Эмпирическое моделирование и прогноз динамики социальной активности в условиях пандемии

«Моделирование эпидемий вирусных инфекций»

РФЯЦ-ВНИИТФ РОСАТОМ
ВНИИТФ

ИВМ РАН

ИПМ РАН

ИПФ РАН

НИЦЭМ

ФГБУ «НМИЦ ФПИ» МИНЗДРАВА РОССИИ
НМИЦ ФПИ

Модели из первых принципов

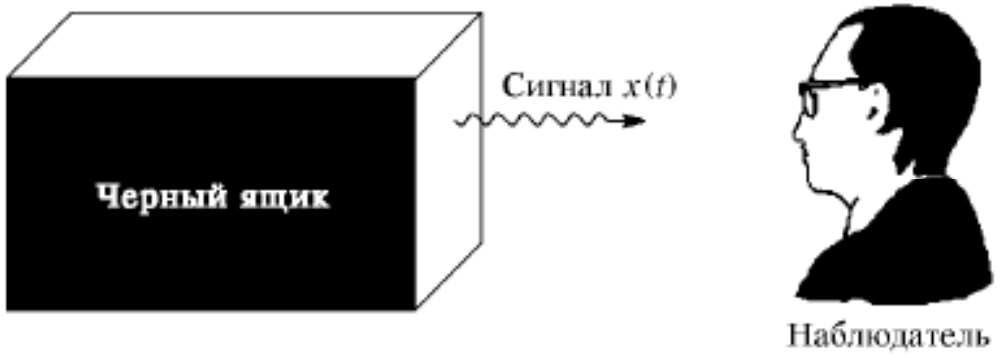
- Возможность исследовать отклик на разные внешние воздействия.
- Возможность исследовать устойчивость поведения модели, зависимость от начальных условий.

- Высокая сложность
- Чувствительность к граничным и начальным условиям
- Нет прямой связи с реальными данными

Эмпирические модели

- Прямая связь с реальными данными

- Неочевидны физические механизмы, которые лежат в основе динамики модели



Модели прогнозирования временных рядов (регрессионные модели, модели машинного обучения, модели фильтрации)

Модели, основанные на дифференциальных уравнениях (SIR, SIER и др.)

Модели эпидемий

ГИС-моделирование

Агент-ориентированное моделирование

Модели прогнозирования временных рядов

- Регрессионные модели:

Неадаптивные регрессионные модели, адаптивные авторегрессионные: ARMA (Autoregressive Moving Average), ARIMA (Autoregressive Integrated Moving Average).

- Модели машинного обучения

Искусственные нейронные сети, скрытые марковские модели.

- Модели фильтрации

Вейвлет-декомпозиция, экспоненциальное сглаживание, фильтр Кальмана

\+ Возможность сделать хороший краткосрочный прогноз (временные ряды слишком коротки для составления какого-либо адекватного долгосрочного прогноза).

– Не дают представления о каналах и способах распространения.

ГИС-моделирование

пространственный автокорреляционный анализ

Рис. 2. Трансформация диаграммы рассеивания П. Морана по числу зараженных COVID-19 за период с марта по июнь 2020 г.

\+ Оценка *пространственной неоднородности* коронавирусной инфекции, поиск *полюсов* ее *роста*, формирующихся *пространственных кластеров* и *зон* их *влияния* с оценкой межтерриториальных взаимосвязей.

– Неоднородность полученных результатов в зависимости от использования разных матриц пространственных весов.

Модели, основанные на дифференциальных уравнениях

- **Логистическое уравнение:** $\frac{dI}{dt} = rI(1 - \frac{I}{N})$

$I(t)$ – количество инфицированных, N – размер популяции, r – скорость роста эпидемии.

- **Компартментные (камерные) модели:**

Модели из первых принципов: закон сохранения масс и особенности передачи инфекции.

SIR-модель

Susceptible
Восприимчивые к заражению

Infectious
Инфицированные

Recovered
Вылечившиеся и умершие

$$\frac{dS}{dt} = -\alpha\frac{SI}{N}$$

$$\frac{dI}{dt} = \alpha\frac{SI}{N} - \beta I$$

$$\frac{dR}{dt} = \beta I$$

$$N = S(t) + I(t) + R(t)$$

$$\mathcal{R}_0 = \frac{\alpha}{\beta}$$

индекс репродукции (заразности)

SEIR-модель

Susceptible
Восприимчивые к заражению

Exposed
Зараженные незаразные

Infectious
Зараженные заразные

Recovered
Вылечившиеся и умершие

$$\frac{dS}{dt} = -\alpha\frac{SI}{N}$$

$$\frac{dE}{dt} = \alpha\frac{SI}{N} - \gamma E$$

$$\frac{dI}{dt} = \gamma E - \beta I$$

$$\frac{dR}{dt} = \beta I$$

$$N = S(t) + \mathrm{E}(t) + I(t) + R(t)$$

\+ Учет особенностей инфекции (инкубационный период, иммунитет).

– Сложно оценить межтерриториальные взаимоотношения. Учет ограничительных мер и мутаций вируса затруднен.

Агент-ориентированные модели

- Статистическое моделирование характерного поведения различных групп населения
- Возрастные и профессиональные группы:
  - школьники и дошкольники,
  - студенты,
  - работники предприятий и офисов,
  - работники сферы жизнеобеспечения,
  - работники сферы обслуживания,
  - пенсионеры.
- Зоны контакта (ячейки):
  - квартира (дом),
  - место работы или учебы,
  - транспорт,
  - магазины и торговые центры (3 типа – гипермаркеты , супермаркеты, магазины шаговой доступности).

