Сети глубокого обучения для построения виртуальных датчиков технологических процессов нефтепереработки

АСПИРАНТ 2 Г.О. ЛАЗУХИН И.С.,
К.Ф-М.Н., ДОЦЕНТ ПЕТРОВСКИЙ М.И.,
Д.Ф-М.Н., ПРОФЕССОР МАШЕЧКИН И.В.

Актуальность. Цифровые двойники

Современный производственный комплекс:

- Один или несколько технологических процессов (синтез выходов-продуктов)
- Повсеместный мониторинг и запись показаний физических датчиков системы
- Лабораторный анализ показателей качества продуктов
- Управление системой на основе дифференциальных линейных моделей

«Цифровой двойник»:

- Производственный процесс как *временной ряд*
- Точные *прогнозные модели* производственных процессов;
- Оптимальное управление на основе моделей
- Визуализация зависимостей, прогнозов, результатов управления

Лабораторные данные

Физические данные – показания датчиков системы, получаемые в режиме онлайн

Лабораторные данные – показания лаборатории, не могут быть получены физическим датчиком:

- Не могут быть получены точно в каждый заданный момент времени, *разреженые данные*. Типичный период сбора показаний – раз в сутки
- Критически важная составляющая процесса – *показатели качества* продукции
- Существующие интерполяции, используемые в производстве: *линейные и кусочно-постоянные аппроксимации*.

Виртуальные анализаторы (англ. *Soft Sensors*) – дополненные в каждый момент времени, тем или иным способом, лабораторные данные.

Пример: октановое число продукта на выходе номер 1.

Данные. Виртуальные анализаторы

Размерность набора данных	8569 на 322
MV переменных	5
LAB переменных	33
Период показаний	01.10.2022-23.09.2023
Частота показаний	1 час

Целевая переменная	
Объем показаний	326
Периодичность	24 часа
Воспроизводимость	3.5 градуса

Набор данных[1], предоставленный для построения виртуальных анализаторов:

- Сильная *коррелированность* как физических так и лабораторных показаний
- Критически *небольшой объем показаний* целевой переменной
- *Воспроизводимость показаний* – средняя «погрешность» лабораторных исследований

Авторами исследовалась *задача получения химических показателей производственного процесса в режиме онлайн на основе значений физических датчиков*

[1]A data set provided to the Moscow State University by the PJSC Lukoil Oil Company as a part of the corresponding research contract No. ИТС 1-22-26сп dated 16.10.2023.

Задача моделирования

Дискретная сетка во времени:

$$T = t_1, t_2, \dots, t_M, \text{где } t_{i+1} - t_i = \tau = const$$

Технологический *набор данных*:

$$F(t) = f_1(t), f_2(t), \dots, f_N(t), \text{ пусть } X = \{f_j(t_i)\}_{MxN}$$

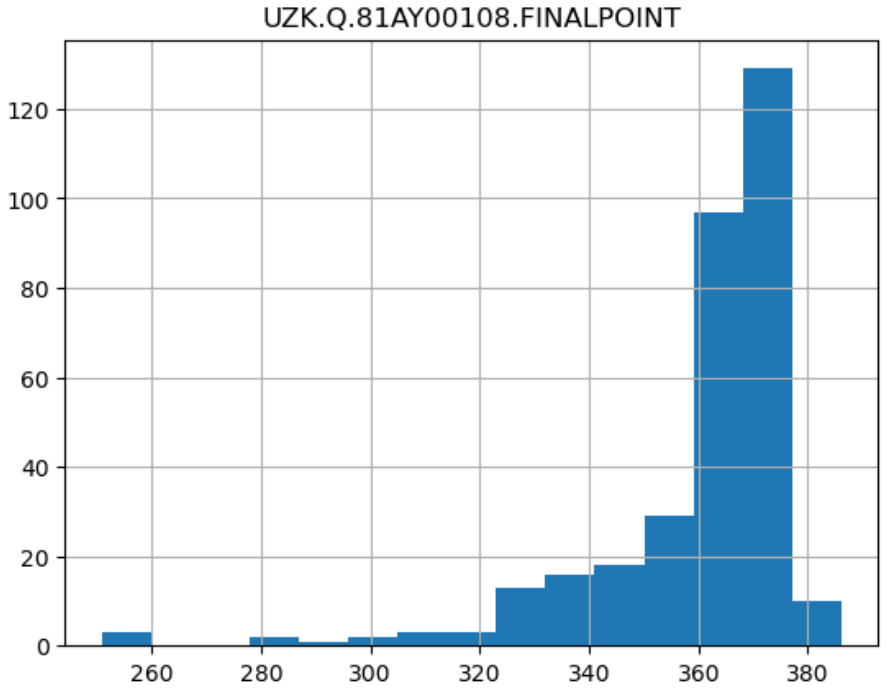
Многомерная *модель технологического процесса* во времени:

$$X_{t+k}, \dots, X_{t+k-r} \approx \tilde{F}(X_t, \dots, X_{t-r}), \text{где } t \geq r$$

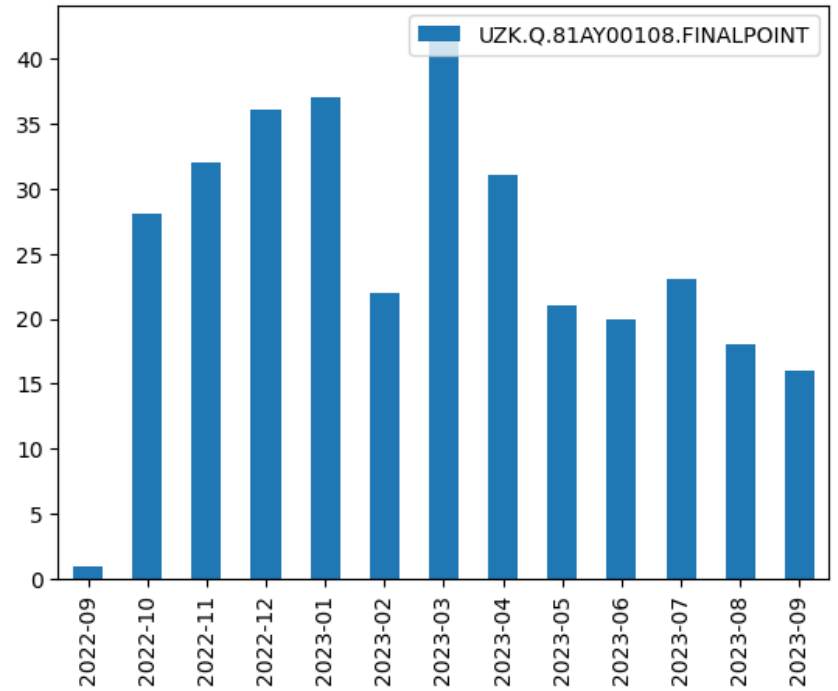
В случае сырых лабораторных данных, $f_{lab}(t)$, соответствующая лабораторной переменной lab, будет иметь пропуски (*NaN*), что будет учтено при обучении конкретных моделей $\tilde{F}$ путем отсеивания таких тренировочных векторов

