Slightly paramilitary задача предсказания временных рядов

Артём В. Васильев*, ИТМФ МГУ

29.02.2020

1 Постановка задачи

- ... "Что окончательно отобрать студентов в научную группу предлагаем вам решить такую задачу. Описание исходных данных: временной ряд длиной 1024, содержащий значения мощности передающей аппаратуры (вещественные значения). Цель: провести статистический анализ, определить закономерности, выполнить прогноз целевой переменной на 20 отсчетов вперед. В качестве ответа необходимо предоставить последовательность из 20 вещественных значений мощности в следующие 20 отсчетов. Ограничений на использование средств программирования, визуализации, библиотек нет. Ждем ваши решения к вечеру среды 4 марта" ....

Из предложенного текста задания не ясно какая метрика для оценки качества прогноза будет использоваться. Доопределим задачу, считая что прогноз должен будет минимизировать MSE (как будет показано ниже, на самом деле это не имеет значения).

2 Предварительный анализ

Будем обозначать W_n значение временного ряда в момент отсчета t_n. Начнём анализ преложенного временного ряда с наивного визуального осмотра.

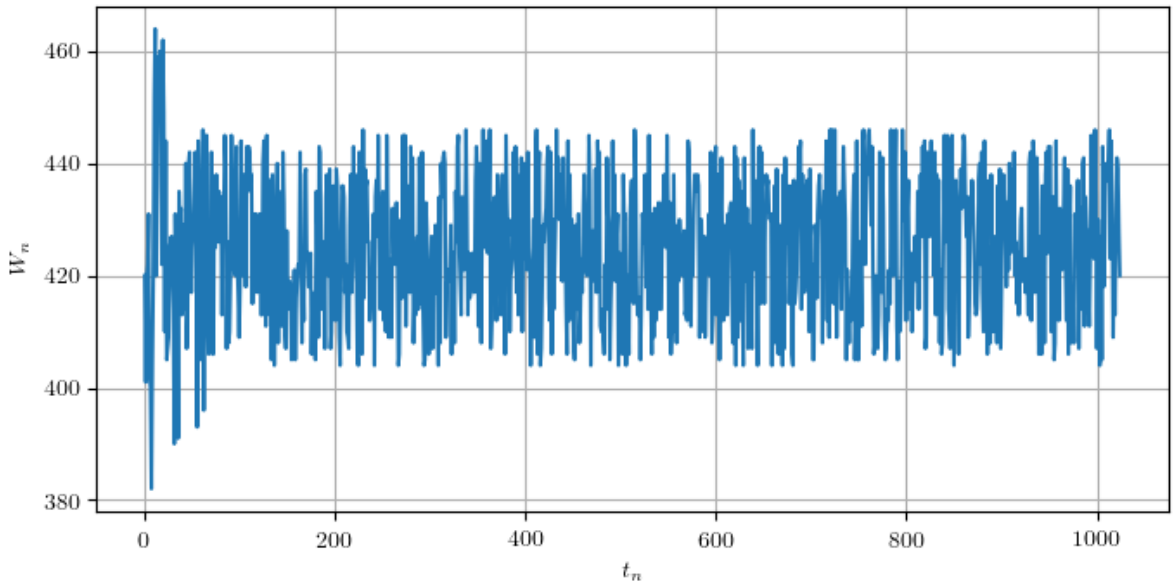

Рис. 1: График временного ряда W_n во всех предоставленных для анализа точках.

Не считая начальной области в районе $[0, \dots, 65]$ отсчетов, ряд W_n колеблется гомоскедастичностно вокруг своего среднего значения $\overline{W} = 424.9$. Какой-либо очевидный тренд не наблюдается. На масштабах $100 - 200$ можно наблюдать похожее на периодическое изменение сигнала.

В ходе первой встречи с представителями института в ответ на мой вопрос о характере данных с которыми предстоит работать, было рассказано о временных рядах данных с разнообразных датчиков с равномерным шагом 20 секунд. Если предположить что предоставленные для данной работы данные тоже соответствуют временному ряду с таким шагом, то не будет оснований ожидать наличия в сигнале сезонных, месячных или даже суточных значащих изменений.