\+ Возможность явного учета ограничительных мер, появления новых штаммов. Хорошие результаты как краткосрочного, так и долгосрочного прогнозов.

– Для качественного моделирования территории нескольких регионов или в масштабе целой страны необходимо собрать большой объем данных.

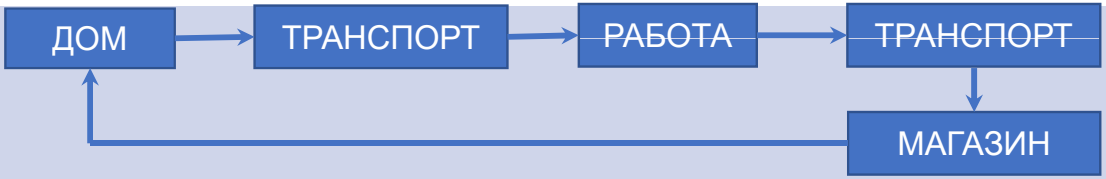
Агент-ориентированные модели

Сценарии для жителей города

Для каждого жителя города каждый день выполняется моделирование посещений закрепленных за ним ячеек, в которых вычисляется количество вирусоносителей.

Группы населения в агентной модели

Группы населения	Количество, %
школьники и дошкольники	21
студенты	10
работники предприятий и офисов	24
работники сферы обслуживания	25
работники сферы жизнеобеспечения	3
пенсионеры	17

Социальный статус	Схема посещений закрепленных за каждым человеком ячеек
РАБОТНИК	
ШКОЛЬНИК	
ПЕНСИОНЕР	

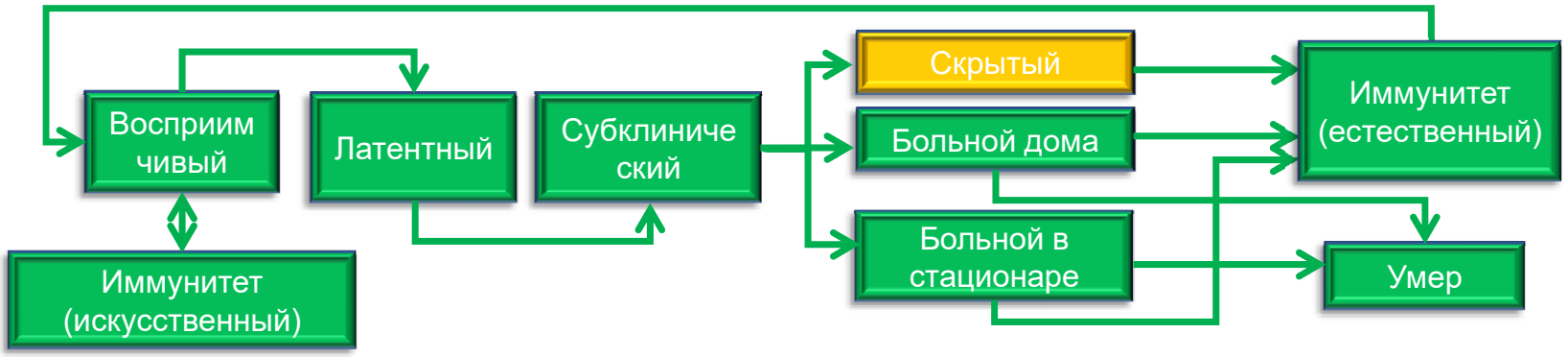
Моделирование заражения жителя

Вероятность заражения жителя в закрепленных за ним ячейках определяется таблицей времен пребывания и параметрами ячеек

$$p(t) = 1 - e^{-\frac{p_{зар}(N_{\inf}, S, R_{\inf})}{T_0} t}$$ - вероятность заражения в ячейке

Социальная группа	Время пребывания в ячейке в течении дня, часы			
РАБОТНИК	ДОМ	ТРАНСПОРТ	РАБОТА	МАГАЗИН
	12	2	9	1
ШКОЛЬНИК	ДОМ	ШКОЛА		
	15	9		
ПЕНСИОНЕР	ДОМ	ТРАНСПОРТ	МАГАЗИН	
	21	2	1	

Фазы заболевания



Фаза болезни	Описание фазы	Период [день]
ВОСПРИИМЧИВЫЙ	Здоровый человек, может заразиться	-
ЛАТЕНТНЫЙ	Инфицирован, но не заразен	3
СУБКЛИНИЧЕСКИЙ	Заразен, но не проявляет симптомов болезни	2
СКРЫТЫЙ БОЛЬНОЙ	Больной не выявлен, заразен	4
БОЛЬНОЙ ДОМА	Больной изолированный дома, заразен для семьи	10
БОЛЬНОЙ В СТАЦИОНАРЕ	Изолированный больной в больнице	12
ИММУНИТЕТ	Здоровый человек, не может заразиться	180/120-δ

