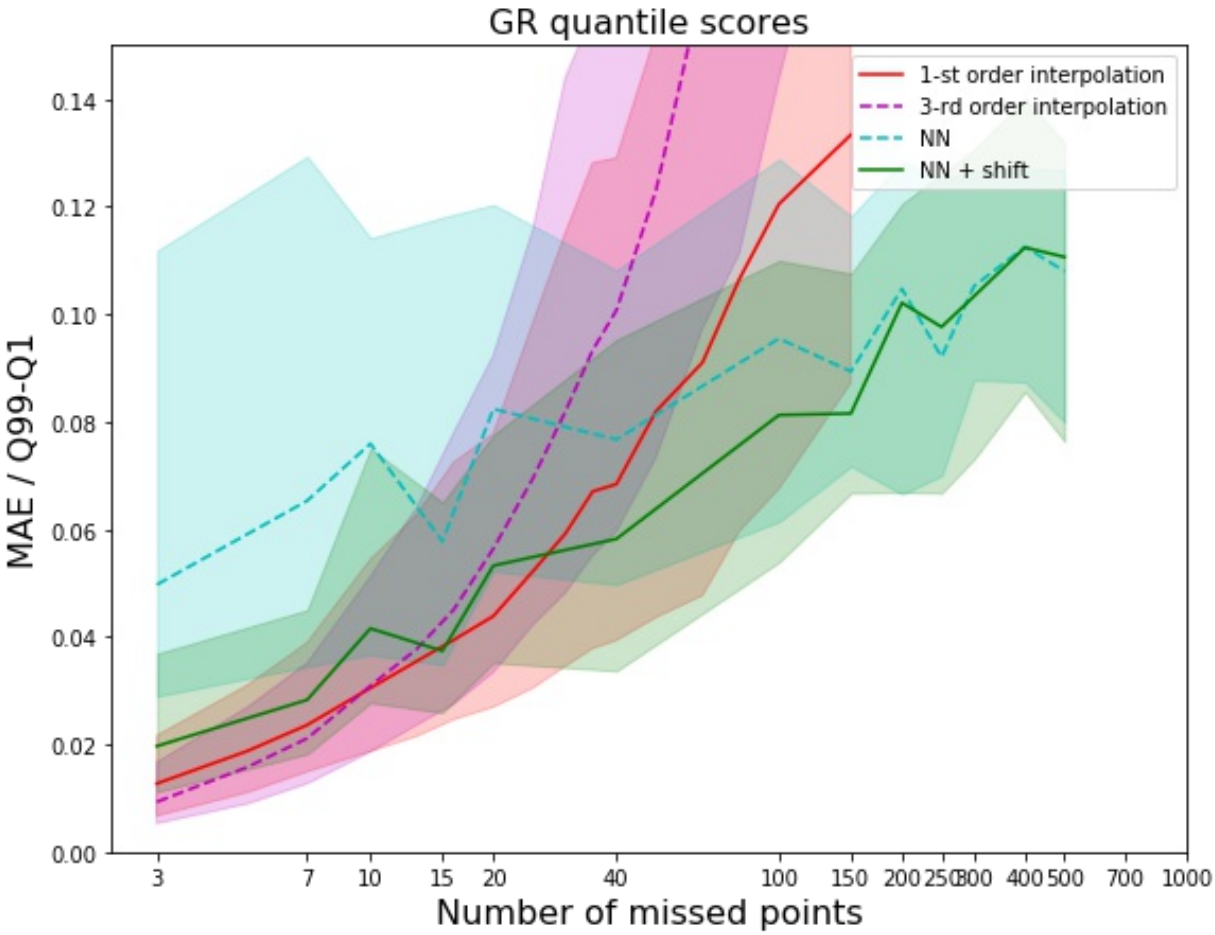

Рис. 5 Квантильное сглаживание

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.