Вычислительный эксперимент

Цель

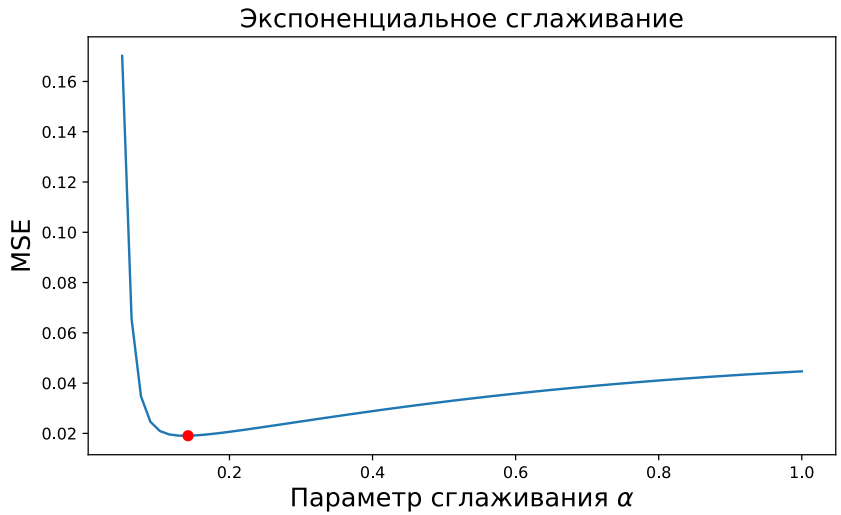
Сравнить качество прогноза суперпозиций прогностических моделей с качеством прогноза базовых моделей.

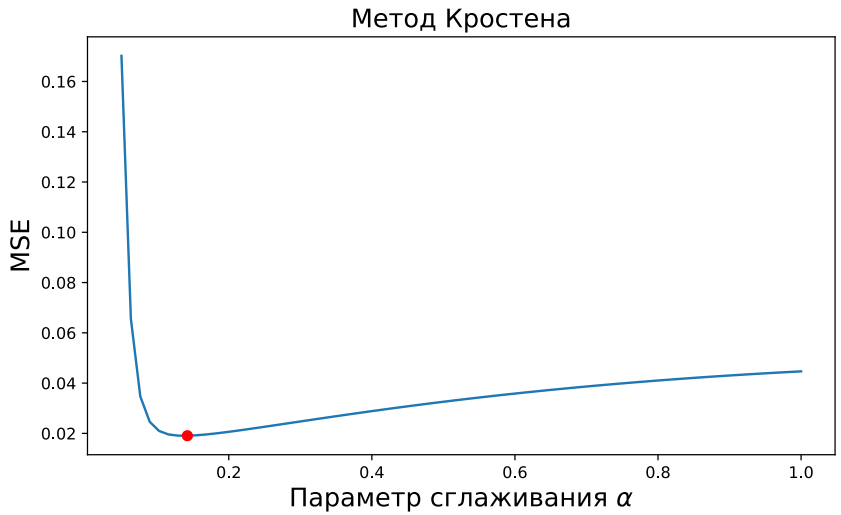
- На реальных данных
- На синтетических данных в зависимости от степени асимметричности распределения.

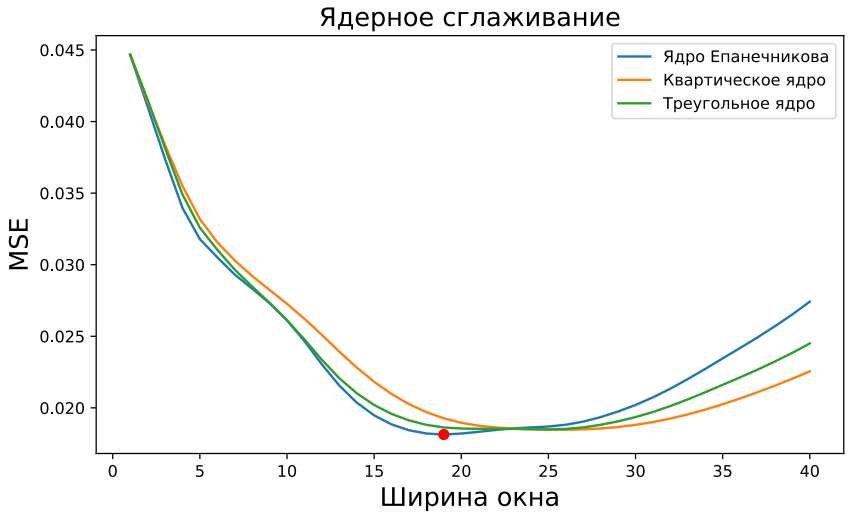
Данные

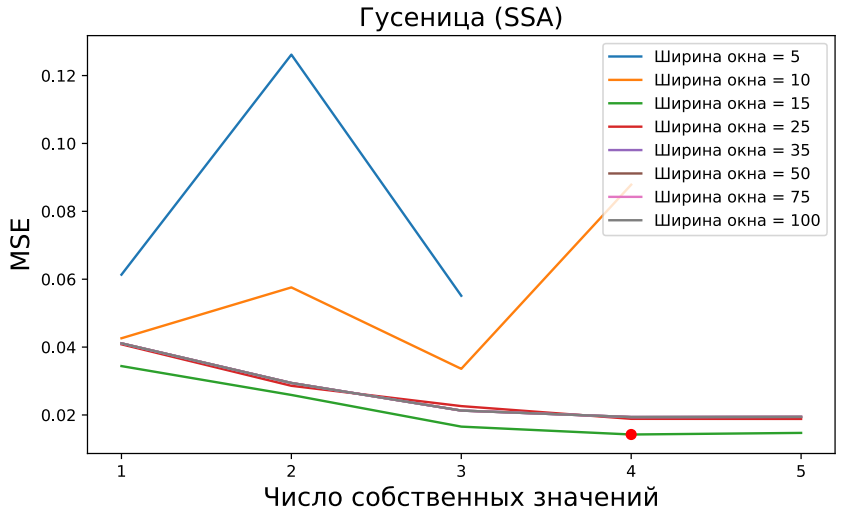
1. Пассажирские авиаперевозки по месяцам,
2. Объём железнодорожных перевозок нефти в Омской области по дням,
3. Цена на электричество в Германии по часам,
4. Потребление электроэнергии в Польше по часам.

Зависимость качества от структурных параметров



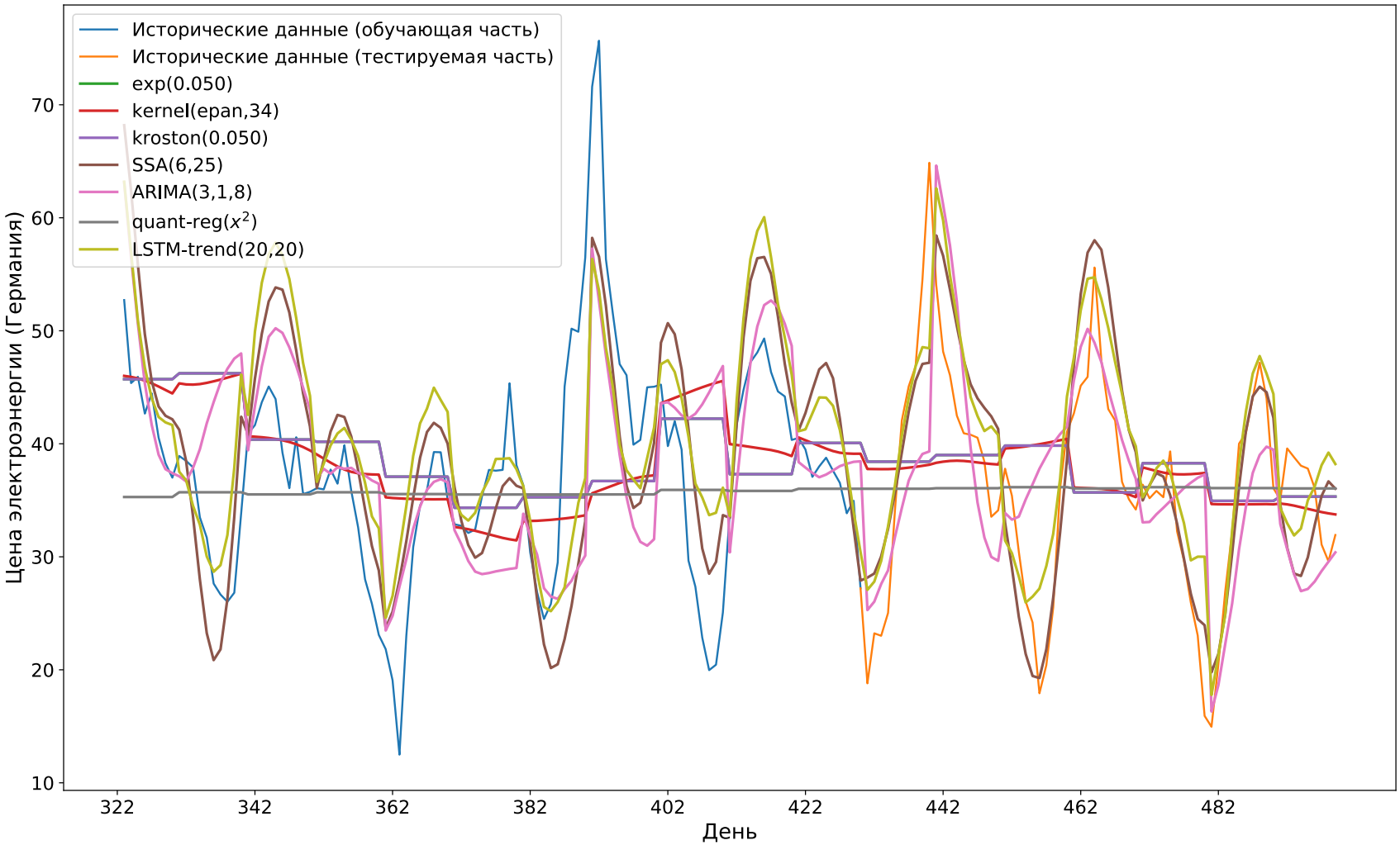






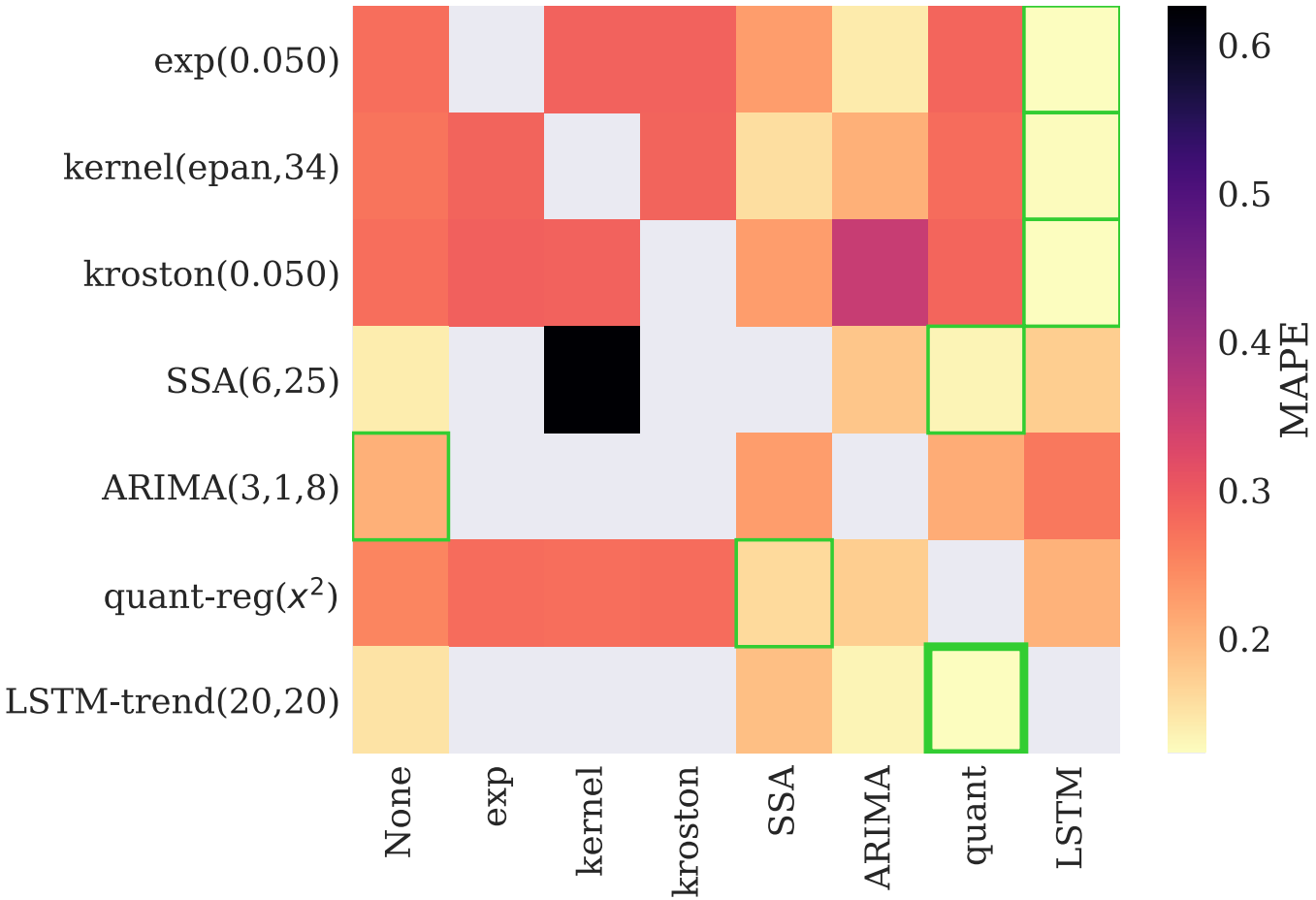
Выпуклый характер зависимости показывает, что используемые сетки структурных параметров позволяют в эксперименте определить значения, близкие к оптимальным.