Подготовка данных. Распределение

РАСПРЕДЕЛЕНИЕ ЦЕЛЕВОЙ ПЕРЕМЕННОЙ

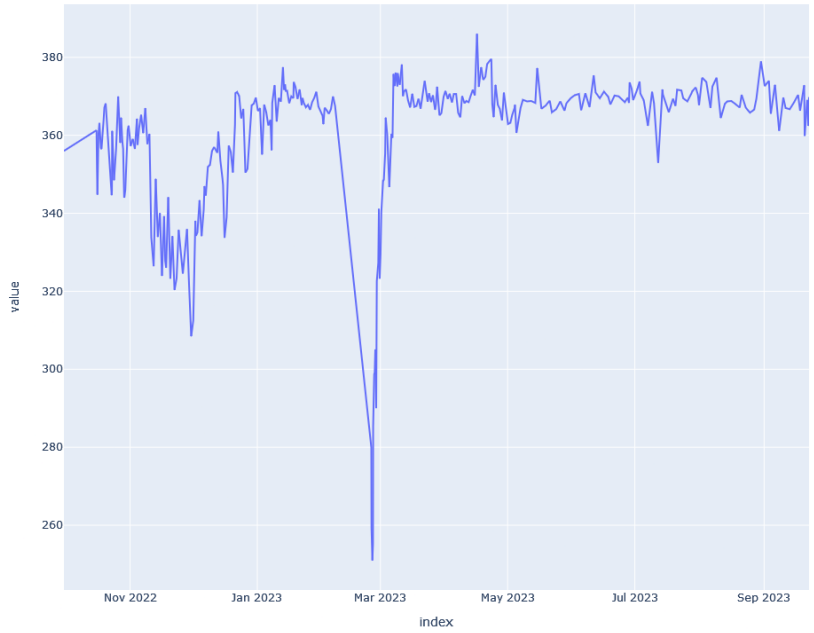


ПЕРИОДИЧНОСТЬ СБОРА ПОКАЗАНИЙ

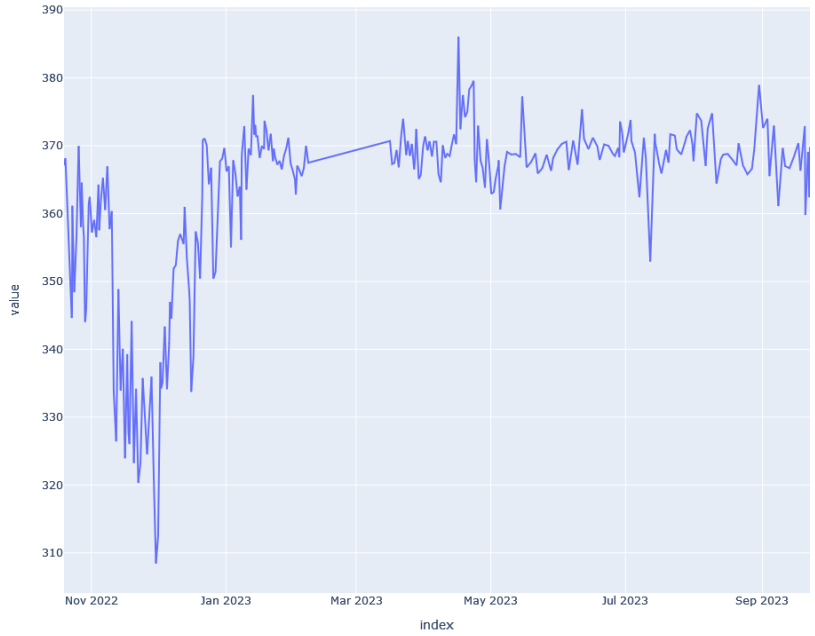


Подготовка данных. Периоды стабильности

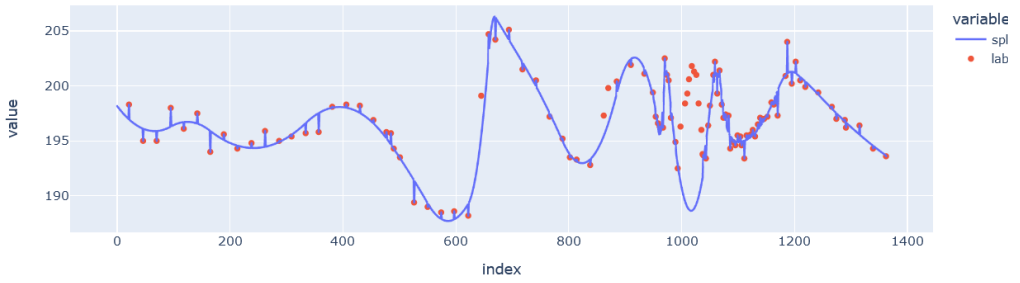
ИСХОДНЫЕ ПОКАЗАНИЯ ЦЕЛЕВОЙ ПЕРЕМЕННОЙ

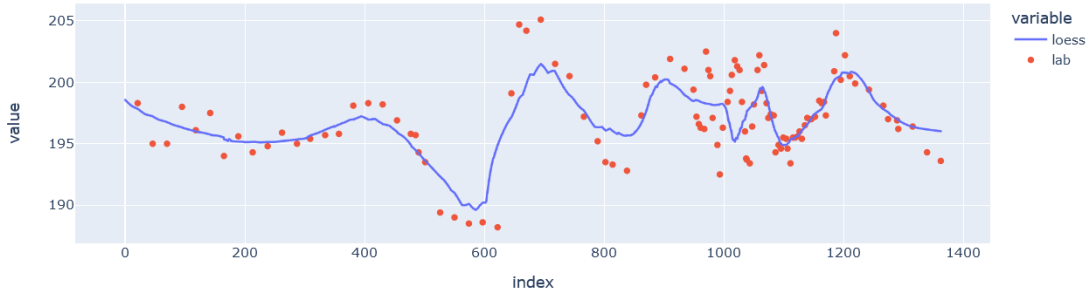


ПОЛУАВТОМАТИЧЕСКИЙ ОТБОР ПЕРИОДОВ СТАБИЛЬНОСТИ

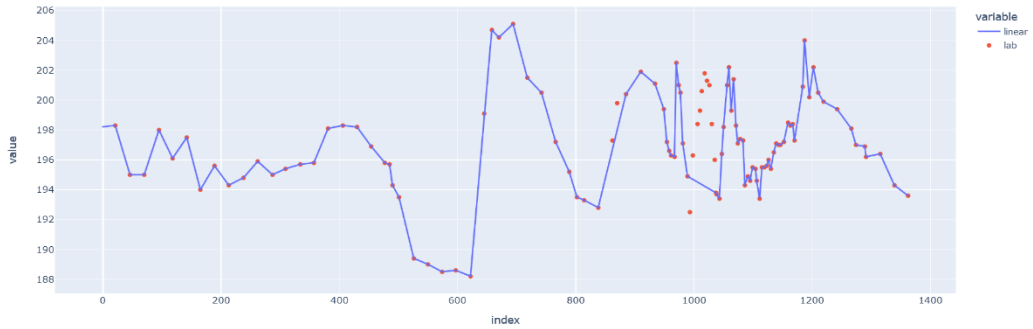


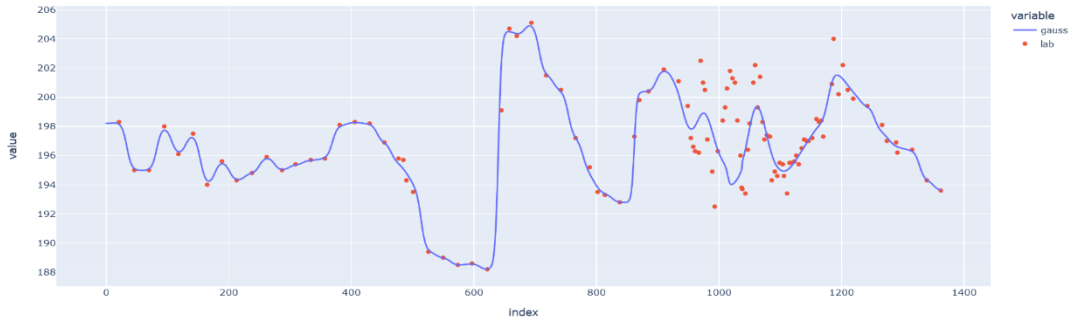
Подготовка данных. Интерполяция





В качестве основных методов базовой интерполяции были выбраны **Сплайн** второго порядка (*сверху слева*) и **Loess-**сглаживание (*сверху справа*). Линейная интерполяция (*снизу слева*) и гауссово ядро (*снизу справа*) так же были рассмотрены, но не зарекомендовали себя в рамках основного набора данных