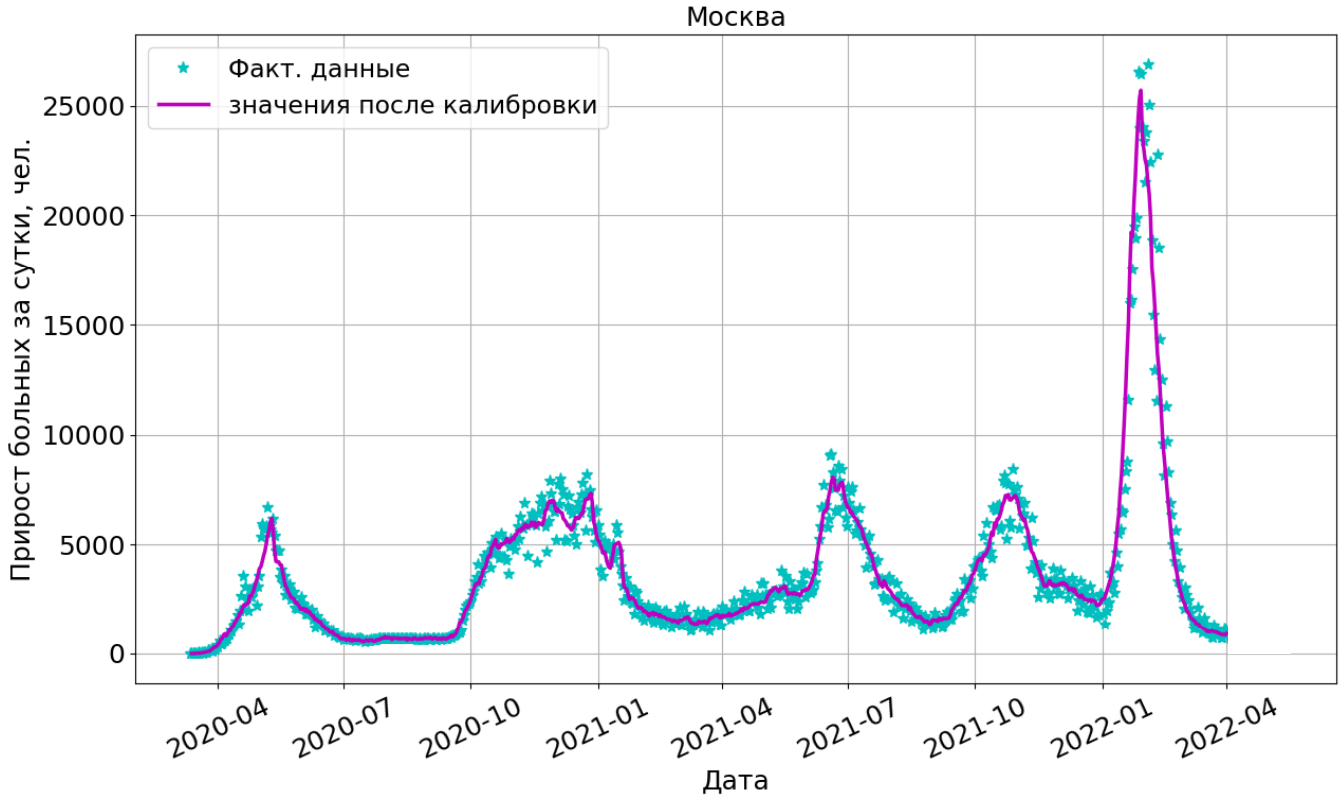
Входные данные для модели

- Параметры вируса
  - Длительность фаз заболевания
  - Контагиозность штамма
- Параметры городской среды
  - Структура ячеек
  - Площадь ячеек
- Параметры популяции
  - Разбиение на группы по социальному статусу
  - Схема посещения ячеек для каждого соц. статуса
- Активность населения
  - Общая активность
  - Активность по типу ячеек
- Количество заболевших за сутки в городе