Прогнозы основных моделей



Предсказания всех моделей после подбора оптимальных структурных параметров адекватно описывают периодическую компоненту временного ряда (компоненту f).

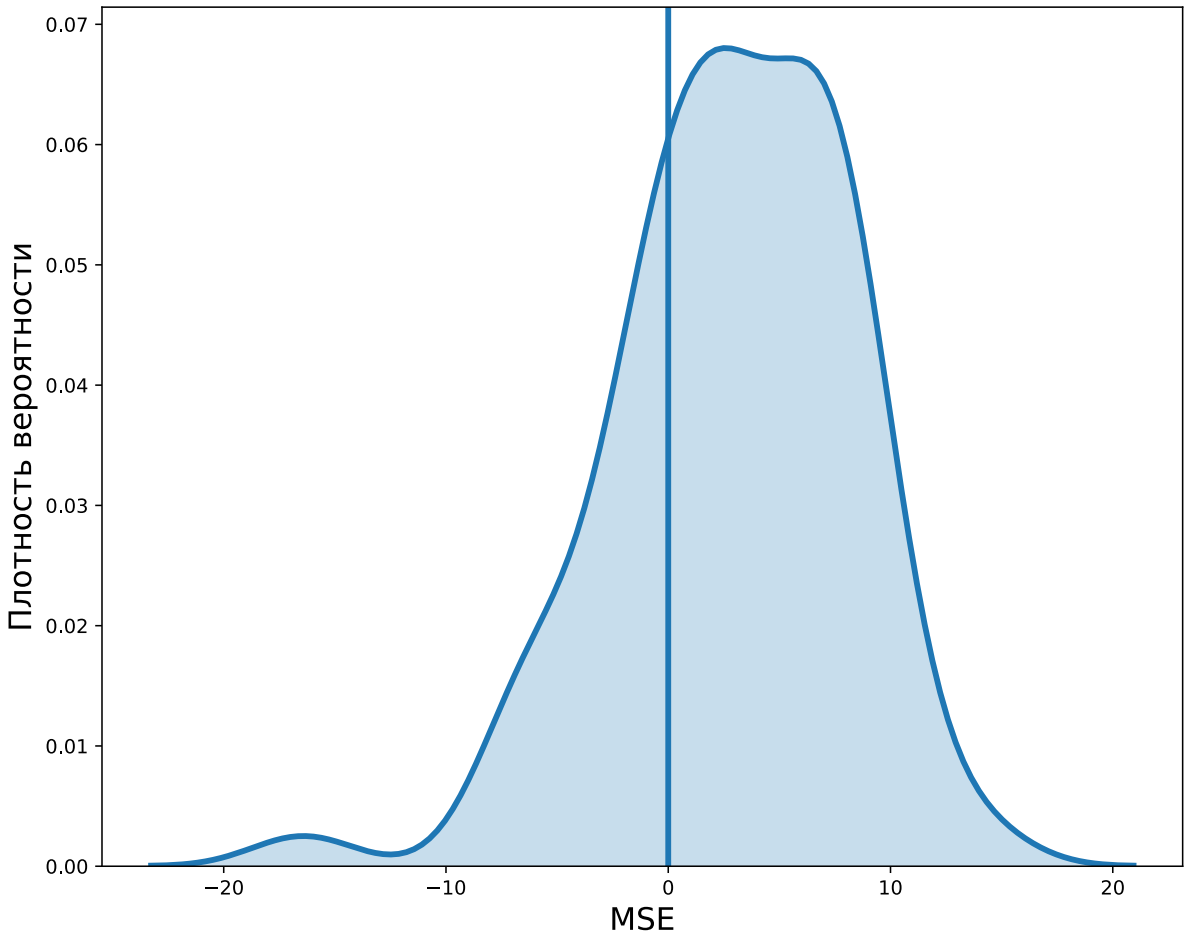
Матрица качества суперпозиций



Лучшие базовые модели – SSA и LSTM.

Лучшие модели остатков – квантильная регрессия и LSTM.

Эмпирическая функция распределения ошибки

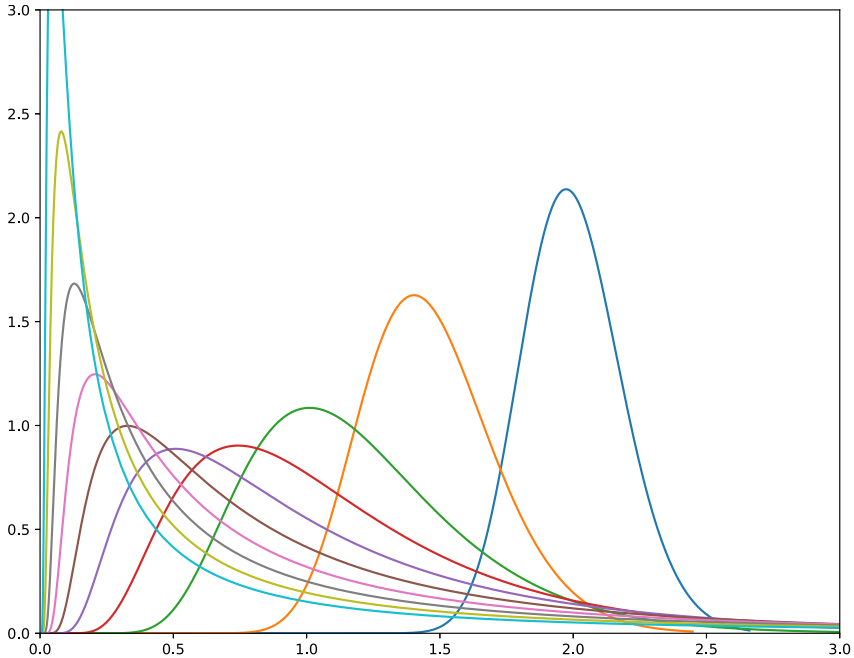


Наблюдаемое распределение ошибки имеет ненулевое матожидание и асимметрично.

Вычислительный эксперимент на синтетических данных

Однопараметрическое (параметр l) семейство распределения шума: распределение Вальда с параметрами

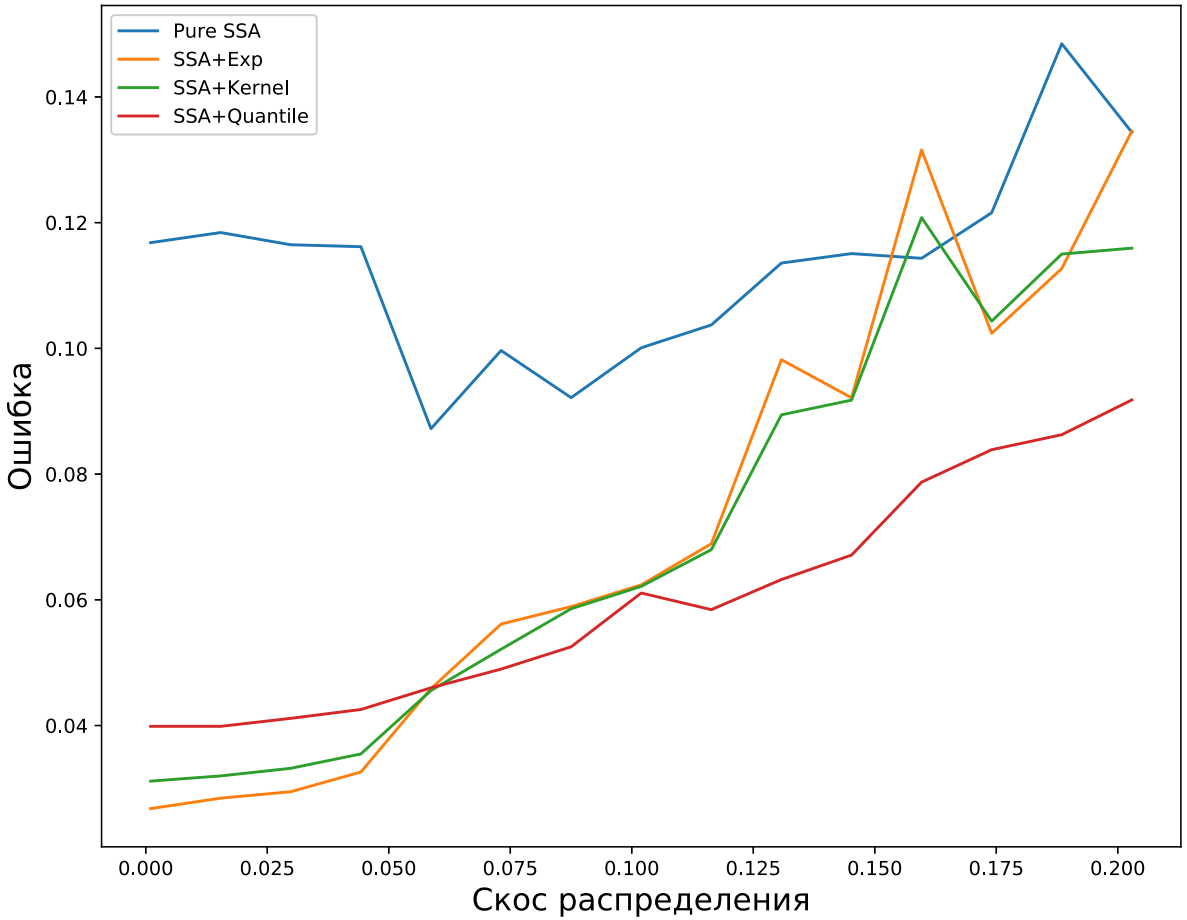
$$\mu = 1 + \frac{l}{30}, \quad \lambda = \frac{l^2}{4}, \quad f(x, \mu, \lambda) = \left[\frac{\lambda}{2\pi x^3}\right]^{1/2} \exp\left\{\frac{-\lambda(x-\mu)^2}{2\mu^2 x}\right\}.$$