*vasiliev.av15@physics.msu.ru

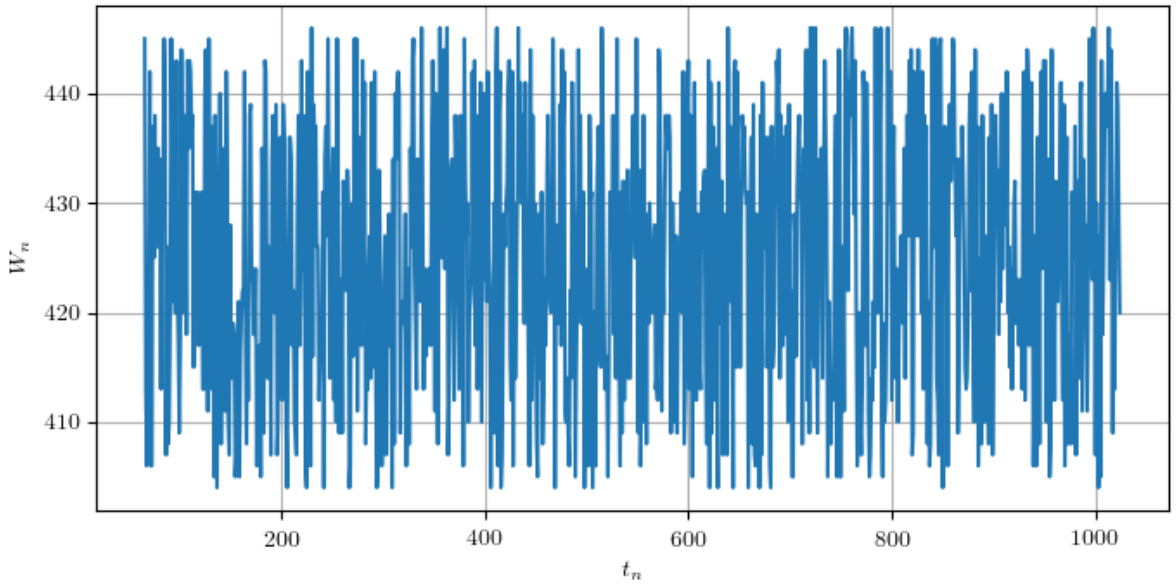

Рис. 2: График отрезка временного ряда W_n после исключения начальной области.

Поскольку нет никакой дополнительной достоверной информации о причинах отклонения значений ряда в области $[0, \ldots, 65]$ отсчетов и для решения задачи прогноза предполагается использование исключительно статистических методов, данная область будет объявлена незначащим единичным отклонением в далеком прошлом и во всем последующем анализе будет игнорироваться.

3 Построение автокорреляционных функций.

Построим несколько первых значений автокорреляционной функции ряда W_n:

$$r_\tau = \frac{\sum_{n=1}^{N-\tau} \left(W_n - \overline{W}\right)\left(W_{n+\tau} - \overline{W}\right)}{\sum_{n=1}^{N} \left(W_n - \overline{W}\right)^2}. \tag{1}$$

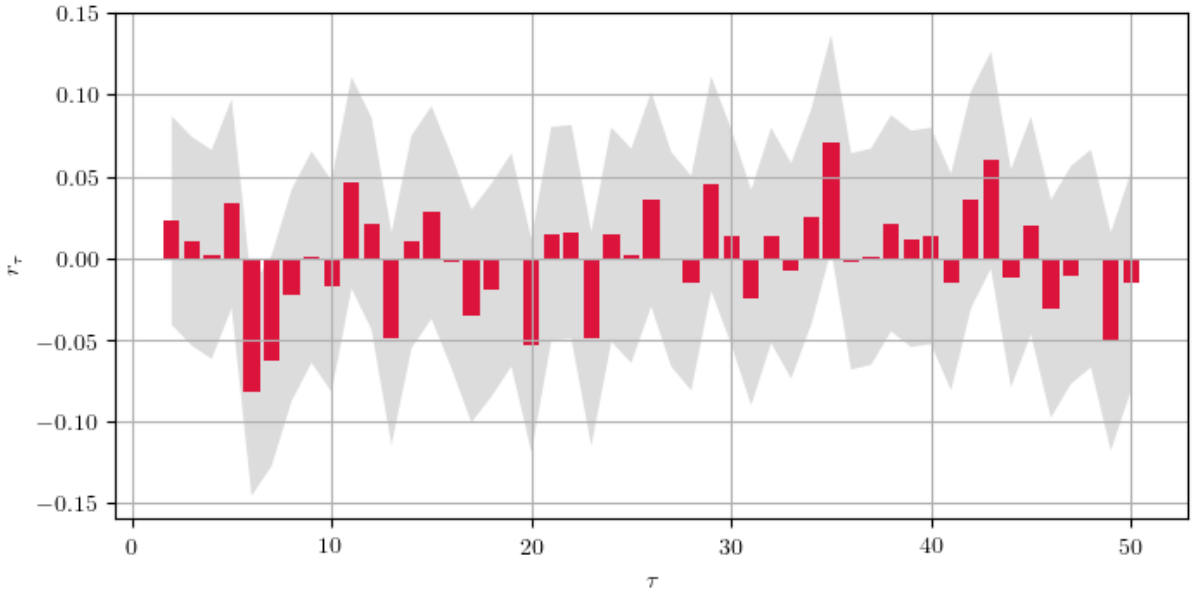

Рис. 3: Значения автокорреляционной функции с лагом $\tau = \overline{1, 50}$. Серым цветом указан доверительный интервал с уровнем доверия 0.95