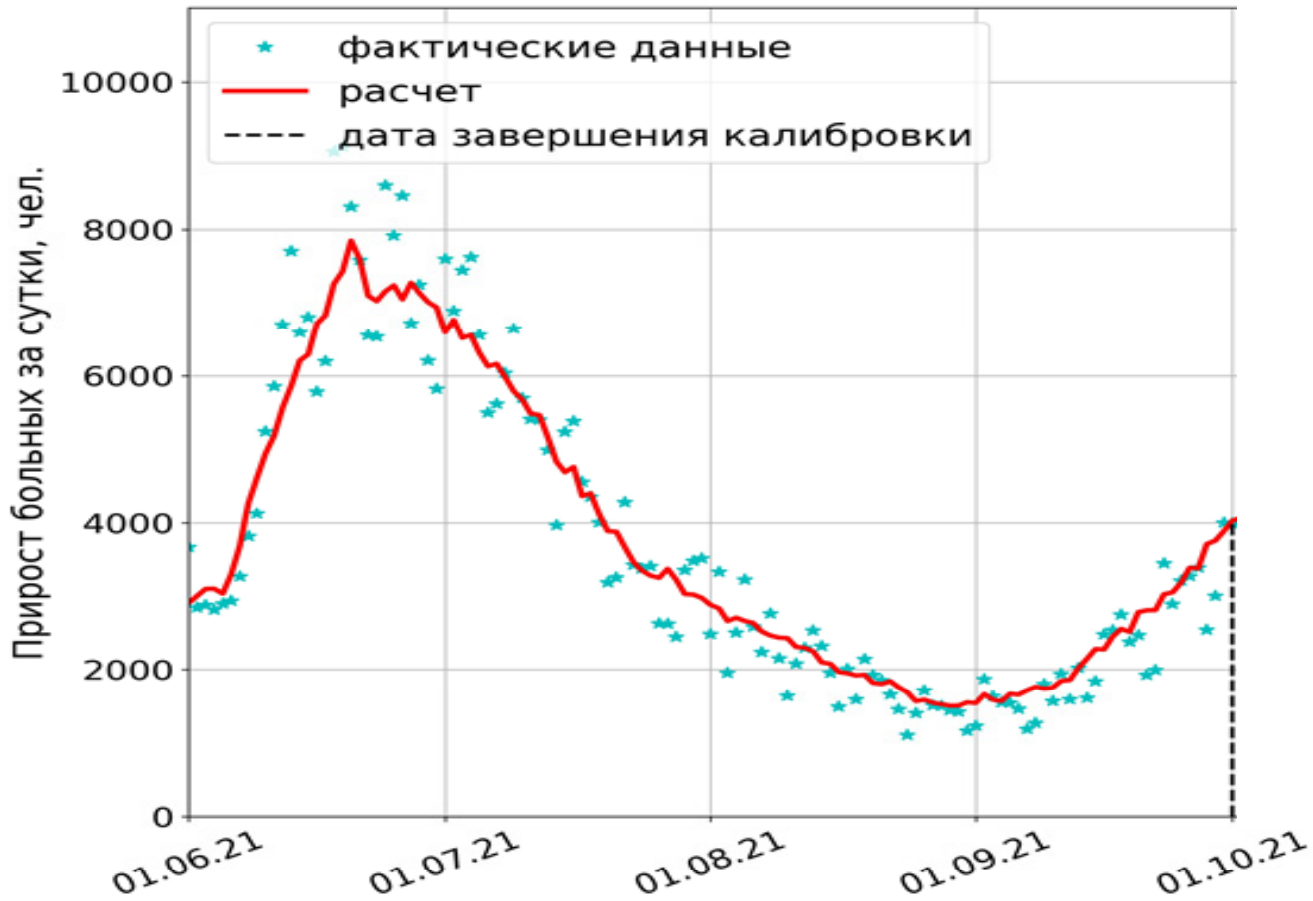
Вероятность выполнения суточного сценария жителя, иначе агент проводит 24 ч. в ячейке «дом»

Калибровка модели

- По факт. данным калибруется параметр модели «T_0»
- «T_0» – мат. ожидание времени нахождения в стандартной ячейке с одним больным, необходимого для заражения человека без иммунитета



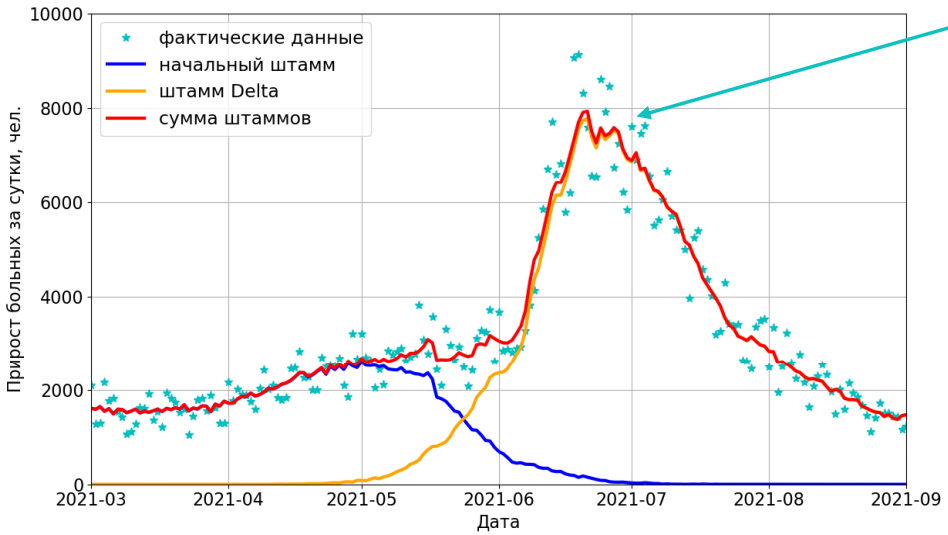
Поведение модели на интервале калибровки



Причины некорректного прогноза

Появление нового штамма

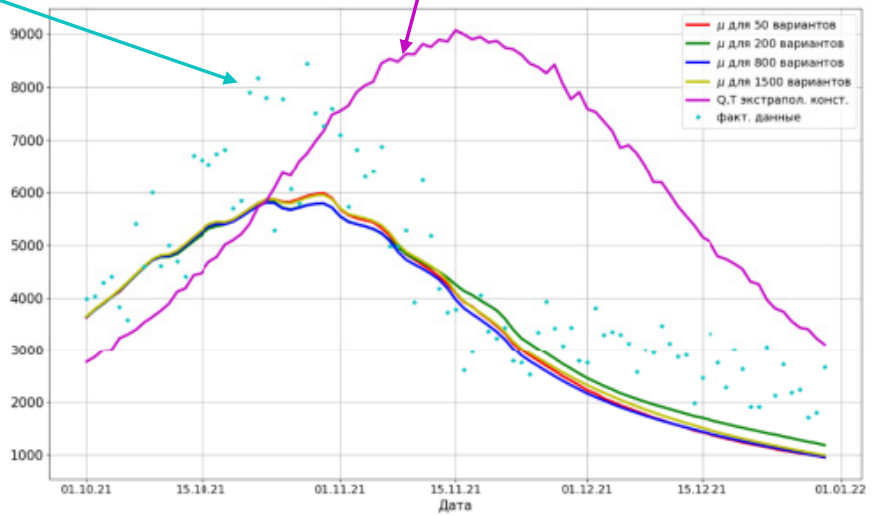
- Получение данных о новом штамме от вирусологов
- Корректировка входных параметров модели