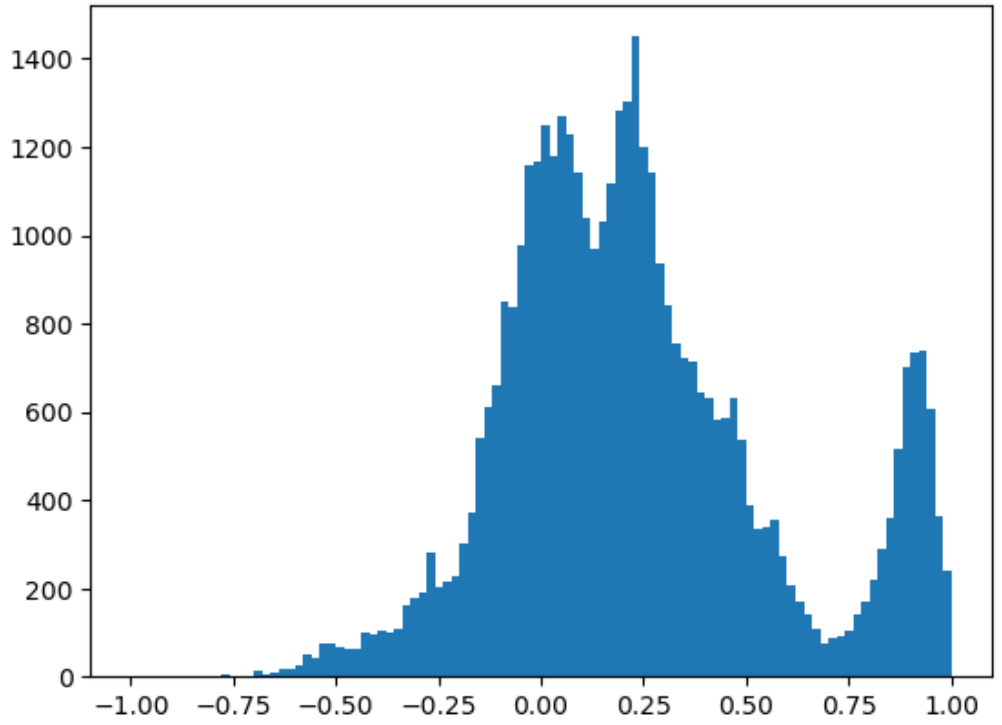




Отбор признаков (1)

Специфика задачи:

- Сильная корреляция внутри набора показаний физических переменных
- Необходимость со стороны технологов в явном представлении зависимостей между переменными
- Необходимость учитывать мнение эксперта области при построении моделей



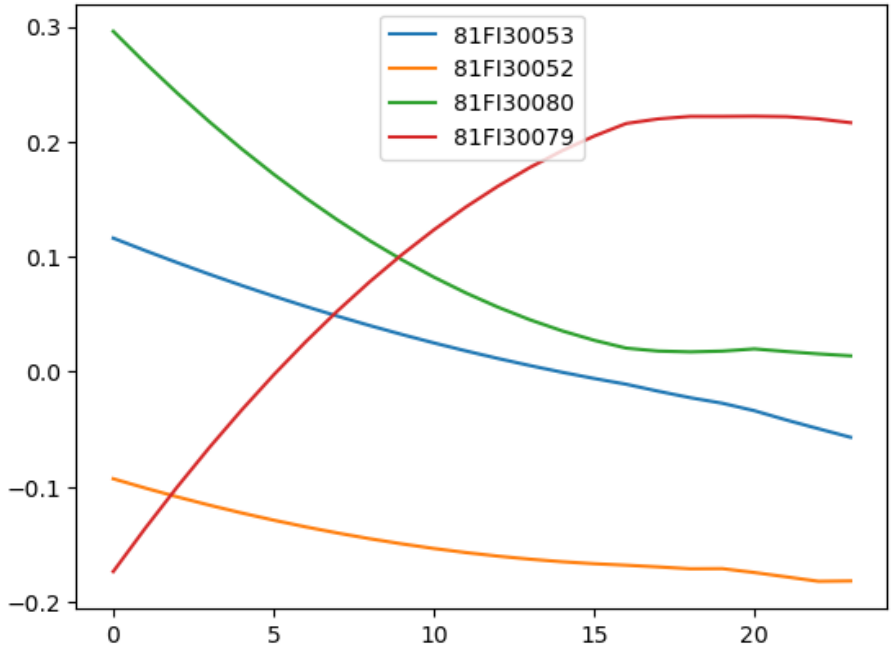
Ненулевые попарные коэф-ты корреляции физических переменных набора

Отбор признаков (2)

Метод	Описание	Пороговая метрика	Простота
PLS-кластеризация	Иерархическая кластеризация на основе регрессии главных компонент	Глубина кластеризации	1 гиперпараметр
Сети Байеса	Авто-генерируемые сети Байеса с корнем в узле целевой переменной	Глубина обхода корня	1-2 гиперпараметра
StemGNN	Графовая нейросетевая модель с механизмами внимания	Число переменных	1-2 гиперпараметра
Регуляризация на основе отклика	Штраф модели на основе модуля корреляции с целевой переменной	-	2 гиперпараметра
Ансамбли деревьев	Важность признаков на основе числа вхождений в ансамбль деревьев	Число вхождений	1+ гиперпараметров
LASSO	Важность признаков как модуль весов LASSO модели.	Число переменных	1 гиперпараметр
Корреляция (baseline)	-	-	-

Отбор признаков. Временная компонента

1. **Неявная.** Попарные корреляции с переменным сдвигом целевой и физических переменных для моделей, не учитывающих временную компоненту в явном виде. Исследуется переменная, сдвинутая во времени до своей наибольшей корреляции с откликом (*PLS, Bayes, Lasso, регуляризация* на основе отклика).
2. **Явная.** Исследуются явные зависимости во времени между физическими и лабораторными переменными, двумерное пространство признаков вида {<переменная, лаг>} (*Ансамбли деревьев*).
3. **Автоматическая**. Компонента времени учитывается внутри модели без дополнительного внимания со стороны пользователя (*StemGNN*).



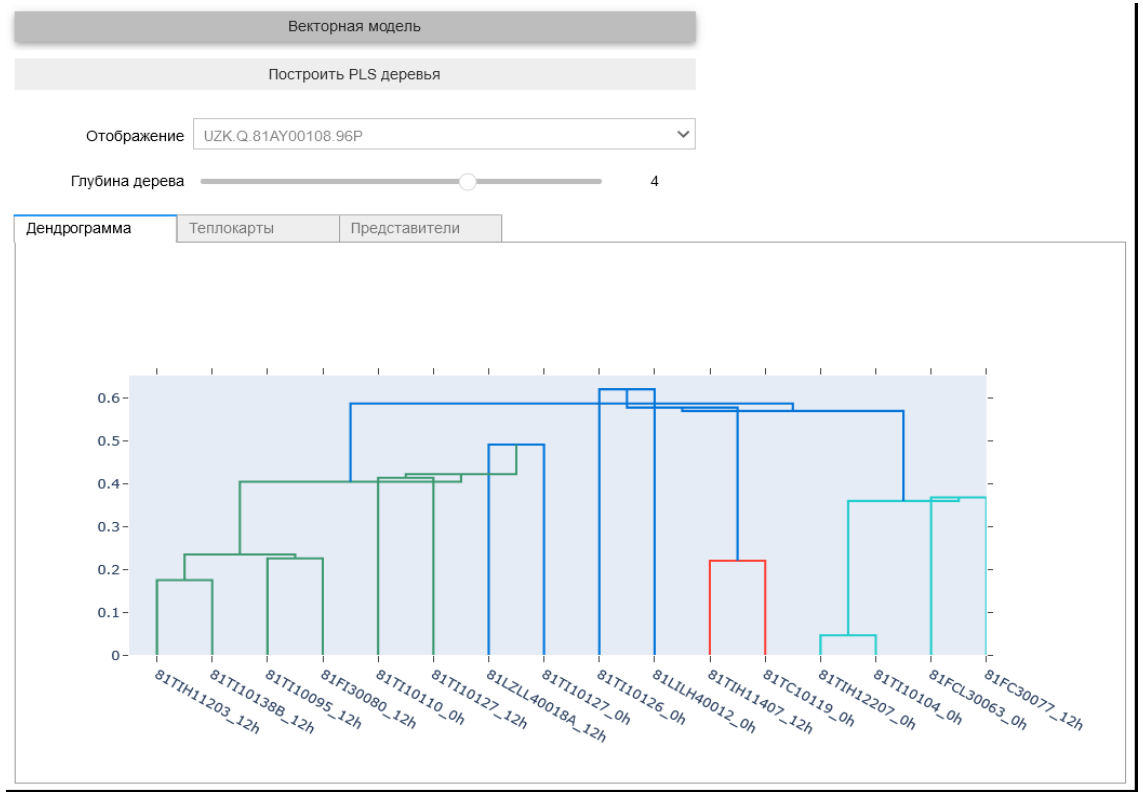
Корреляция переменных с откликом в зависимости от времени

Отбор признаков. PLS-кластеризация

Экспертные переменные выбираются как главные в кластерах, построенных по разбиению на гланые компоненты

Характеристики:

- Общий/групповой/индивидуальный отбор;
- Неявная компонента времени;
- Дендрограмма как визуальная интерпретация;
- Учитывается мнение эксперта офлайн (т.е. во время построения моделей);
- Гарантированное включение обязательных переменных;
- Гиперпараметры (1): порог обрезания дерева.