Все полученные значения автокоррекции значимо неотличимы от нуля. Полученный результат даёт основания полагать, что исследуемый отрезок ряда просто представляет собой шумовой сигнал с ненулевым средним.

Для проверки предположения о том что сигнал является белым шумом используем стандартный статистический критерий Q-теста Льюнга — Бокса[1], который позволяет рассматривая совокупность последовательностей значений корреляционной функции выбрать одну из двух конкурирующих гипотез: сигнал является белым шумом, сигнал является не случайным.

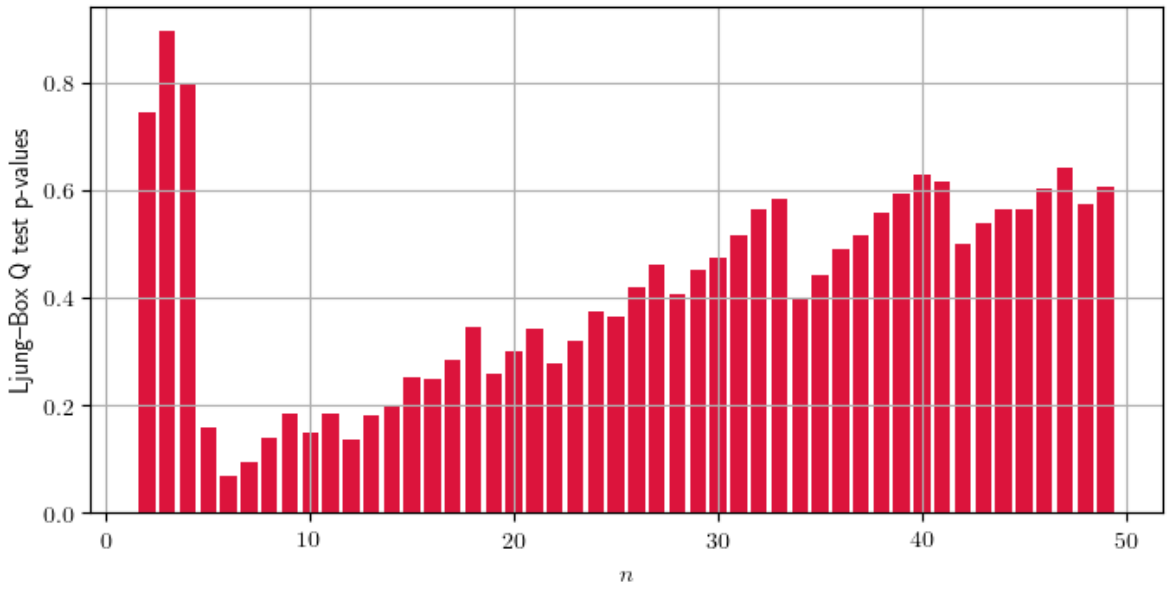

Рис. 4: P-уровень значимости для Q-теста Льюнга — Бокса порядка $n = \overline{2, 50}$.

Полученные p-значения Q-теста Льюнга — Бокса значимо отличаются от нуля и потому гипотеза о неслучайности сигнала должна быть отвергнута. На масштабе 50-ти отсчетов сигнал представляет собой просто шум.

Для исследования автокоррекцией более высоких порядков длина датасета может оказаться недостаточной. Тем не менее, попробуем посчитать значения автокорреляционной функции и p-значения Q-теста Льюнга — Бокса вплоть до 200-го лага (порядка).