Изменение активности населения

фактические данные

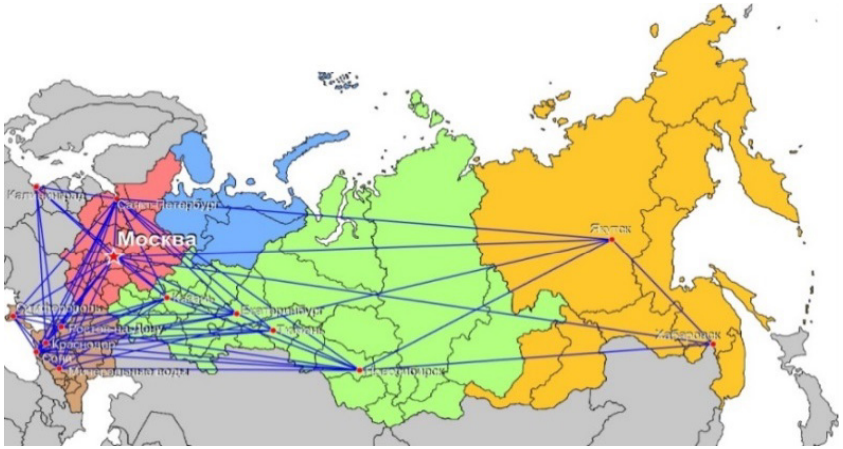
экстраполяция параметра активности населения константой



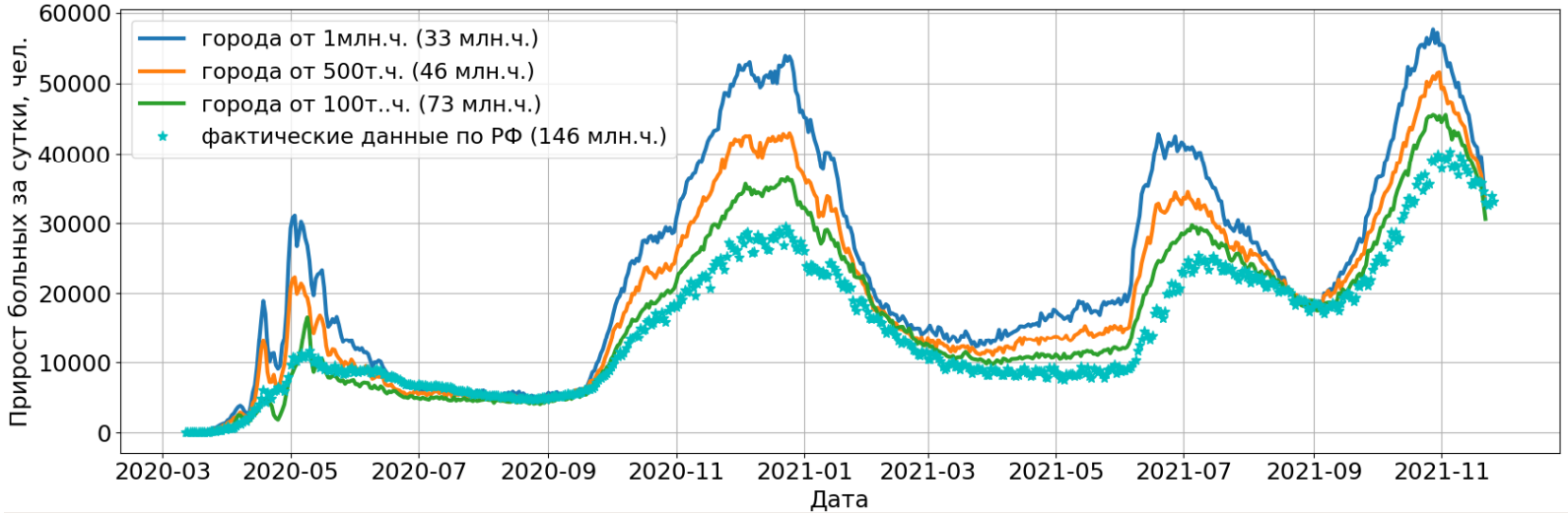
Моделирование страны

- В каждом городе локальная модель эпидемии
- Россия представляется сетью городов
- Учтены крупные транспортные узлы (ж.д., авиа)

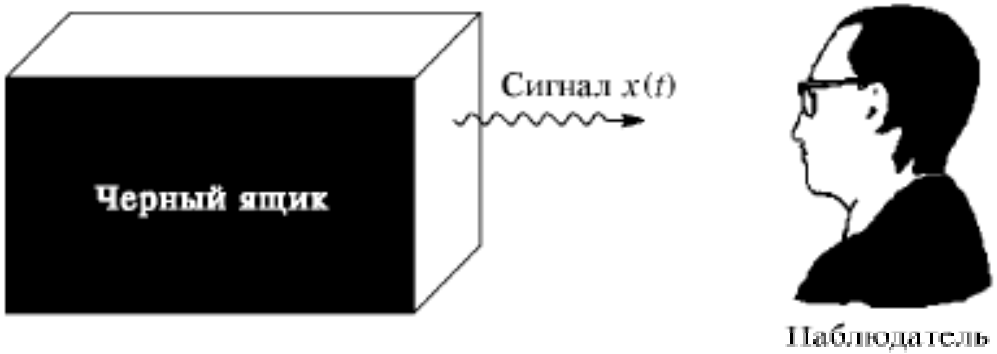
Для качественного моделирования территории нескольких регионов или в масштабе целой страны необходимо собрать большой объем данных.