Мера скоса распределения

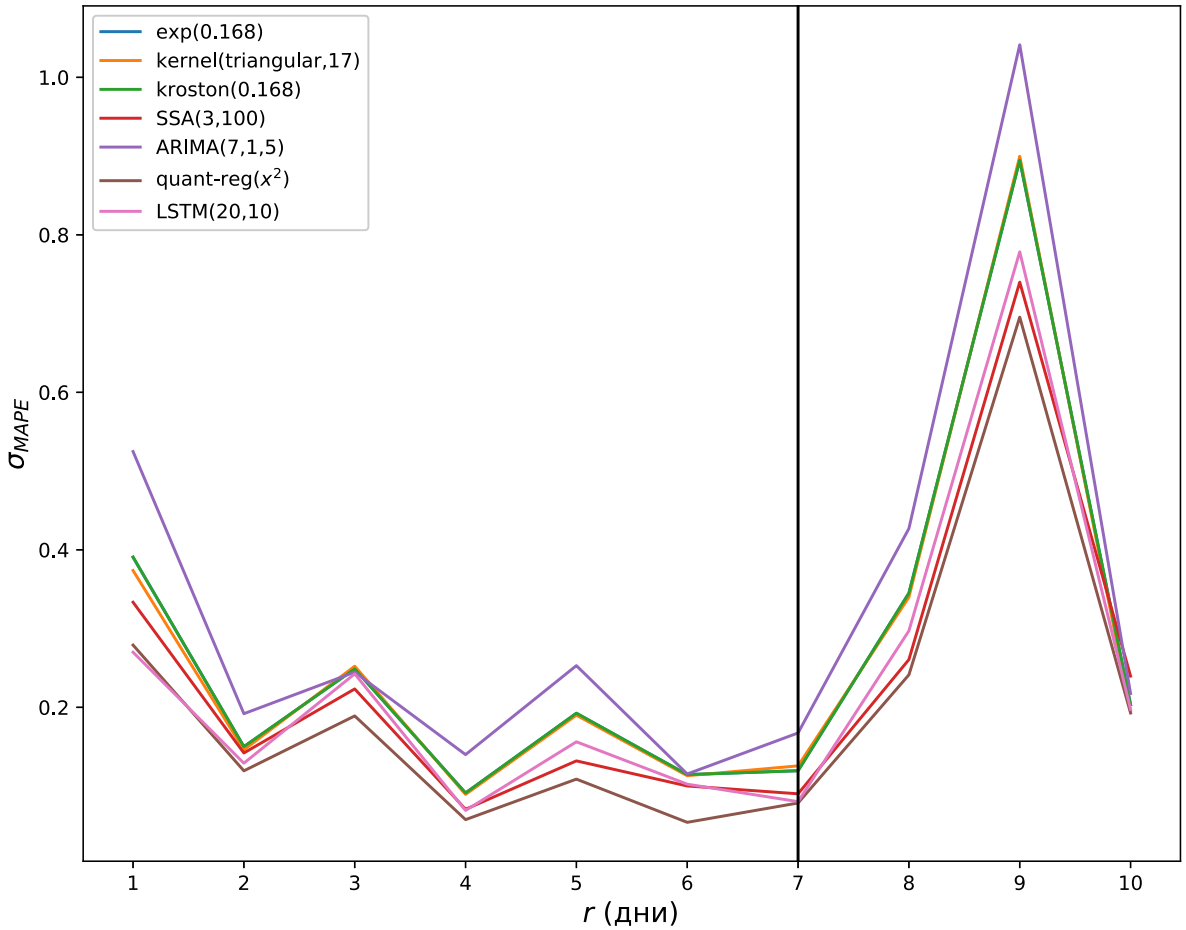
$$S = \left|\mathbb{P}\left(x > x_{\text{н.в.}}\right) - \mathbb{P}\left(x < x_{\text{н.в.}}\right)\right|.$$

Зависимость ошибки от скоса распределения



Суперпозиция с квантильной регрессией даёт меньшую ошибку по сравнению с базовой моделью при прогнозировании ряда с асимметричным шумом.

Зависимость ошибки от горизонта прогнозирования



Горизонт прогнозирования определяется по «правилу сломанной трости»; например, на рисунке $h = 7$.

Результаты

- Предложено два способа построения суперпозиций базовых алгоритмов (экспоненциальное сглаживание, метод Кростена, SSA, ARIMA, квантильная регрессия, LSTM).
- Исследованы свойства суперпозиций прогностических моделей при условии симметричного и асимметричного распределения регрессионных остатков.
- Построены матрицы качества суперпозиций, распределение остатков модели, определён горизонт прогнозирования.
- Показано, что использование суперпозиций может повышать качество прогноза.
- Исследована зависимость ошибки от степени асимметричности распределения шумовых остатков.
- Статья по теме дипломной работы подана в журнал «Вестник московского университета, сер.15. Вычислительная математика и кибернетика»

Вычислительный эксперимент

Цель

Восстановить плотности распределения пространственных ориентаций различных пар вида аминокислота-лиганд.

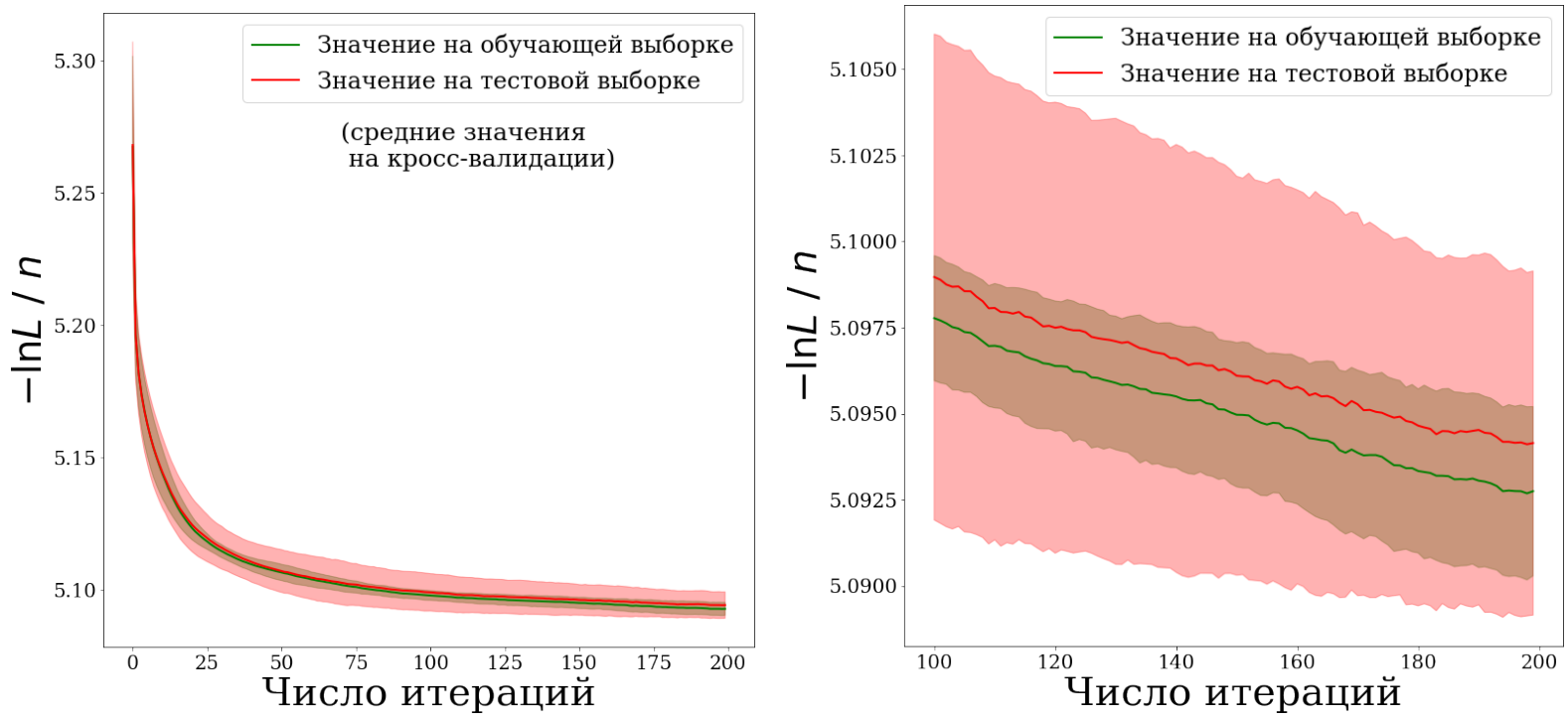
Данные

Данные представляют собой 47916041 пятерку значений, элементы каждой пятерки: a — индекс аминокислоты, b — индекс лиганда и тройка r, θ, φ. Индексы аминокислоты и лиганды образуют 840 пар и используются для разделения данных на 840 выборок (r, θ, φ), каждая из которых соответсвует своей взаимодействующей паре.

Описание эксперимента

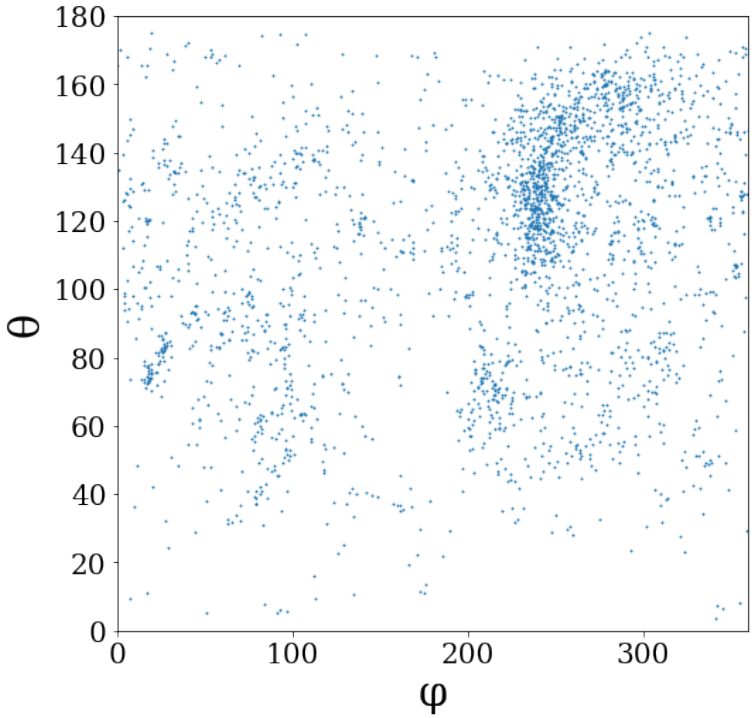
Для каждой из 840 выборок строится восстановленная плотность $\hat{p}^{a,b}(r, \theta, \varphi) = p(r, \theta, \varphi | \boldsymbol{w}^*, \mathbf{U}^*)$.

Иллюстрация свойств алгоритма

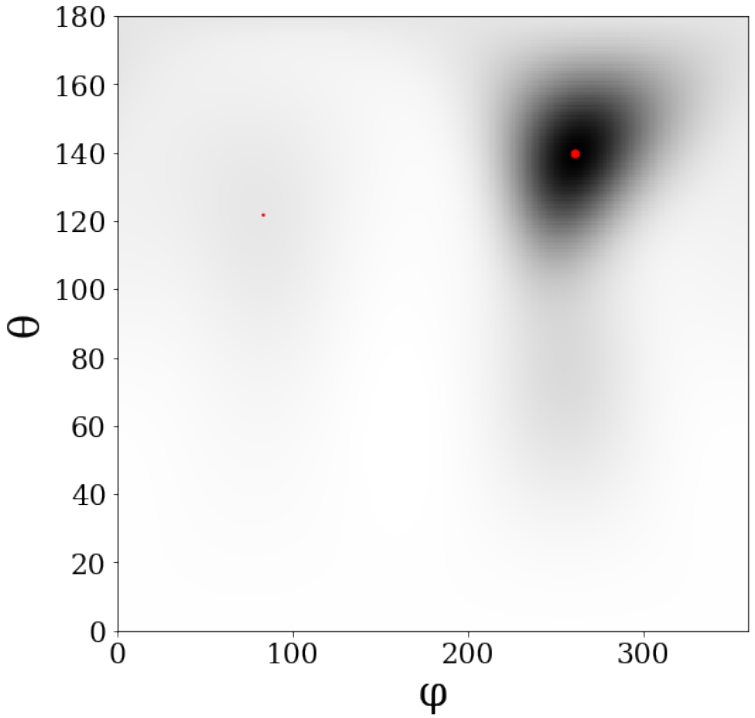


Среднее на кросс-валидации значение отношения логарифма правдоподобия к объёму, соответственно, обучающей и тестовой выборок. График иллюстрирует сходимость алгоритма и отсутствие переобучения.