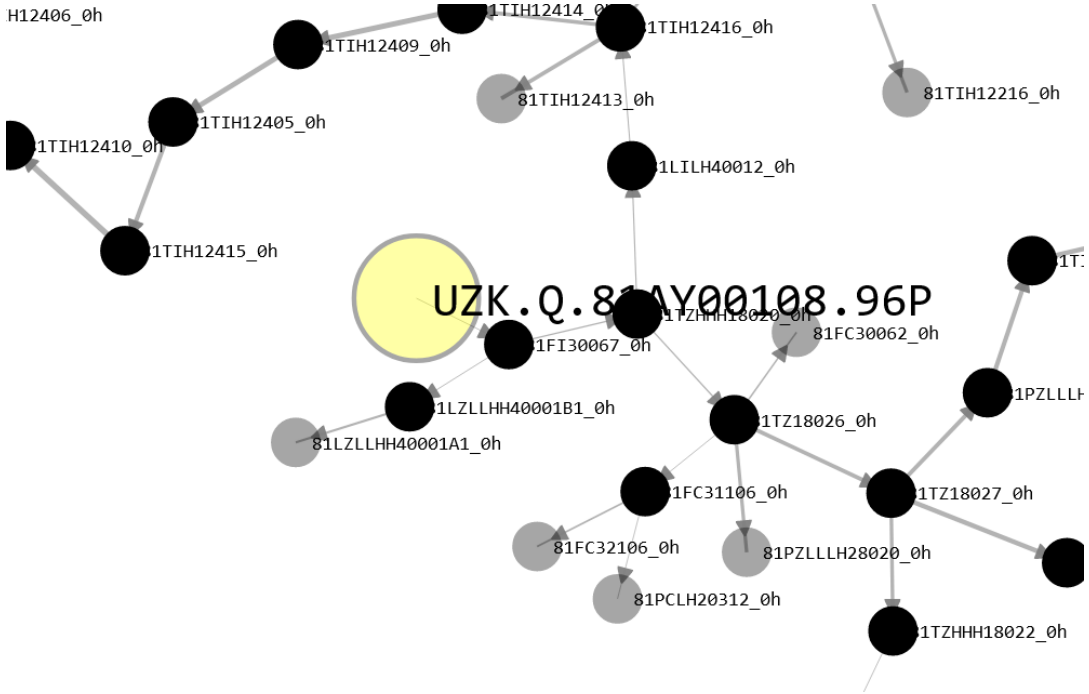
Визуализация работы с *PLS*

Отбор признаков. Сети Байеса

Отбор по поддеревьям, построенных по *Chow-Liu* с корнем в целевой переменной

Характеристики:

- Индивидуальный отбор;
- Неявная компонента времени;
- Древесное представление зависимостей;
- Учитывается мнение эксперта офлайн;
- Гарантированное включение обязательных переменных;
- Гиперпараметры (1-2): число отобранных признаков [, порог p-value].



Визуализация работы с деревьями

Отбор признаков. StemGNN

Отбор на основе графовой нейросети, использующей механизмы внимания *StemGNN*. Модификация: ложный вход для лабораторных данных. Только интерполированные данные.

Характеристики:

- Общий контекстуальный во времени отбор переменных;
- Автоматическая временная компонента – отсутствует лаг как параметр;
- Визуализация зависимостей в виде теплокарты внимания;
- Учитывает мнение эксперта онлайн/офлайн;
- Гарантированное включение обязательных переменных;
- Опциональная трансформация признаков;
- Гиперпараметры (0-2): [число слоев] [, число переменных].

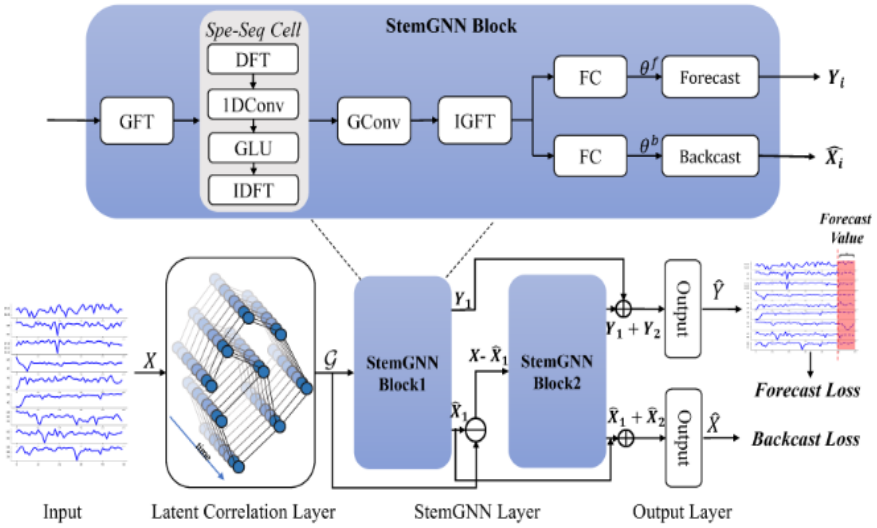


Figure 1: The overall architecture of Spectral Temporal Graph Neural Network.

StemGNN в оригинальной статье

[2]Cao, Defu, et al. "Spectral temporal graph neural network for multivariate time-series forecasting." Advances in neural information processing systems 33 (2020): 17766-17778.

Виртуальные анализаторы. Отбор признаков. Регуляризация

Вектор корреляции (выбранного типа) физических и лабораторных показаний (индивидуально) как множитель для *регуляризации нейросетевой модели* лабораторных исследований. Эксперт опционально проставляет единицы для физически значимых зависимостей.

Характеристики:

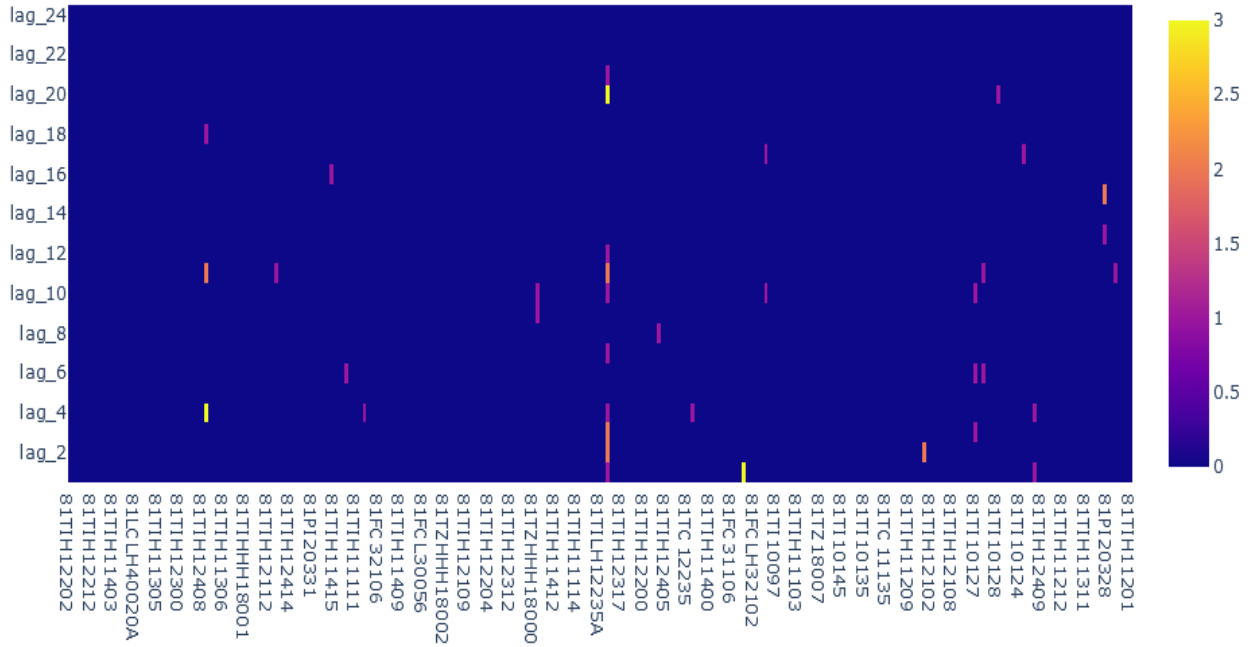
- Индивидуальный отбор;
- Неявная временная компонента;
- Теплокарта корреляций как визуализация зависимостей;
- Учитывает мнение эксперта офлайн;
- Гарантированное включение обязательных переменных;
- Неявное воздействие на участвующие переменные;
- Гиперпараметры (2): тип корреляции, множитель регуляризации.