Расчетные результаты приближаются к фактическим данным по мере уточнения структуры страны



Эмпирическое моделирование



- **Построение моделей систем напрямую по данным наблюдений**
- **В общем случае не предполагается наличие какой-либо информации об устройстве исследуемой системы, которая рассматривается как «черный ящик»**
- **Акутально в ситуации, когда уравнения из “первых принципов”, лежащие в основе исследуемой системы, неизвестны, либо не могут быть эффективно применены (живые системы, климат, социо-экономические системы)**

Эмпирическая модель

1. Фазовые переменные

Наблюдаемые данные

$\mathbf{x}_1, \dots, \mathbf{x}_N \quad \mathbf{x} \in \mathbb{R}^D$

Фазовые переменные

$\mathbf{y}_1, \dots, \mathbf{y}_N \quad \mathbf{y} \in \mathbb{R}^d$

$d < D$

2. Оператор эволюции

Стохастический оператор эволюции

$\mathbf{y}_{n-L}, \dots, \mathbf{y}_{n-1} \rightarrow \mathbf{y}_n \qquad (\mathbb{R}^{d \times L} \rightarrow \mathbb{R}^d)$

$\mathbf{y}_n = \mathbf{f}(\mathbf{y}_{n-L}, \dots, \mathbf{y}_{n-1}) + \hat{\mathbf{g}}\,\boldsymbol{\xi}_n$

Детерминированная часть

Стохастическая часть **(плохо разрешенные процессы), ξ** – гауссов, дельта-коррелирован

Типы параметризаций модели

$\mathbf{z}_n = (\mathbf{y}_{n-L}, \dots, \mathbf{y}_{n-1})$

Линейная параметризация

$\mathbf{f}(\mathbf{z}_n) = \mathbf{A}_n \mathbf{z}_n + \mathbf{B}_n \mathbf{c}_n,$

Нелинейная параметризация - нейронная сеть

$\mathbf{f}(\mathbf{z}_n) = \sum_{l=1}^{m} \boldsymbol{\alpha}_l \cdot \tanh(\boldsymbol{\omega}_l \mathbf{z}_n + \boldsymbol{\delta}_l \mathbf{c}_n + \gamma_l),$

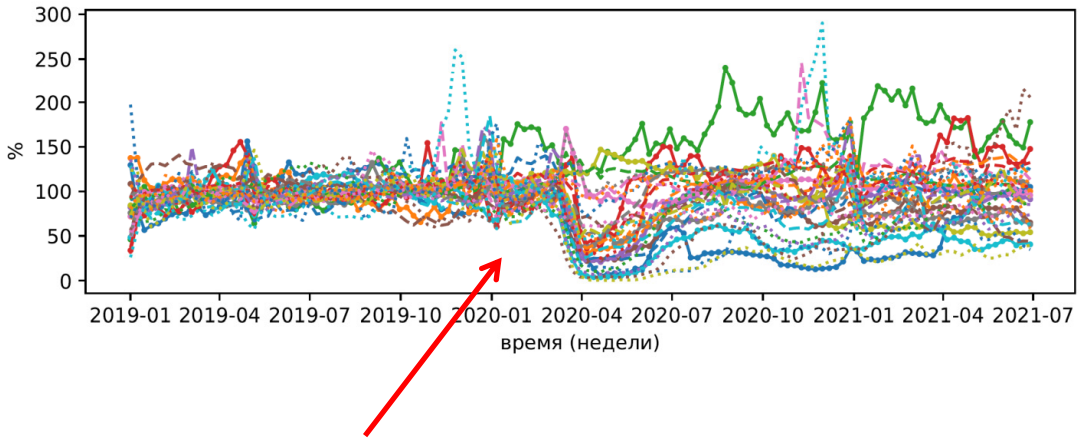
Внешнее воздействие

Описание данных и внешних воздействий (форсингов)

[1]**Данные:** временные ряды потребительской активности от банка Tinkoff на интервале с января 2019 года по июль 2021 года в 6 крупнейших городах Российской Федерации – Москве, Нижнем Новгороде, Калининграде, Екатеринбурге, Новосибирске, Хабаровске.