Результаты восстановления, пара $0 - 2$, $r = 7$Å

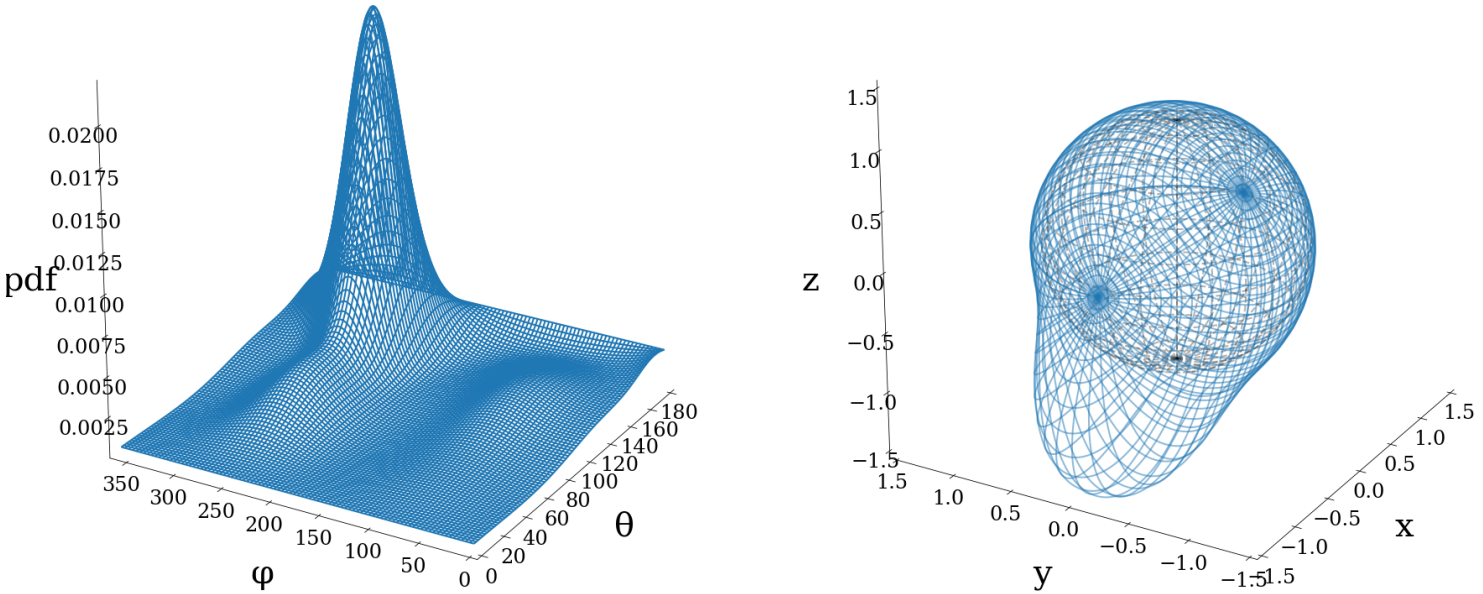


Множество элементов выборки в диапазоне расстояний $r = 7 \pm 0.5$, спроецированное на плоскость (φ, θ).



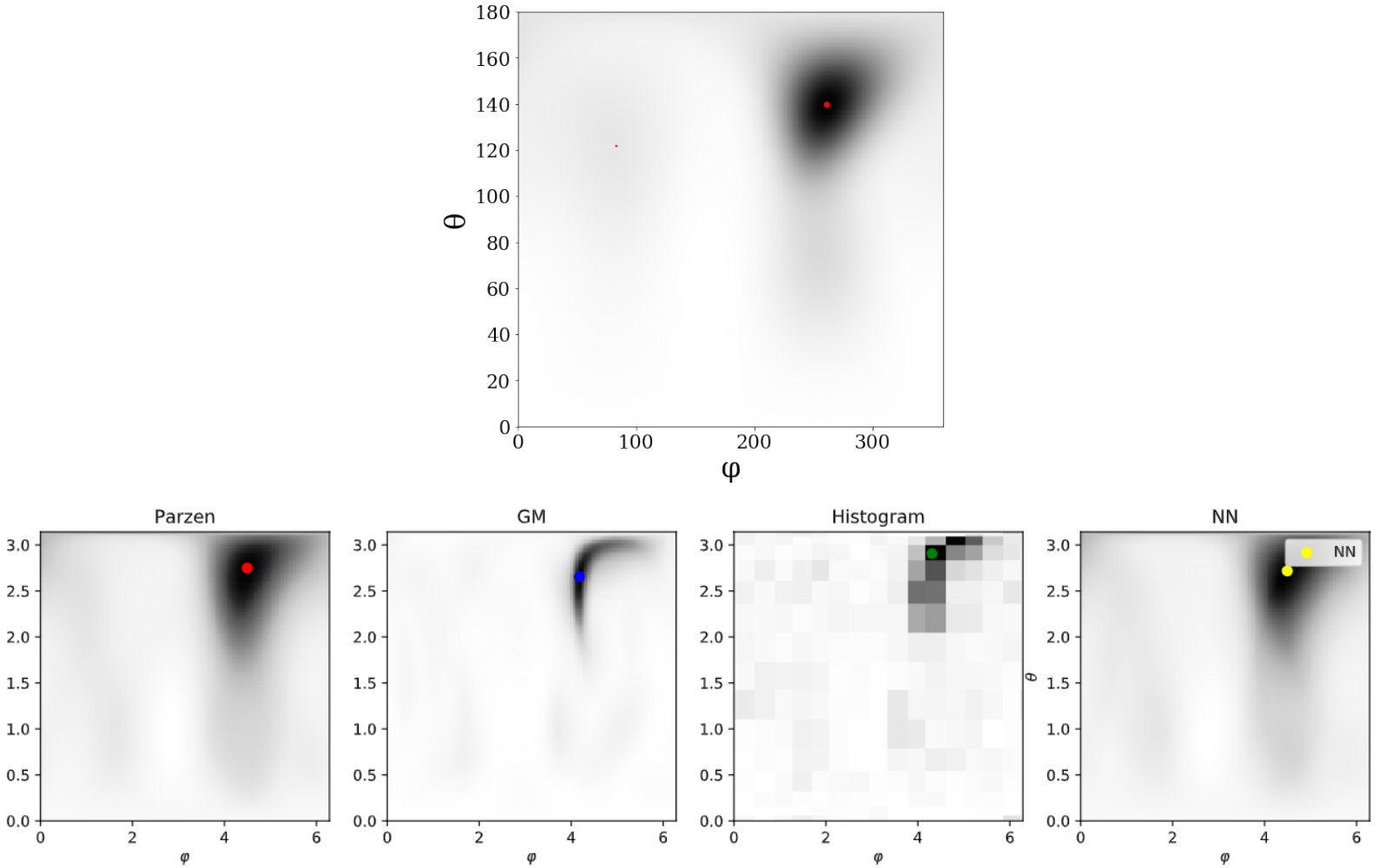
Двумерное полутоновое изображение восстановленной плотности $\hat{p}(r = 7\text{Å}, \theta, \varphi)$; красная точка соответствует максимуму, попавшему в диапазон $r = 7 \pm 0.5$.

Результаты восстановления, пара $0-2$, $r = 7$Å



Трёхмерное изображение восстановленной плотности $\hat{p}(r = 7\text{Å}, \theta, \varphi)$: в виде графика функции переменных (θ, φ) (слева) и в виде поверхности (справа).

Соответствие результатам простых моделей, $r = 7$Å



Соответствие восстановленной плотности (сверху) результатам, полученным с помощью других моделей восстановления (снизу).

Выносится на защиту

1. Предложен алгоритм нахождения параметров смеси распределений Кента для моделирования параметров химической связи пары аминокислота-лиганд
2. Проведен анализ восстановленных плотностей, установлено соответствие найденных максимумов с результатами, полученными с помощью более простых моделей.

Цель эксперимента и данные

Цель

Сравнить качество прогнозов, полученных предложенными алгоритмами согласования, с качеством независимых прогнозов и согласованных прогнозов, полученных при помощи существующих алгоритмов согласования, для различных типов иерархических структур.

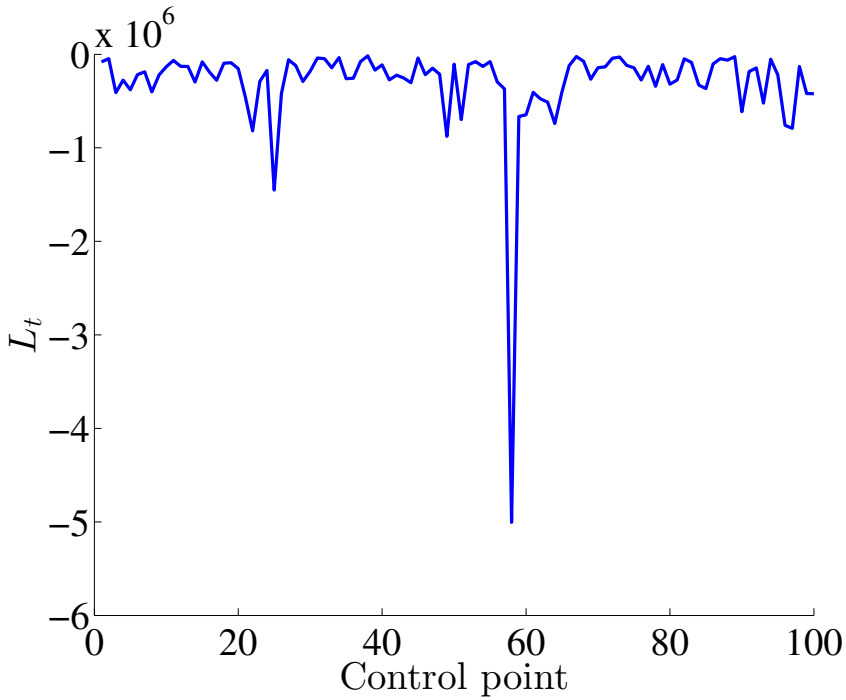
Данные

- **Трехуровневая иерархия:** данные о посуточной загруженности узлов РЖД. 37 типов грузов, 98 ЖД веток.
- **Двухуровневая иерархия:** данные о почасовом потреблении электроэнергии в 20 регионах Канады (Global Energy Forecasting Competition 2012).

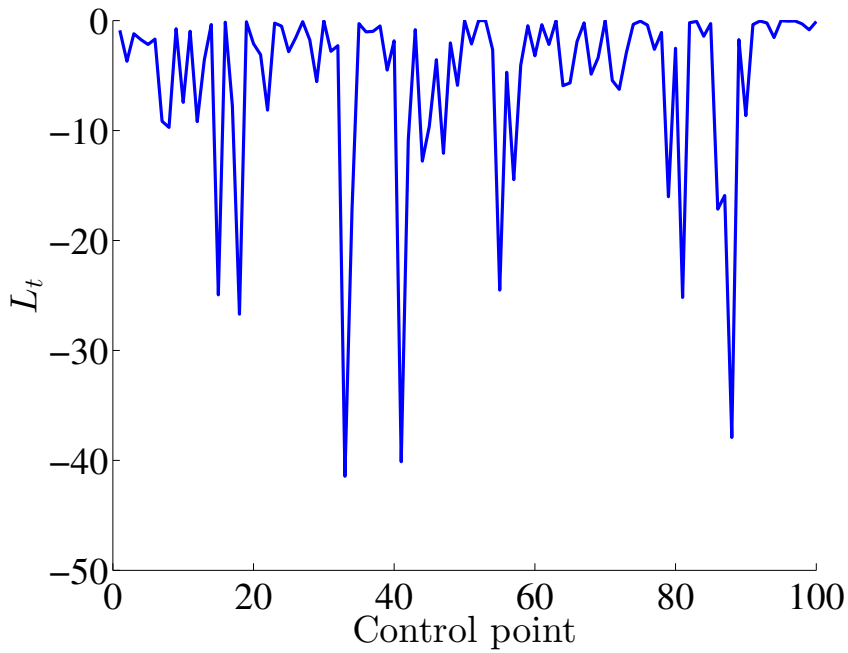
Для согласования прогнозов $H = 100$ последних точек истории решалась оптимизационная задача $\hat{\varphi} = \arg\min_{\chi \in \mathcal{A} \cap \mathcal{B}} \|\chi - \hat{\chi}\|_2^2$.

Изображена величина
$L_t = \|\chi_t - \hat{\varphi}\|_2^2 - \|\chi_t - \hat{\chi}\|_2^2, \quad t = (T - H + 1), \ldots, T.$

Во всех контрольных точках потери уменьшились.