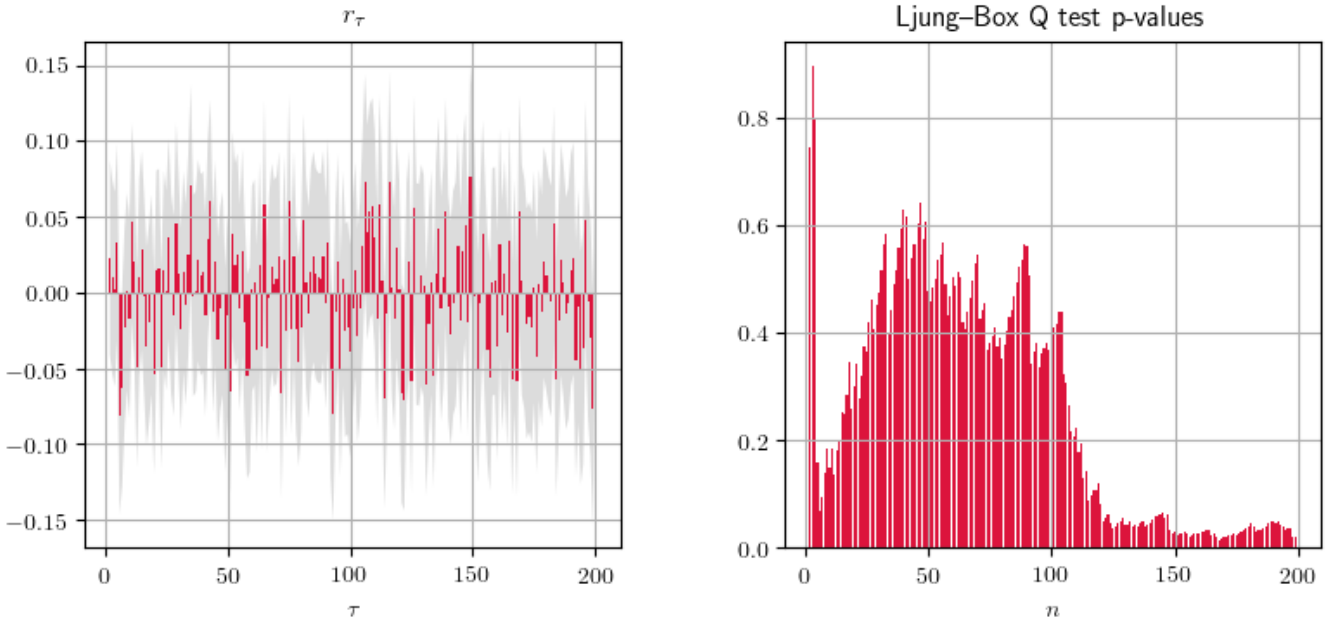

Рис. 5: Автокорреляции высоких порядков.

Не сложно заметить, что на масштабе 200 отсчетов сигнал также не является скорректированным. Чтобы окончательно убедиться в том что предлагаемый для исследования сигнал является шумом, построим оценку для спектральной плотности мощности сигнала, сделав БПФ с оконной функций Хэмминга.

Получаем спектр белого шума. Строго говоря, на этом этапе анализ сигнала нужно завершить. Как и решение задачи в целом.

[1] arXiv:1312.2240 [math.ST]

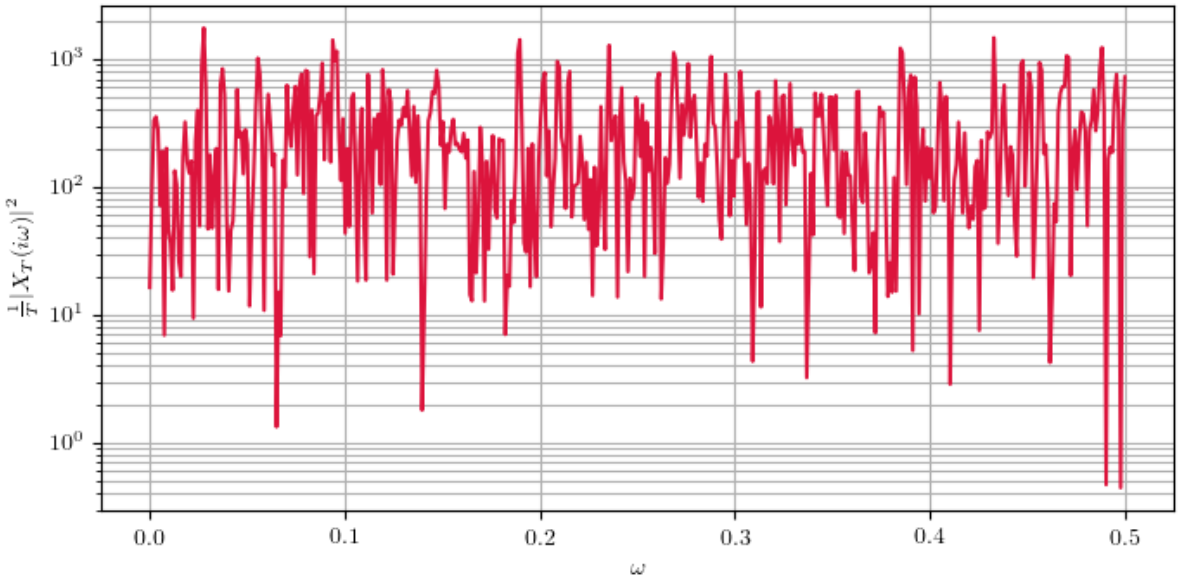

Рис. 6: Оценка для спектральной плотности мощности сигнала (периодограмма).