Размерность данных: D=32 категории поребительской активности

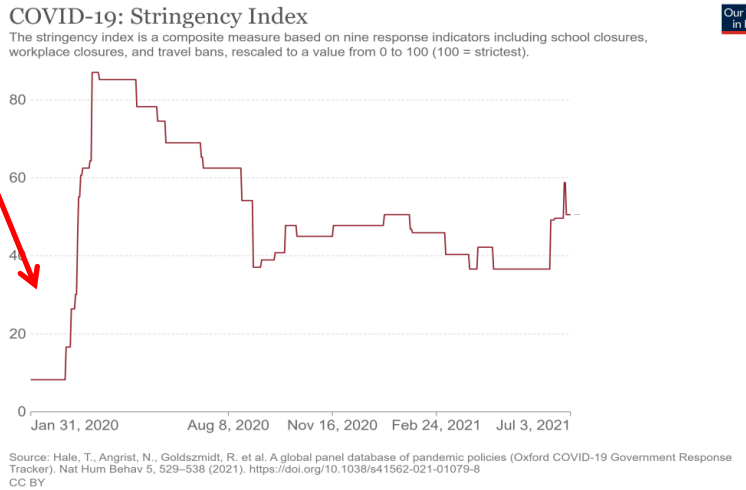
Длина данных: N=131 неделя



Начало пандемии в РФ

[2]**Ограничительные меры:** обобщенный индекс ограничений из базы данных база данных Oxford COVID-19 Government Response Tracker

ID	Name	Type	Targeted/ general?
Containment and closure			
C1	School closing	Ordinal	Geographic
C2	Workplace closing	Ordinal	Geographic
C3	Cancel public events	Ordinal	Geographic
C4	Restrictions on gathering size	Ordinal	Geographic
C5	Close public transport	Ordinal	Geographic
C6	Stay-at-home requirements	Ordinal	Geographic
C7	Restrictions on internal movement	Ordinal	Geographic
C8	Restrictions on international travel	Ordinal	No
Health systems			
H1	Public information campaign	Ordinal	Geographic



1. https://www.tinkoff.ru/
2. A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker) / Thomas Hale, Noam Angrist, Rafael Goldszmidt, Beatriz Kira, Anna Petherick, Toby Phillips, Samuel Webster, Emily Cameron-Blake et al. // Nature Human Behaviour. — 2021. — Vol. 5, no. 4. — Pp. 529–538.

Редукция данных на основе метода главных компонент (ЭОФ-разложение)

Редукция данных - выделение из данных относительно небольшого числа переменных, содержащих информацию о ключевых свойствах наблюдаемой динамики

Данные $\mathbf{x}_1, \dots, \mathbf{x}_N$ → *Главные компоненты* $\mathbf{y}_1, \dots, \mathbf{y}_N$

$$\mathbf{x}_n \in \mathbb{R}^D,$$

$$\mathbf{y}_n \in \mathbb{R}^d, d \ll D$$

Разложение по базису эмпирических ортоганальных функций (ЭОФ разложение) - проекция данных на линейные многообразия меньшей размерности

$$\mathbf{x}_n = \sum_{i=1}^{d} \mathbf{V}_i \cdot y_{i,n} + \boldsymbol{\eta}_n$$

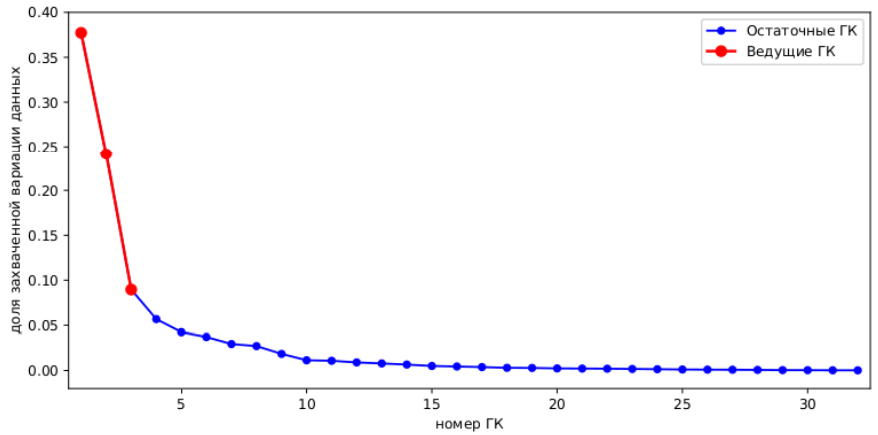
$$\mathbf{y}_n = \mathbf{V}^T \mathbf{x}_n, \mathbf{V}_i \in \mathbb{R}^D$$

Собственные векторы ковариационной матрицы данных

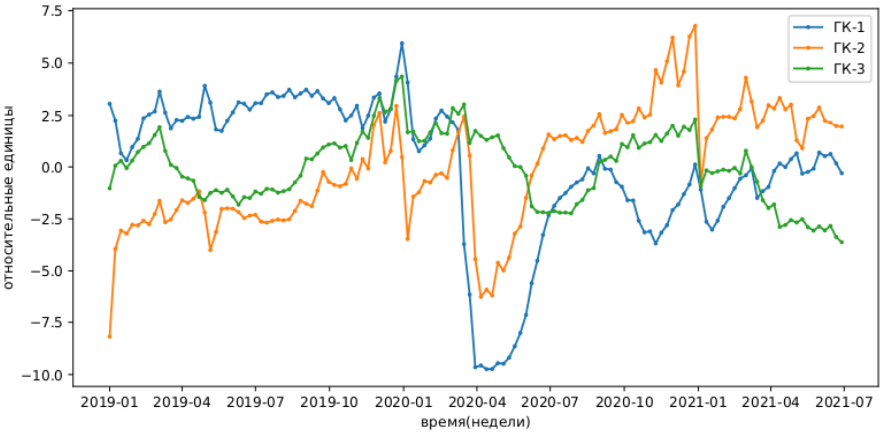
Редукция данных - необходимый шаг при построении модели по высокоразмерным данным наблюдений