Для РЖД

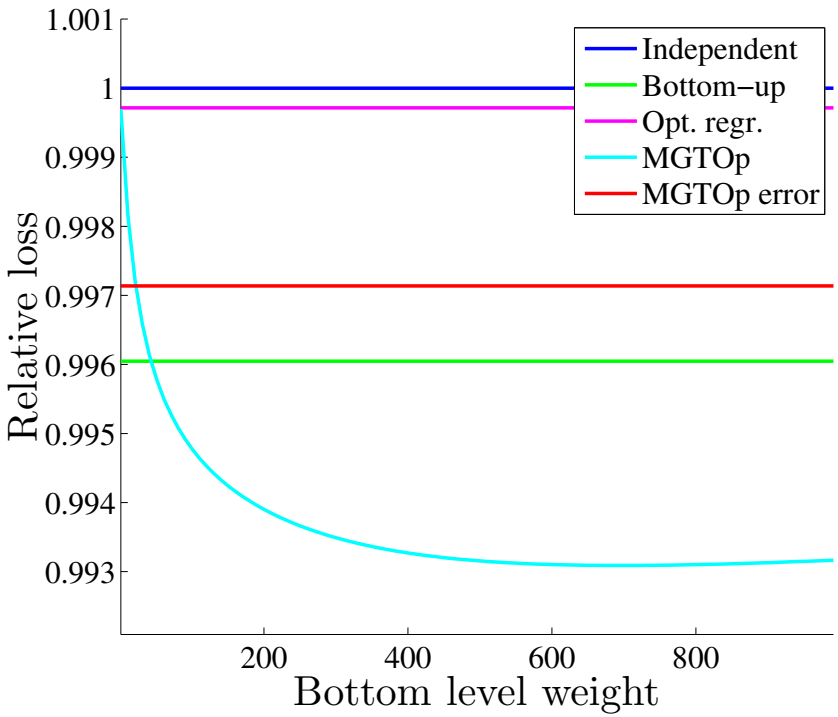


Для электроэнергии

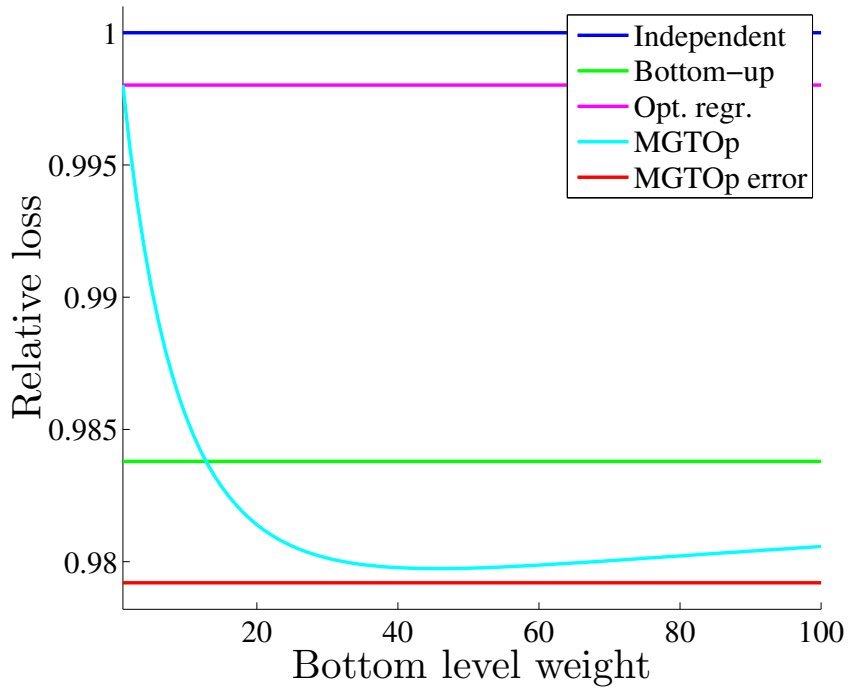
Сравнение качества согласованных и независимых прогнозов

$$\text{Relative loss}(\textit{algorithm}) = \frac{\sum_{t=T-H+1}^{T} \|\chi_t - \hat{\varphi}_{\textit{algorithm}}\|_2^2}{\sum_{t=T-H+1}^{T} \|\chi_t - \hat{\chi}\|_2^2}.$$

$$\hat{\varphi}_{\text{теор.игр.}} = \arg\min_{\chi \in \mathcal{A} \cap \mathcal{B}} l_r(\chi, \hat{\chi}), \quad l_r(\chi, \hat{\chi}) = \sum_{i=1}^{d} w_i \left(\chi(i) - \hat{x}(i)\right)^2.$$



Для РЖД



Для электроэнергии

Функция потерь $l_h(\boldsymbol{\chi}_t, \hat{\boldsymbol{\chi}}) = \|\boldsymbol{\chi}_t - \hat{\boldsymbol{\chi}}\|_2^2$.

Средние потери прогнозирования отгрузки в узлах РЖД, $\times 10^8$

Уровень иерархии	Независимые прогнозы	Восходящее согласование[1]	Оптимальная регрессия[2]	Модиф. теор.-игр. согл. (веса 700)
Вся иерархия	10.038	9.999	10.035	9.969
Верхний уровень	2.858	2.868	2.856	2.840
Средний уровень, ветки	2.549	2.486	2.545	2.487
Средний уровень, грузы	2.338	2.351	2.340	2.348
Нижний уровень	2.294	2.294	2.294	2.294

Средние потери прогнозирования потребления электроэнергии

Уровень иерархии	Независимые прогнозы	Восходящее согласование	Оптимальная регрессия	Модиф. теор.-игр. согл. (погрешности)
Вся иерархия	2727	2683	2722	2670
Верхний уровень	2083	2039	2076	2029
Нижний уровень	644	644	646	642

[1] Albert B. Schwarzkopf, Richard J. Tersine, John S. Morris *Top-down versus bottom-up forecasting strategies.* The International Journal Of Production Research, 26(11):1833—1843, 1988.

[2] Rob J. Hyndman, Roman A. Ahmed, George Athanasopoulos, Han Lin Shang. *Optimal combination forecasts for hierarchical time series.* Computational Statistics and Data Analysis, 55(9):2579–2589, 2011.

Выносится на защиту

1. Предложен алгоритм теоретико-игрового оптимального согласования прогнозов иерархических временных рядов, который сводит задачу согласования к задаче оптимизации.
2. Доказано, что алгоритм не требует несмещенности независимых прогнозов и оценок их погрешностей. На независимые прогнозы накладываются только физические ограничения.
3. Показано, что алгоритм позволяет работать с иерархическими структурами любой сложности.
4. Результаты экспериментов подтверждают, что качество прогнозов, согласованных с помощью предложенного алгоритма и его модификации, превосходит качество независимых прогнозов, а также качество согласованных прогнозов, полученных с помощью существующих алгоритмов согласования.

Вычислительный эксперимент

Цели:

1. Изучение зависимости качества скоринга от объема обучающей выборки и от ядра сглаживания гистограм признаков
2. Сравнение качества скоринговой функции с лучшими существующими методами

Данные:

- Модельные структуры с соревнований CASP[5-11]
- По 300 NMA моделей белков для каждой нативной из CASP в RMSD диапазоне $[0.5, 6]\text{A}$ на 100 первых нормальных модах

Зависимость качества скоринговой функции от сглаживания

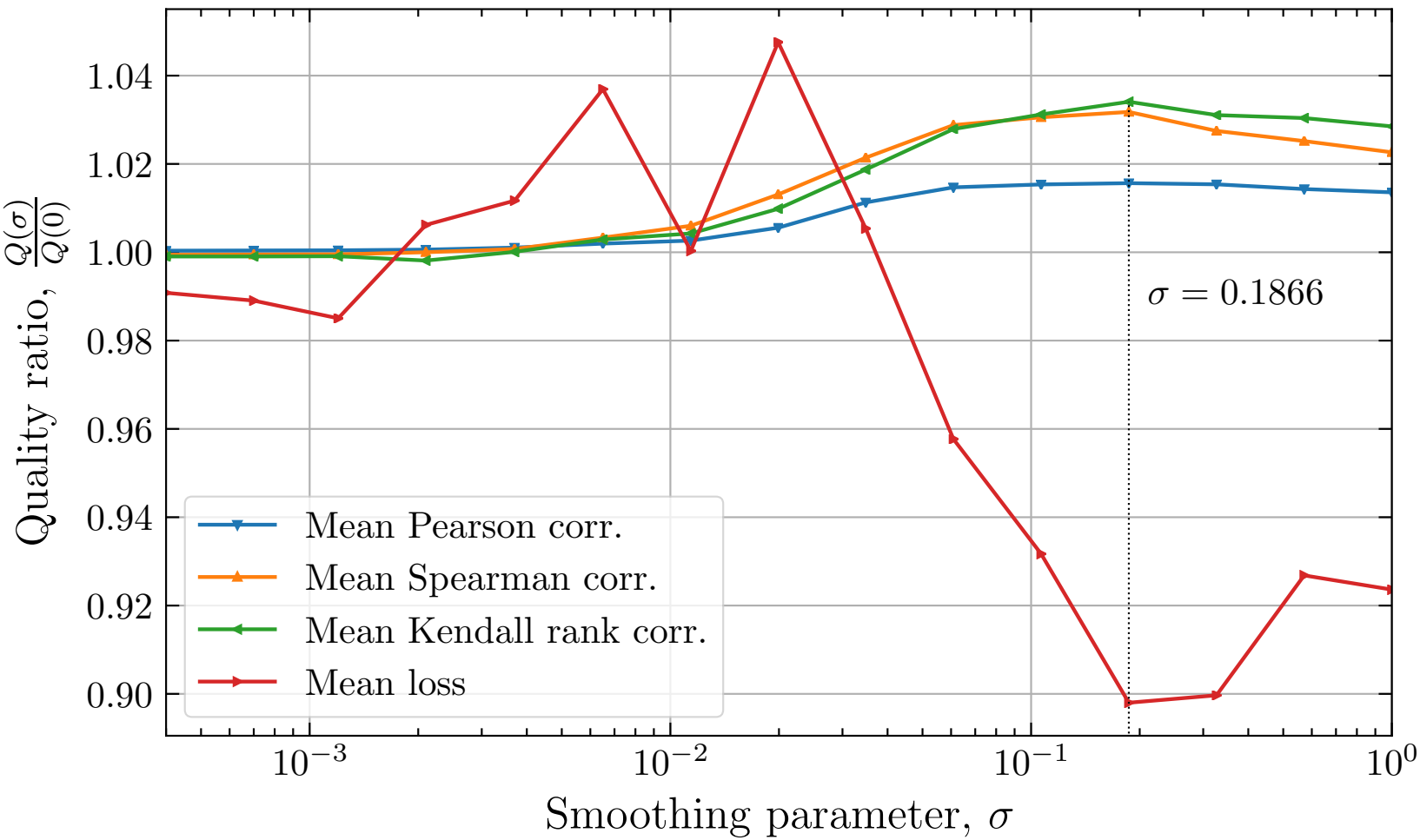


Рис.: Оценка качества структур на выборке CASP10 (stage1 и stage2 вместе) от ширины ядра сглаживания $\sigma^a = \sigma^r = \sigma^h = \sigma^s = \sigma$ при обучении на выборках CASP[5-9] без сглаживания ($\sigma = 0$).