Виртуальные анализаторы. Отбор признаков. Ансамбли деревьев

Градиентный бустинг для отбора (*LightGBM)*

Характеристики:

- Индивидуальный отбор;
- Явная временная компонента (отбор по матрице шаг времени-переменная);
- Теплокарта важности в модели как визуализация зависимостей;
- Негарантированное включение обязательных переменных;
- Гиперпараметры (1): число отобранных переменных (порог отбора).



Теплокарта зависимостей для лагов

Отбор признаков. Эксперимент

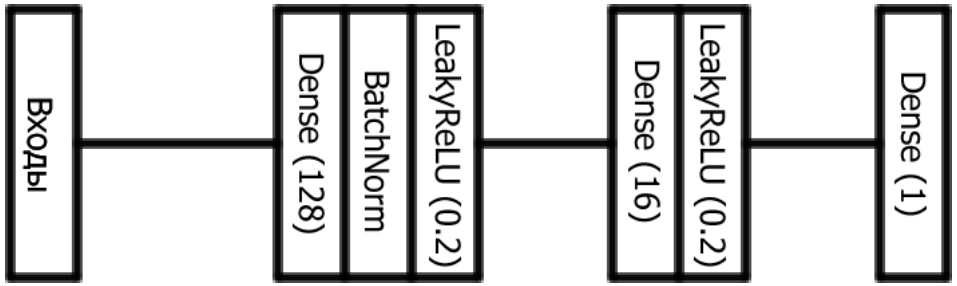
Переменные	PLS	Bayes	Lasso	LGBM	StemGNN
	0,865768	4	0,045041	3	
		4	0,029486		0,947341
	0,963523	4	0,123929		
		2	0,034464		0,334994
	0,79364	1		3	
		4		2	0,414463
	0,994848		0,032323		
	0,930119	2			
		3	0,137878		
	0,979286	4			
		4	0,203014		
		3	0,021009		
		2			0,361444
		3	0,001983		
		4	0,05296		
		3	0,017038		

Прогнозирование

Метод	Описание	Данные для обучения
Базовая модель MLP+Boosting	Полносвязная нейросетевая модель (baseline модель)	Исходный ряд
Простая RNN модель	Простая рекуррентная модель с масштабирующим слоем (простая модель)	Исходный ряд/Spline/LOESS
Seq2seq	Рекуррентная многошаговая модель для физических переменных в ансамбле с полносвязной моделью лаб. исследований (модель прогноза на несколько шагов)	Spline/LOESS
StemGNN	Сложная GNN модель, использующая механизмы внимания (известная сложная модель)	Spline/LOESS
GAN	Генеративно-состязательная сеть, создающая новые тренировочные образцы в процессе обучения (модель борьбы с небольшим тренировочным набором)	Исходный ряд

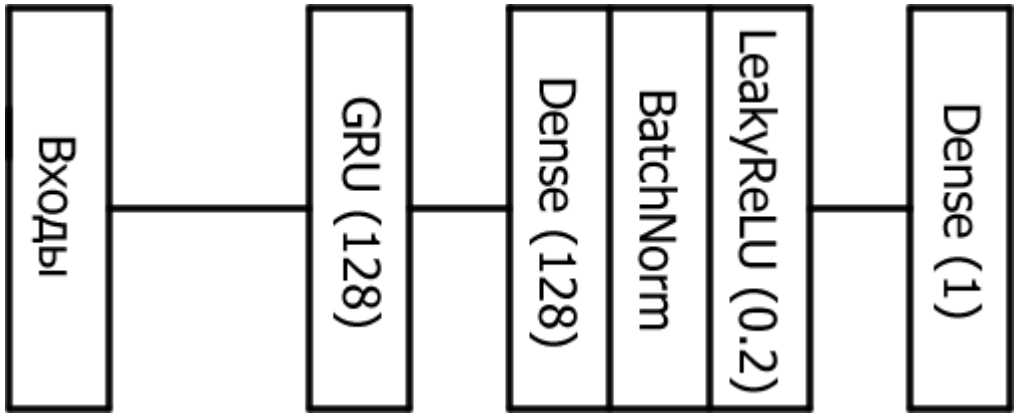
Прогнозирование. Простые модели

Отбор переменных MLP: ансамбли деревьев (*LightGBM*) – плоские признаки на входе

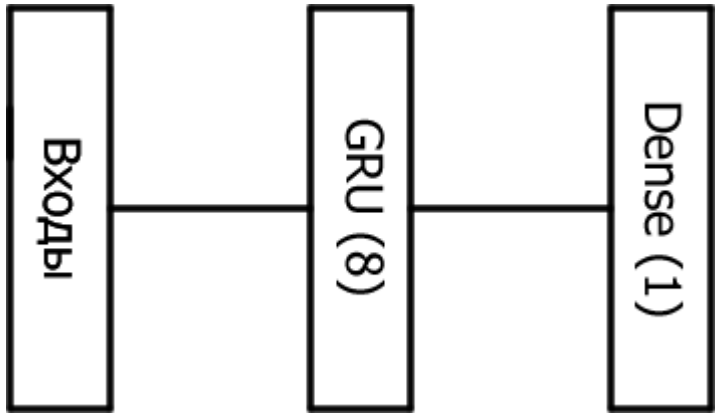


Подобранная архитектура (1)

Отбор переменных RNN: *StemGNN, PLS, Lasso, Bayes, регуляризация* по отклику



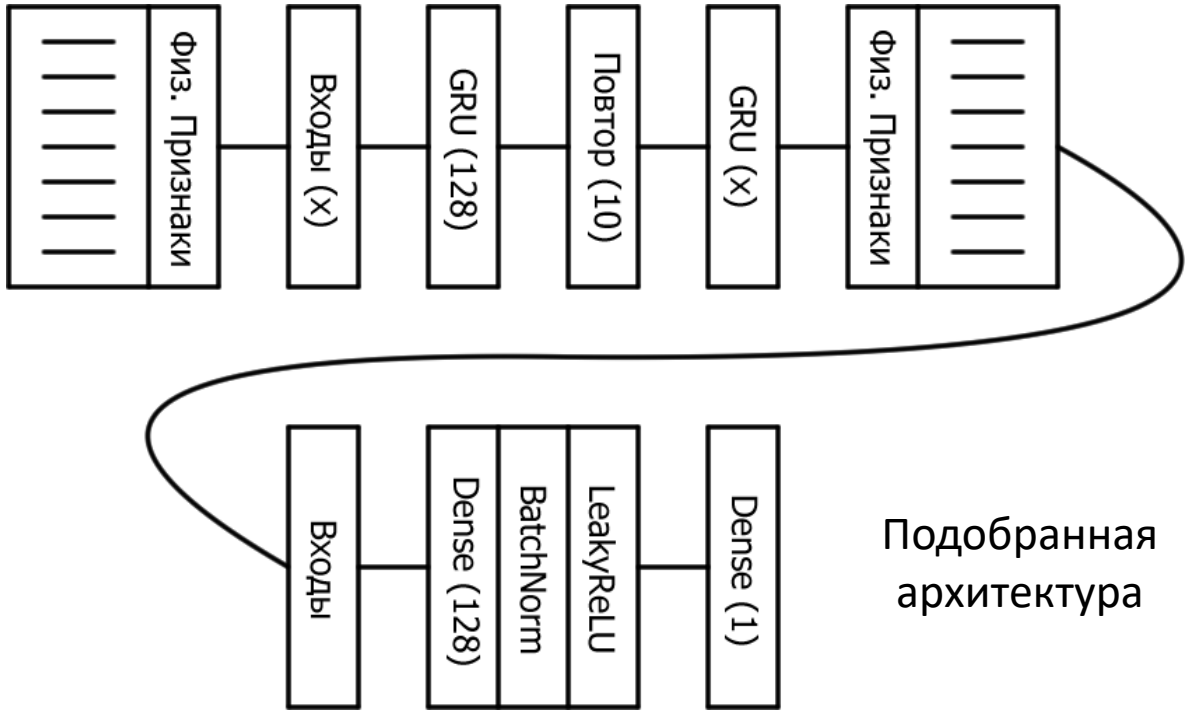
Подобранная архитектура (2)