4 Построение предсказания

Первоначальная формулировка задачи могла давать основания полагать, что для построения прогноза можно попробовать воспользоваться одной из моделей класса авторегрессия — скользящеее среднеее (ARMA, ARIMA, SARIMAX и другие), некоторыми более продвинутыми линейными методами или нелинейными, такими как одномерные свёрточные сети. К сожалению, предложенный датасет не содержит какой-либо содержательной информации (за исключением среднего значения за всё время).

Для предсказания сигнала, являющегося суммой константы и шума с нулевым средним, достаточно просто дать оценку средним значением ряда за известное прошлое. Такой же результат даст и вырожденная ARMA(0,0) модель.

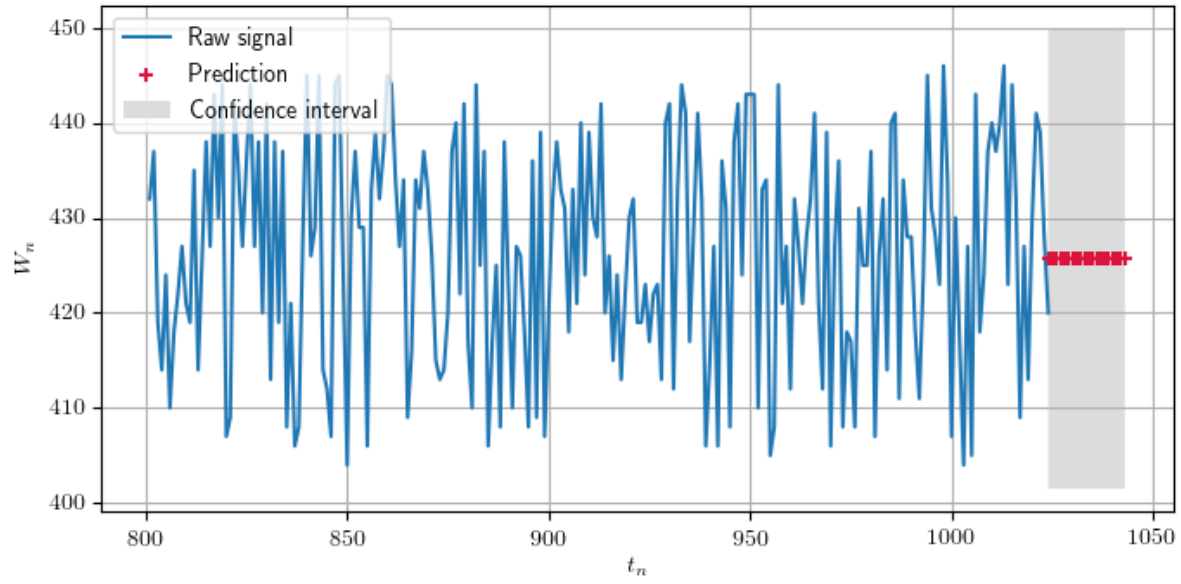

Рис. 7: Предсказание средним значением.

5 Оптимизация гиперпараметров модели

Не требуется. Модель нужно просто научить на среднее значение ряда за предыдущее время.

6 Ответ

timestamp	value	Confidence interval bounds
1024	425.7300	(401.5182, 449.9596)
1025	425.7300	(401.5190, 449.9606)
1026	425.7400	(401.5198, 449.9616)
1027	425.7400	(401.5206, 449.9627)
1028	425.7400	(401.5215, 449.9637)
1029	425.7400	(401.5223, 449.9648)
1030	425.7400	(401.5231, 449.9658)
1031	425.7400	(401.5239, 449.9669)
1032	425.7400	(401.5247, 449.9679)
1033	425.7400	(401.5256, 449.9690)
1034	425.7400	(401.5264, 449.9700)
1035	425.7400	(401.5272, 449.9710)
1036	425.7500	(401.5280, 449.9721)
1037	425.7500	(401.5288, 449.9731)
1038	425.7500	(401.5296, 449.9742)
1039	425.7500	(401.5305, 449.9752)
1040	425.7500	(401.5313, 449.9763)
1041	425.7500	(401.5321, 449.9773)
1042	425.7500	(401.5329, 449.9783)

Страница с проектом на github.