Исследование скоринговой функции

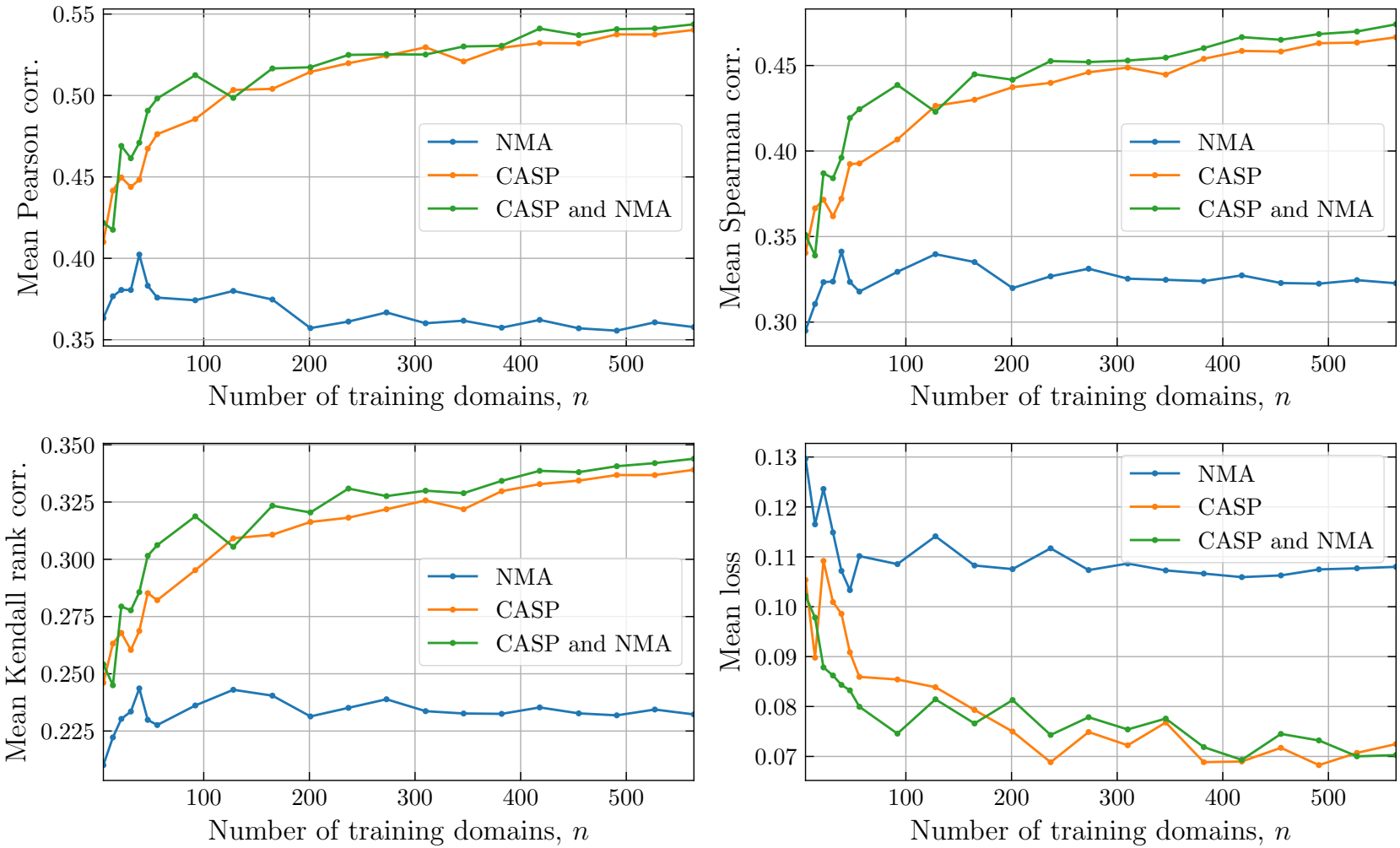


Рис.: Зависимость качества скоринговой структур от объема обучающей выборки. Обучение: случайные подвыборки CASP[5-10]. Контроль: CASP11 (stage1 и stage2 вместе).

Сравнение

QA Method	CASP11 Stage1			CASP11 Stage2		
	Loss	PCC	SCC	Loss	PCC	SCC
This study	**0.083**	0.645	0.522	**0.057**	**0.441**	**0.426**
ProQ2	0.090	0.643	0.506	0.058	0.372	0.366
VoroMQA	0.108	0.561	0.426	0.069	0.401	0.386
Wang-SVM	0.109	**0.655**	**0.535**	0.085	0.362	0.351
Dope	0.111	0.542	0.416	0.077	0.304	0.324
RWplus	0.135	0.536	0.433	0.084	0.295	0.314

Таблица: Качество ранжирования структур выборки CASP11. Метрики качества: Mean metric loss (Loss), коэффициент корреляции Пирсона и Спирмана (PCC и SCC) между оценками качества структур разными методами и функцией близости $\rho_{\text{GDT-TS}}$. Обучение: CASP[5-10].

Вычислительный эксперимент

Эволюционное семейство моделей:

1. Базовая модель (BASE) не учитывает типы ребер, признаки атомов; не использует механизмы работы с несвязанными графами.
2. Модель расширенного молекулярного графа (EG). По сравнению с базовой моделью, использует расширенный молекулярный граф.
3. Модель Трансформер (T). По сравнению с базовой моделью, после сверточных слоев используется преобразование self-attention.
4. Модель EGT Использует обе предложенных модификации.
5. Модель EGTB использует разные типы ребер в соответствии с типом химической связи.
6. Модель EGTBF использует признаки атомов (валентность, заряд и тд).
7. Модель MT_EGTBF использует многозадачное обучение для двух рассматриваемых задач.

Сводная таблица резльтатов

	Product mapping		**Center detection**	
	FM	F_1	*FM*	F_1
BASE	0.21 ± 0.01	0.92 ± 0.002	0.15 ± 0.01	0.502 ± 0.002
EG	0.45 ± 0.01	0.943 ± 0.002	0.40 ± 0.01	0.714 ± 0.002
T	0.36 ± 0.01	0.938 ± 0.002	0.29 ± 0.01	0.643 ± 0.002
EGT	0.47 ± 0.01	0.946 ± 0.002	0.43 ± 0.01	0.731 ± 0.002
EGTB	0.53 ± 0.01	0.950 ± 0.002	0.55 ± 0.01	0.809 ± 0.002
EGTBF	0.59 ± 0.01	0.959 ± 0.002	0.60 ± 0.01	0.838 ± 0.002
MT_EGTBF	0.60 ± 0.01	0.963 ± 0.002	0.61 ± 0.01	0.841 ± 0.002

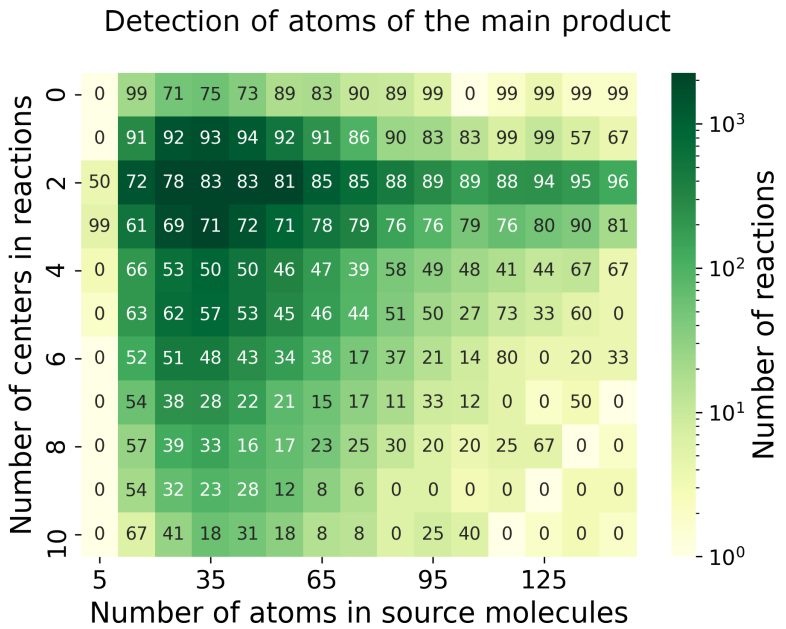
FM среднее значение точности полного совпадения (1, если все метки атомов в реакции предсказаны верно, 0 иначе). F_1 среднее значение F_1-меры между предсказанными и правильными метками атомов в реакции.

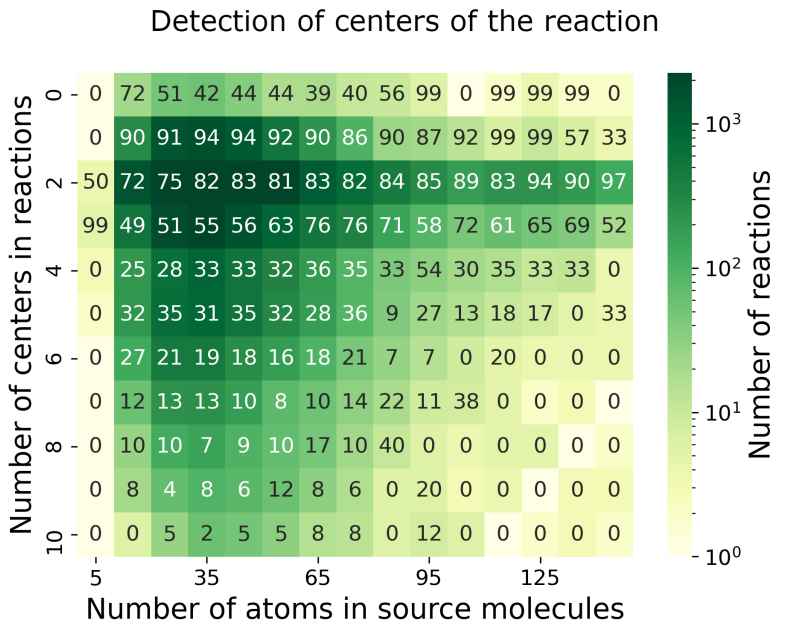
Вывод

Предложенные методы работы с несвязанными графами значительно улучшают качество модели. Использование признаков вершин и ребер молекулярного графа приводит к повышению качества.

Анализ ошибки

На графике представлена совместная зависимость качества модели от количества центров и длины исходных молекул. Цветом указано распределение исходных данных.