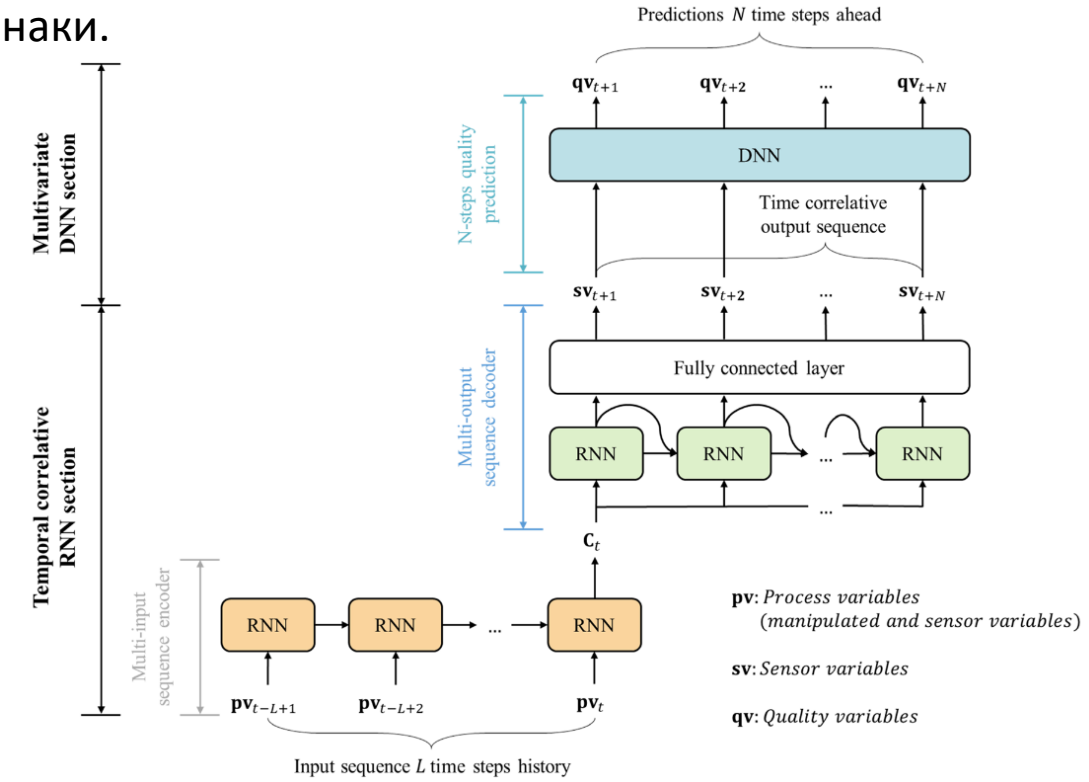
Подобранная архитектура (3)

Прогнозирование. Sequence-to-Sequence

Разновидность модели[3] векторного (в частном случае – пошагового) прогноза физических данных (напр. RNN-RNN). Одновременно обучается более простая модель (MLP) прогнозирования лабораторных исследований, использующая выходы *Seq2Seq* модели как признаки.



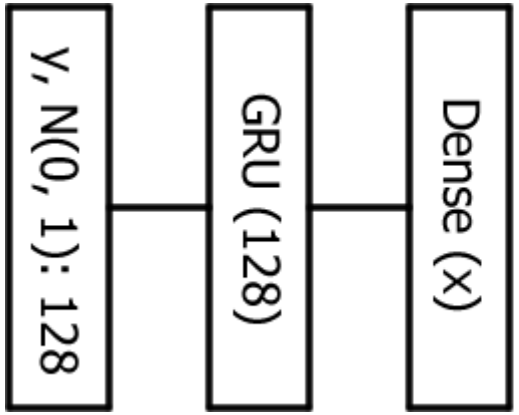
Подобранная архитектура



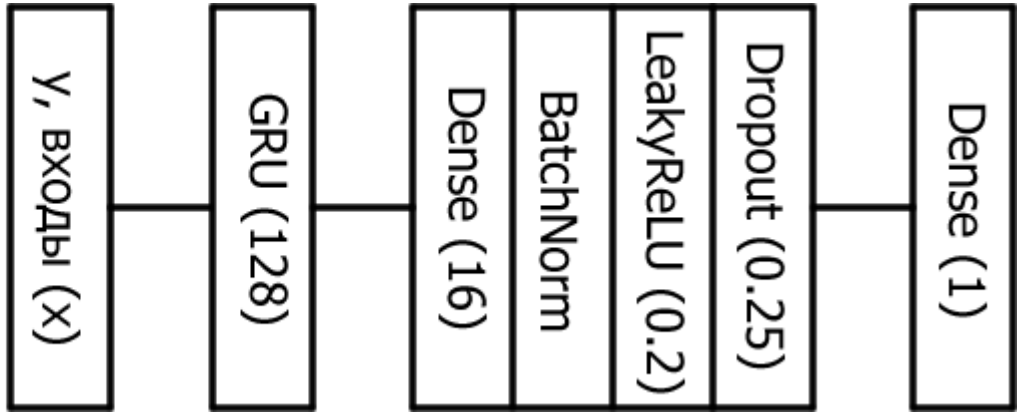
[3]Hong, Seokyoung, et al. "A dynamic soft sensor based on hybrid neural networks to improve early off-spec detection." Engineering with Computers 39.4 (2023): 3011-3021.

Прогнозирование. GAN

Рекуррентный регрессор, обучаемый в рамках GAN. **Отбор переменных**: *StemGNN, PLS, Lasso, Bayes*