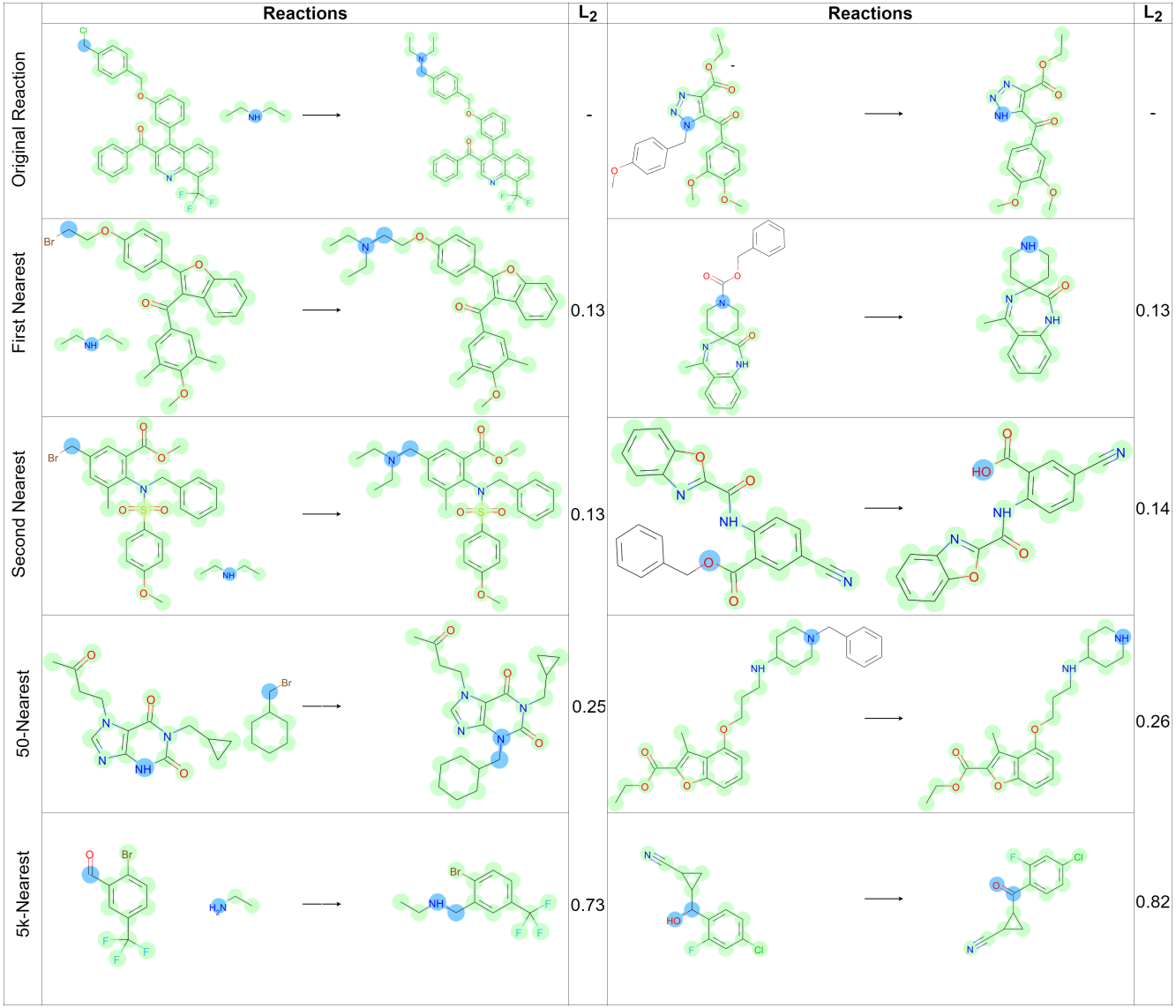


50 The number means full-match accuracy

The color means the number of reactions

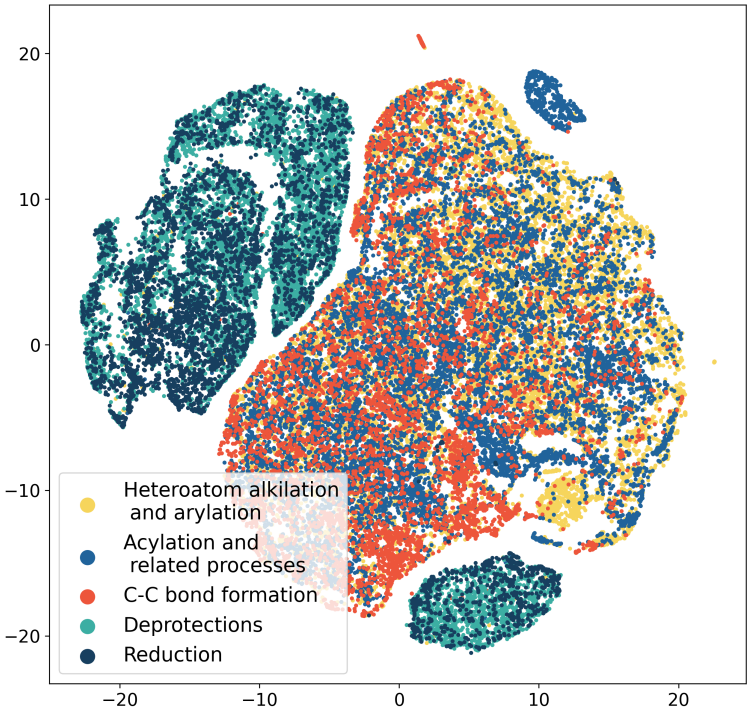
1. С увеличением числа центров качество падает;
2. Зависимость качества от длины исходных молекул выражена слабо.

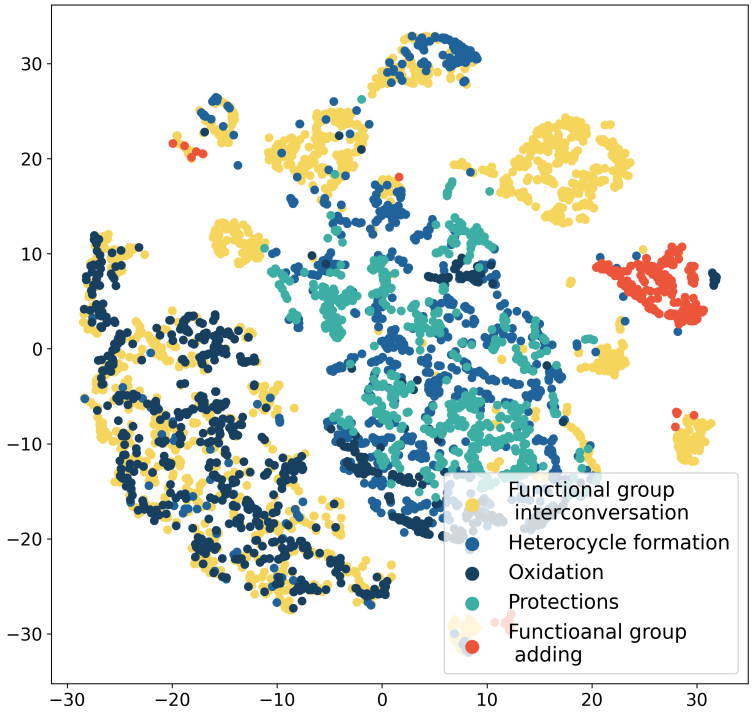
Исследование свойств пространства реакций



Метрически близким векторам, соответствующим состояниям всей реакции, соответствуют реакции с похожим механизмом.

T-SNE карты реакций





Для датасета USPTO_50k (10 классов реакций) построены T-SNE проекции в двумерное пространство.
Кластеры векторов состояний реакции скоррелированы с разметкой по классам химических реакций.

Результаты выносимые на защиту

1. Сформулирована задача предсказания продуктов химической реакции в терминах классификации вершин несвязанного графа.
2. Предложено обобщение графовых нейронных сетей для работы с несвязанными графами.
3. Предложена последовательность вычислительных экспериментов, демонстрирующая необходимость каждой предложенной модификации.
4. Проанализирована полученная модель, исследованы свойства векторных состояний химической реакции, формируемых в модели.

Материалы

1. GitHub репозиторий с задокументированными исходными файлами вычислительных экспериментов.
2. Web-интерфейс предложенной модели, интерактивное представление t-SNE карт реакций.

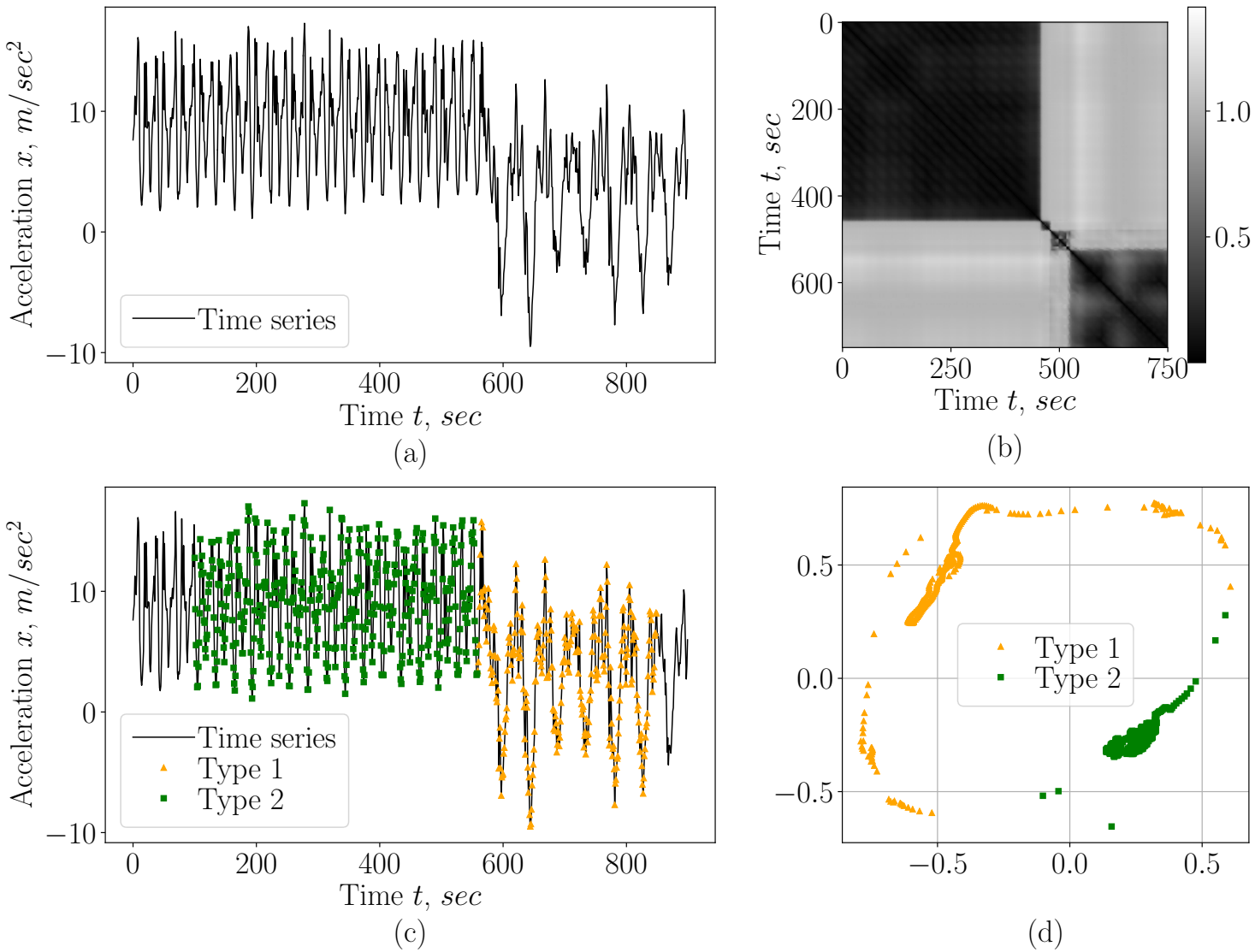
Результаты эксперимента

- Physical Motion — ряды получены при помощи мобильного акселерометра. Характерные действия: ходьба, бег, приседания.
- Synthetic — синтетические временные ряды.

Ряд, $\mathbf{x}$	Длина, N	Сегментов, K	Период, T	Ошибка, S
Physical Motion 1	900	2	50	0.03
Physical Motion 2	900	2	35	0.08
Physical Motion 3	900	2	30	0.09
Physical Motion 4	800	2	50	0.01
Synthetic 1	2000	3	40	0.008
Synthetic 2	2000	2	40	0.06
Synthetic 3	2000	2	40	0.03
Synthetic 4	2000	2	40	0.03
Synthetic 5	2000	2	40	0.04
Simple	1000	2	135	0.14

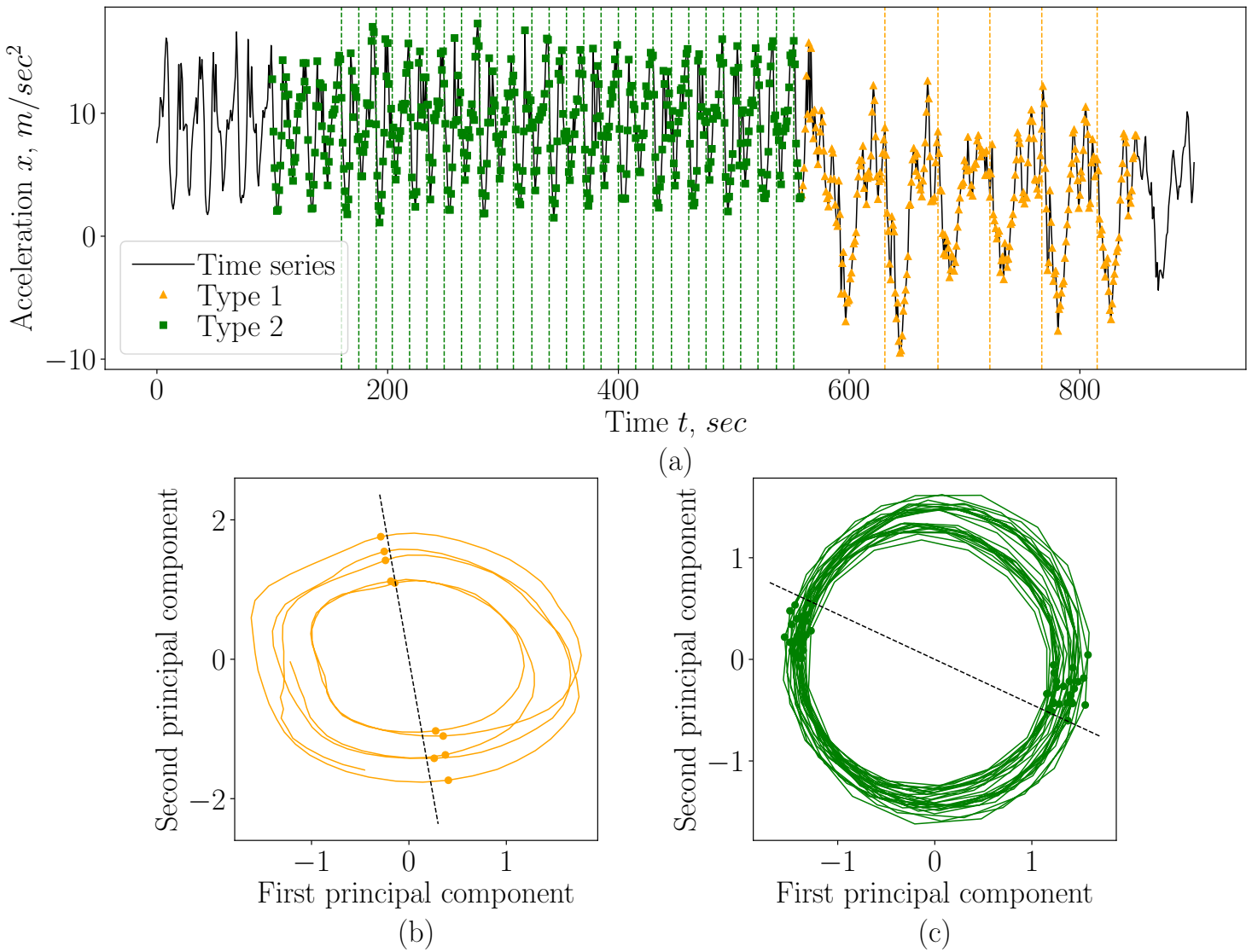
- N — число точек во временном ряде,
- K — число различных действий во временном ряде,
- T — максимальная длина сегмента,
- S — точность кластеризации.

Пример кластеризации точек временного ряда



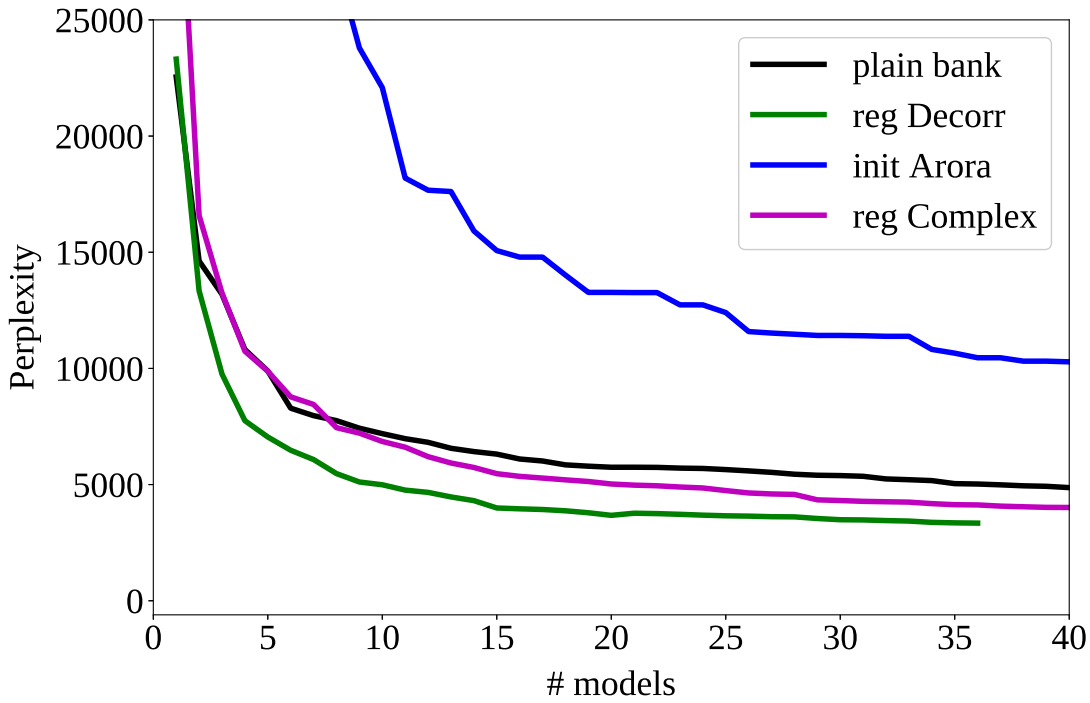
a) начальный временной ряд; b) матрица попарных расстояний; c) кластеризация точек ряда; d) Multidimential Scaling для матрицы попарных расстояний.

Пример сегментации временного ряда



a) сегментация ряда; b) фазовая траектория для второго действия; c) фазовая траектория для первого действия.

Зависимость банка тем от числа обучаемых моделей

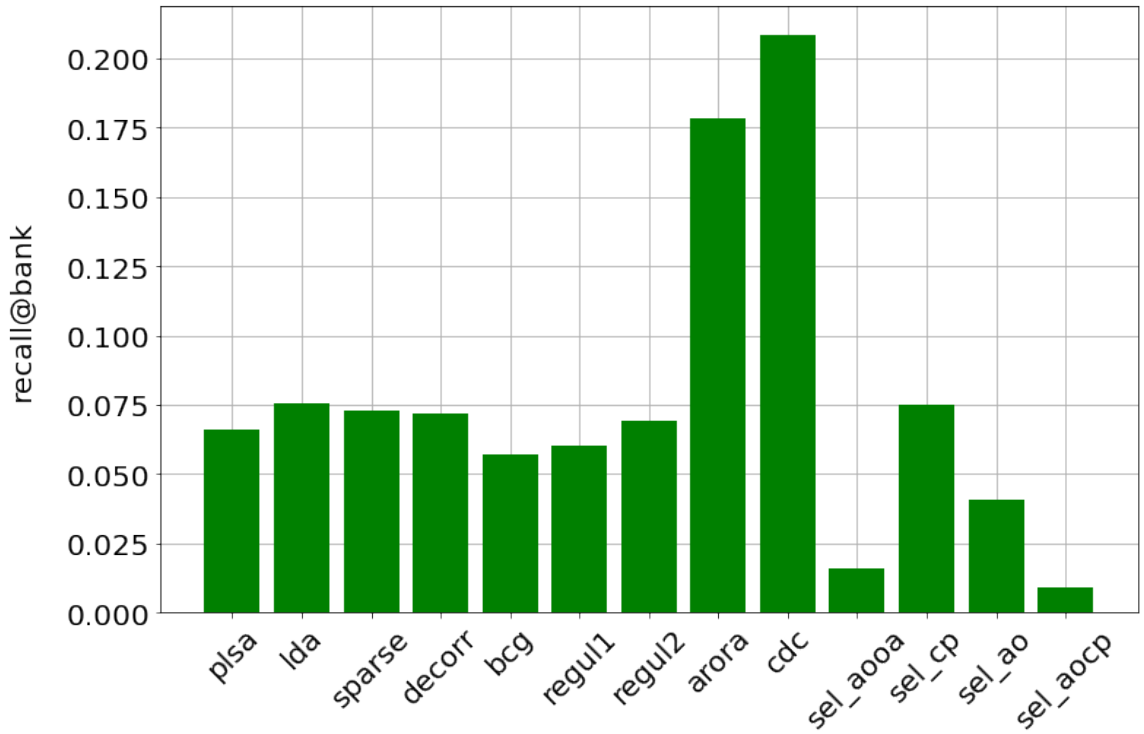


- Наилучшая перплексия — при декоррелировании.
- При добавлении моделей число тем в банке постепенно увеличивается; скорость же пополнения банка снижается.

Vorontsov K. et al. *Additive regularization of topic models*, 2015.
Arora S. et al. *Learning topic models–going beyond SVD*, 2012.
Hofmann T. *Probabilistic latent semantic indexing*, 1999.

Результат, усреднённый по датасетам

Цель: по множеству датасетов $\mathcal{D}$ получить оценки качества моделей recall@bank(m), $m \in \mathcal{M}$.

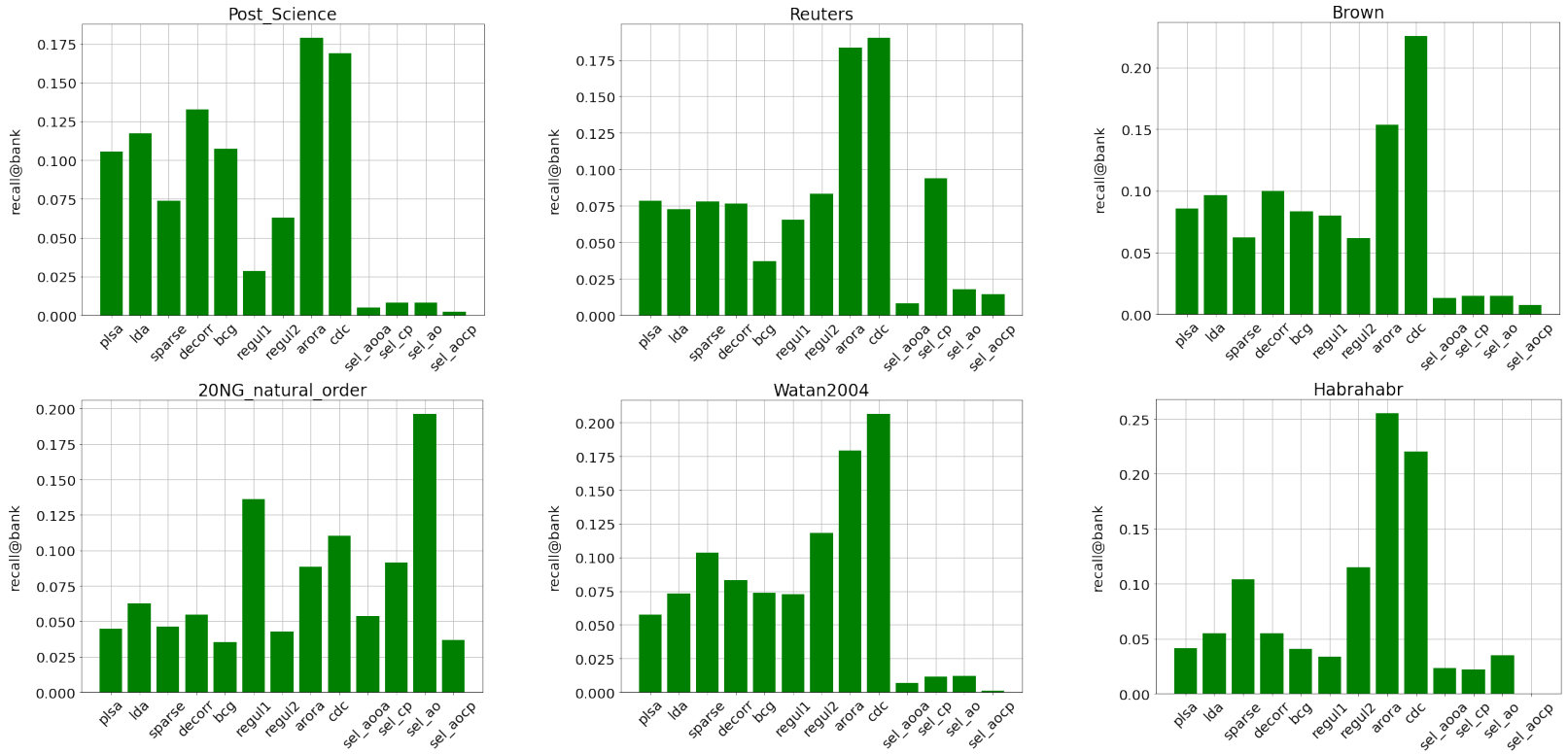


Вывод: банк тем помог из фиксированного ряда моделей $\mathcal{M}$ найти те модели, которые лучше подстраиваются под данные.

Dobrynin V. et al. *Contextual document clustering*, 2004.

Blei D. et al. *Latent dirichlet allocation*, 2003.

Результаты по разным датасетам



Вывод: под каждый набор данных нужна своя модель, но в большинстве случаев модели с неслучайной инициализацией превосходят по качеству остальные модели из $\mathcal{M}$.

---

habr.com, postnauka.ru, nltk.org/book/ch02.html
sites.google.com/site/mouradabbas9/corpora
scikit-learn.org/0.19/datasets/twenty_newsgroups.html

Результаты, выносимые на защиту

- Предложен алгоритм создания банка тем с использованием многократного обучения моделей
- Предложена методика оценивания качества тематических моделей с помощью банка тем
- Реализована система для использования банка тем[1]

Публикации

- Alekseev V. et al. *TopicNet: Making Additive Regularisation for Topic Modelling Accessible*. LREC, 2020.[2]
- Alekseev V. et al. *Topic Modelling for Extracting Behavioral Patterns from Transactions Data*. IEEE, 2019.[3]
- Alekseev V. et al. *Intra-Text Coherence as a Measure of Topic Models' Interpretability*. Dialogue, 2018.[4]

[1] github.com/machine-intelligence-laboratory/OptimalNumberOfTopics
[2] aclweb.org/anthology/2020.lrec-1.833
[3] ieeexplore.ieee.org/abstract/document/9007329
[4] dialog-21.ru/media/4281/alekseevva.pdf