Подобранная
архитектура генератора

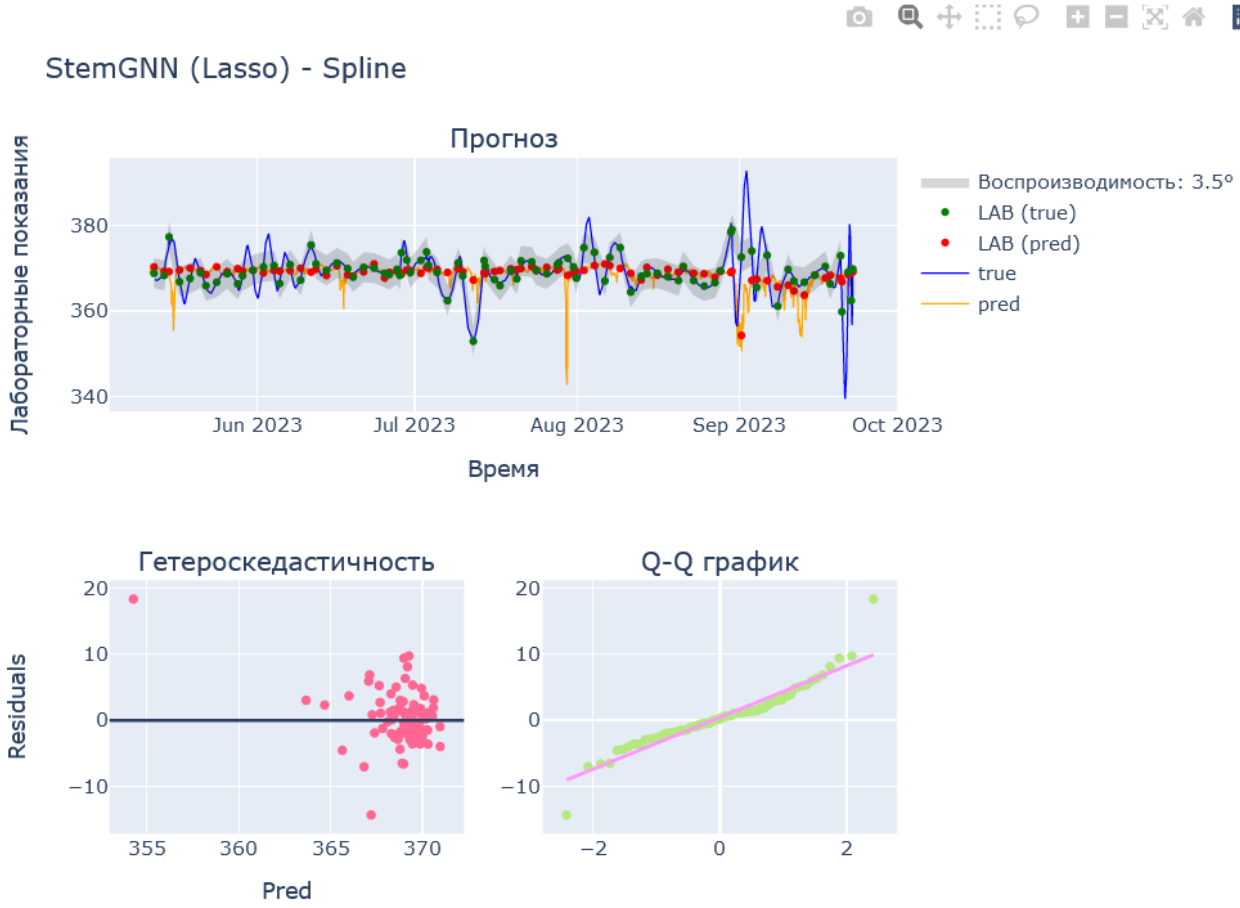


Подобранная архитектура
дискриминатора и регрессора

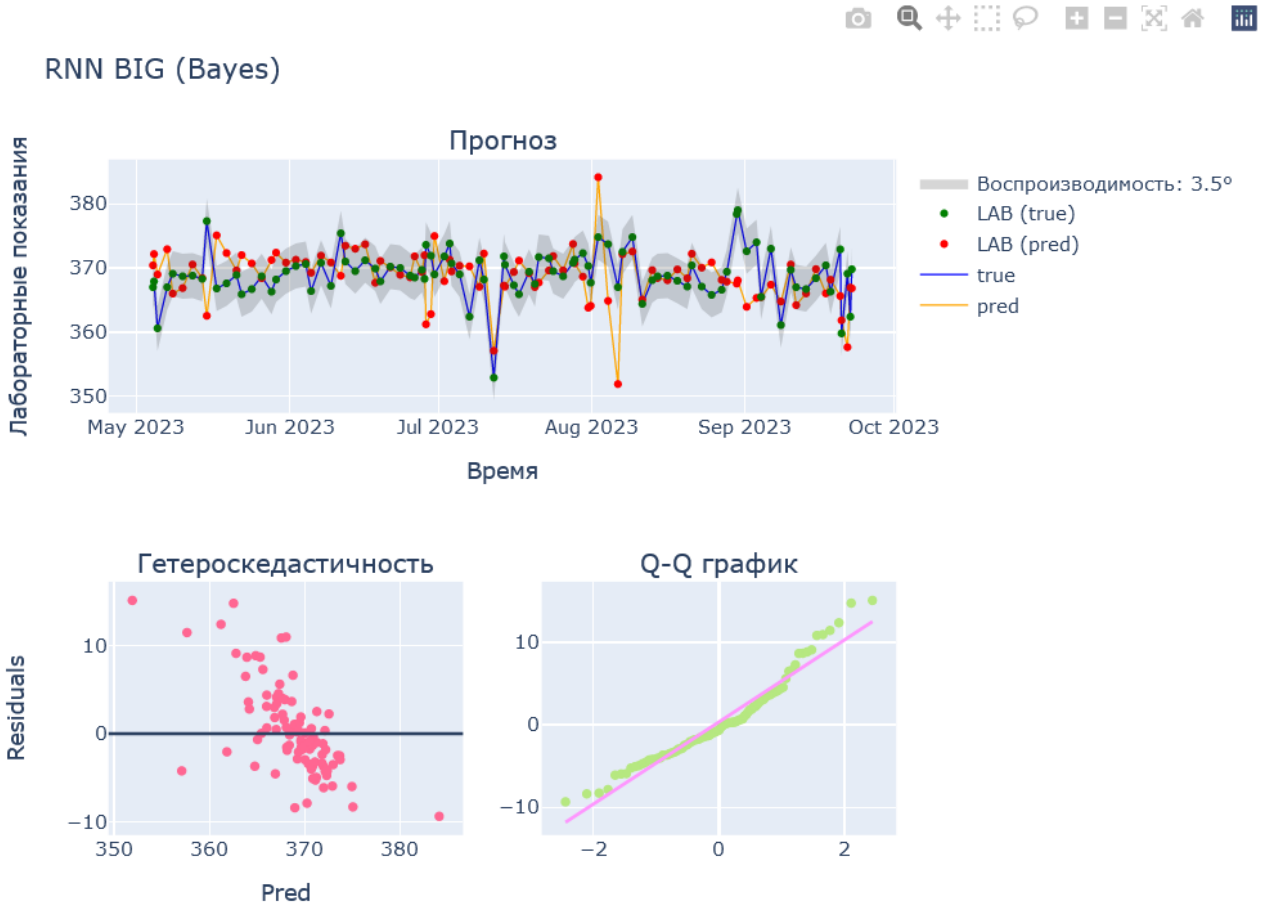
Прогнозирование. Эксперимент (1)

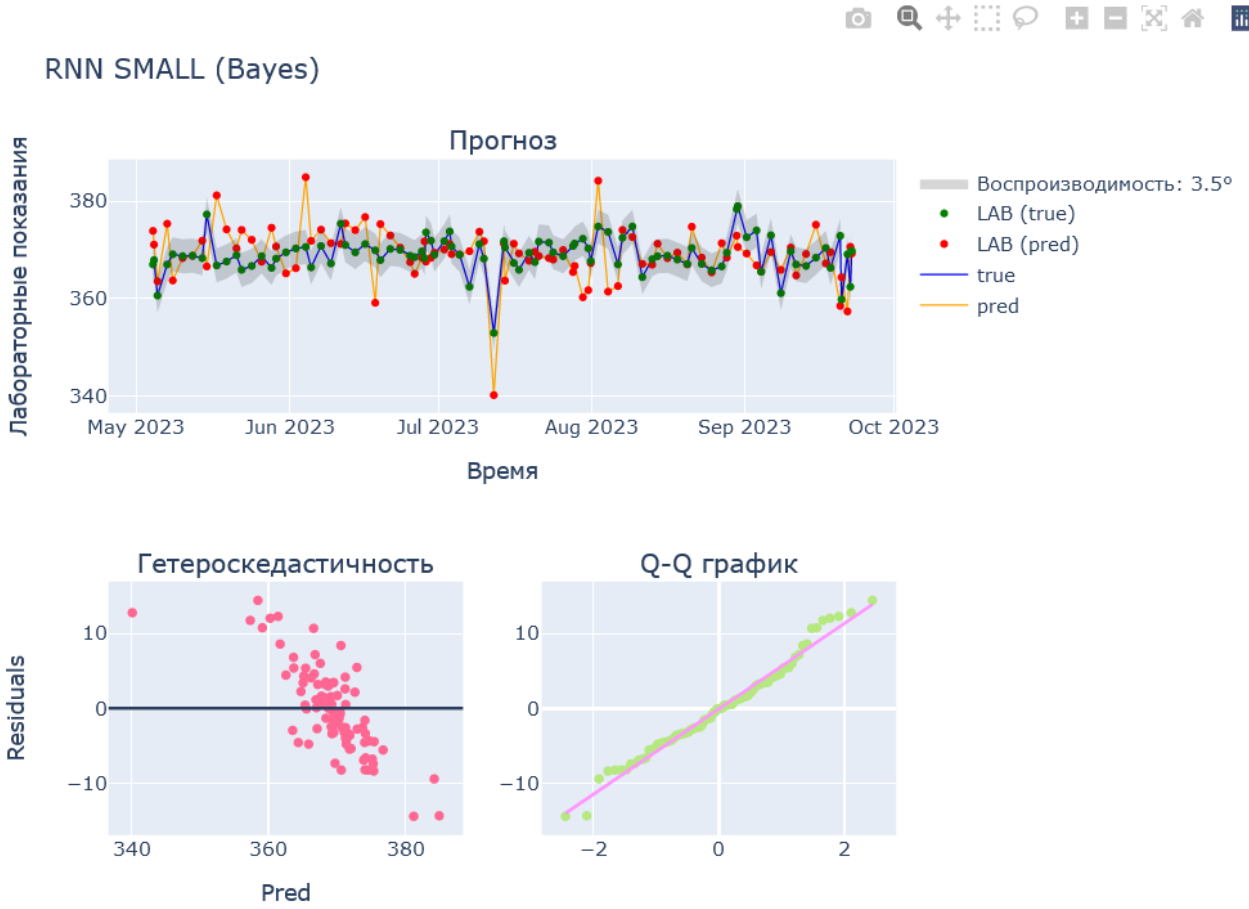
- TRAIN:TEST = 6:4
- Исследуемые методы: *MLP, StemGNN, RNN (small/large), GAN, Seq2Seq*
- Исследуемые методы отбора: *Corr+PLS, Corr+Bayes, Corr+Lasso, StemGNN, LightGBM*
- Максимальный лаг моделей при построении пространства признаков: 24 часа
- Исследуемые методы интерполяции (<24 часов): Spline, LOESS
- Шаг моделей: 1 час вперед от последнего известного значения физических переменных;
- Задача: аппроксимация + прогноз.
- Отбор моделей: p-value корреляции (адекватность), положительный R2, RMSE

Виртуальные анализаторы. Прогнозирование. Эксперимент (2)



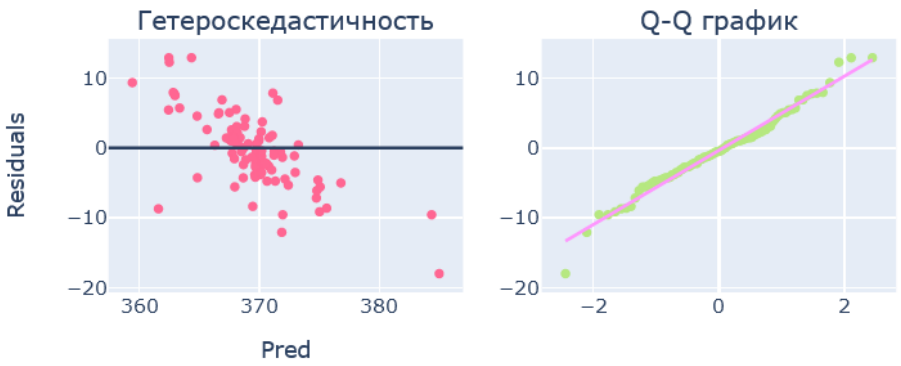
Виртуальные анализаторы. Прогнозирование. Эксперимент (3)



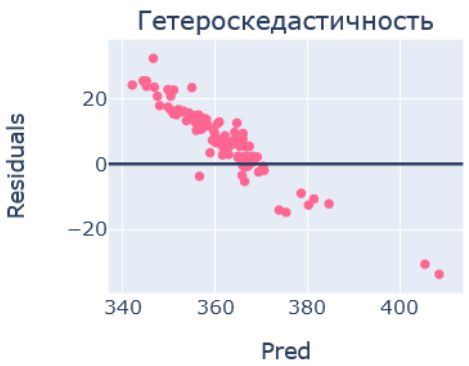


Виртуальные анализаторы. Прогнозирование. Эксперимент (4)

MLP (LightGBM)



GAN (LightGBM)



Виртуальные анализаторы. Прогнозирование. Эксперимент (5)

Результаты

- Были предложены несколько подходов, объединяющих задачи прогноза и заполнения пропусков для моделирования лабораторных исследований. Рассмотрены методы отбора признаков на основе PLS регрессии (иерархическая кластеризация), градиентного бустинга, деревьев Байеса, графовой нейронной сети, базовые методы в виде Lasso и корреляций. Каждый из предложенных методов отбора был адаптирован для возможности учитывать экспертное мнение специалистов области
- Были предложены методы прогнозирования лабораторных исследований на основе графовых нейросетей, генеративных подходов, полносвязных и рекуррентных сетей, рассмотрены аппроксимации отклика с использованием LOESS и сплайнов, а также вариант регуляризации, учитывающей корреляцию с откликом
- Полученные авторами экспериментальные результаты показывают преимущество использования для данной задачи простых рекуррентных сетей, графовых нейросетей с предварительной интерполяцией, отбора признаков на основе деревьев Баейса. Был предложен комбинированный подход, учитывающий адекватность модели, корреляцию ее с истинными значениями лаборатории и стандартные ошибки

Публикации

Статьи:

- Lazukhin I. S., Petrovskiy M. I., Mashechkin I. V. Deep Learning Methods for Tasks of Creating Digital Twins for Technological Processes //Moscow University Physics Bulletin. – 2023. – Т. 78. – №. 1. – С. S3-S15.

Доклады:

- 2023 Deep learning methods for the tasks of creating "digital twins" for technological processes (Приглашенный), Авторы: Petrovskiy Mikhail, Lazuhin Ivan, The 7th International Conference on Deep Learning in Computational Physics DLCP23, Санкт-Петербург, Россия, 21-23 июня 2023;
- 2023 Дискретная модель оптимального управления при ограничениях на основе (Устный), Авторы: Машечкин И.В., Петровский М.И., Лазухин И.С., Ломоносовские чтения - 2023, Секция вычислительная математика и кибернетика, 4-14 апреля 2023, Москва, МГУ, факультет ВМК, Россия, 4-14 апреля 2023.

Спасибо за внимание!

StemGNN

StemGNN[2] (Attention для признаков + Graph Neural Networks) – попытка объединить использование темпоральных и межпризнаковых зависимостей в одной архитектуре. Рассматривают задачу прогноза как исследование графа связей между признаками. В задаче прогноза на несколько шагов предлагают использовать кареточный подход, т.е. последовательно прогнозировать и подавать выход прогноза на вход для расчета следующего шага. Используется дополнительный выход модели для автоэнкодинга.

Темпоральные связи обрабатываются с помощью графового преобразования Фурье (GFT), разлагая по базису собственных значений лапласову матрицу графа.

Так же используется предложенный блок Spe-Seq Cell (имплементирующий дискретное преобразование Фурье – DFT) для разложения выходов GFT по частотному базису.

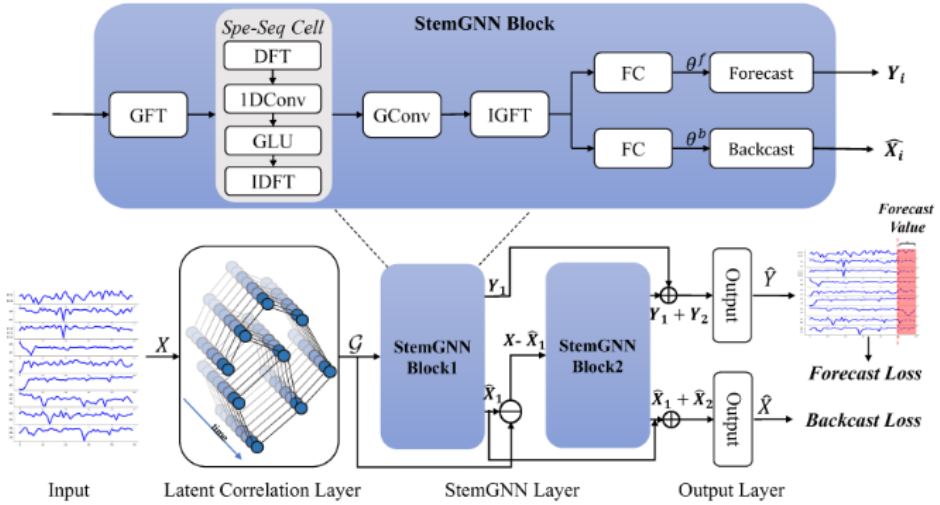


Figure 1: The overall architecture of Spectral Temporal Graph Neural Network.

[2]Cao, Defu, et al. "Spectral temporal graph neural network for multivariate time-series forecasting." Advances in neural information processing systems 33 (2020): 17766-17778.

Метрики, объединенные по лагам и шагам

feature	model	metric	data	rmse	mae	mape	r2	Pearson (p-value)	hinge
-	RNN_BIG	mse	loess	2,535649	1,907155	0,005158	-0,63966	0,78063	0,211197
-	RNN_REG	mse	raw	4,863678	3,567031	0,009662	0,147436	0,922943	1,241325
BAYES	S2S	mse	loess	5,305752	4,21272	0,011398	-1,84303	0,994548	1,560256
-	STEMGNN	mse	loess	5,311611	4,181584	0,011313	-0,1593	0,970011	1,685108
AE	MLP_GREEDY	mse	raw	5,800053	4,234657	0,011468	-0,27718	0,990812	1,751397
AE	MLP	mse	raw	6,823525	5,07043	0,013702	-0,51792	0,945529	2,447496
BAYES	S2S_REG	mse	loess	6,93979	6,517621	0,017634	-0,52716	0,990655	3,181237
PLS	RNN_SMALL	mse	raw	7,070208	5,315755	0,014364	-0,25371	0,625487	2,608204
AE	GAN	mse	raw	14,78332	11,11681	0,03006	-2,14625	0,680998	8,005231
LASSO	GAN_RNN	mse	raw	16,20799	12,94707	0,035028	-2,14692	0,998559	9,712436