ЭОФ - анализ данных покупательской активности (Москва)

Спектр захваченных вариаций

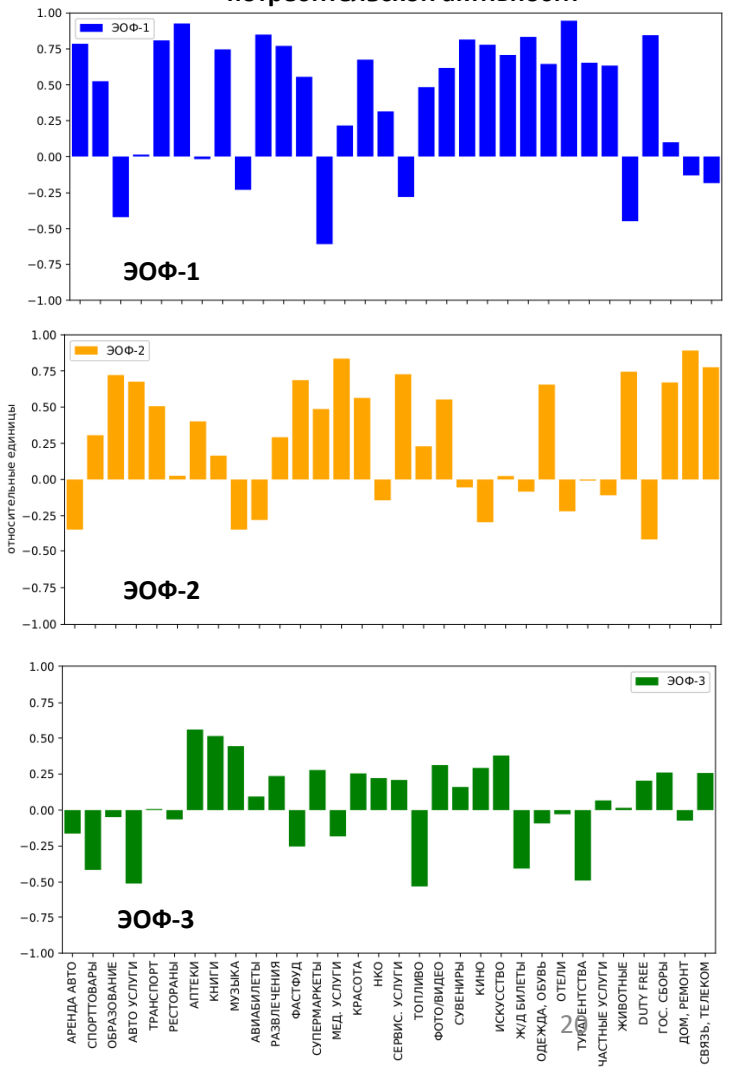


Временные ряды главных компонент

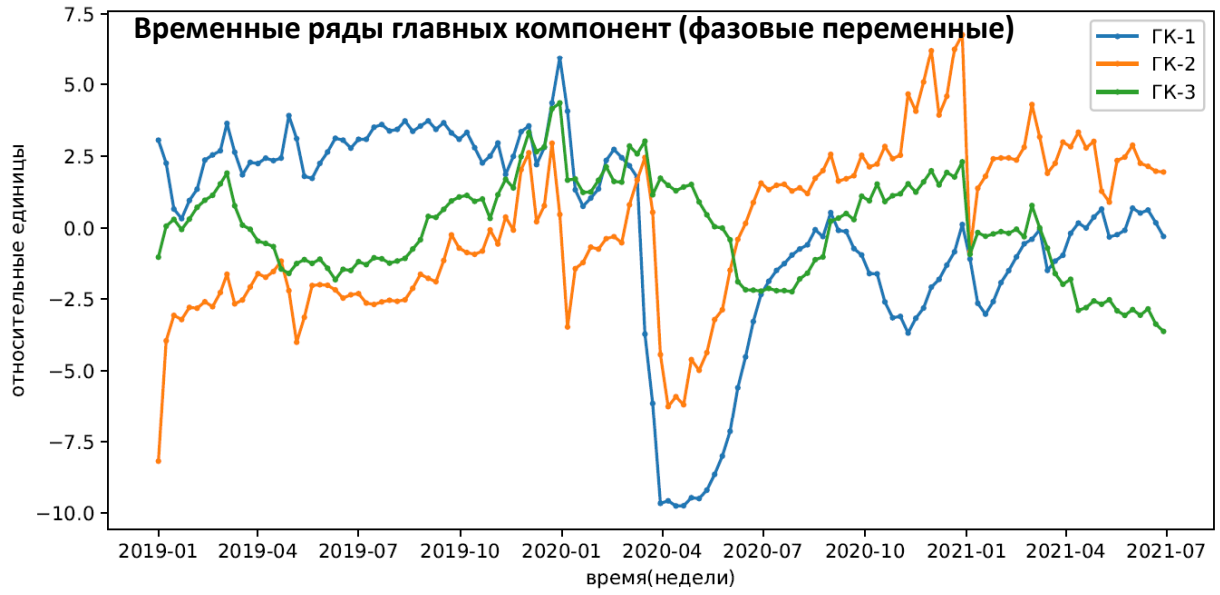


- Данные потребительской активности эффективно описываются тремя ведущими ЭОФ
- Каждая их этих ЭОФ вносит вклад в различные группы категорий потребительской активности

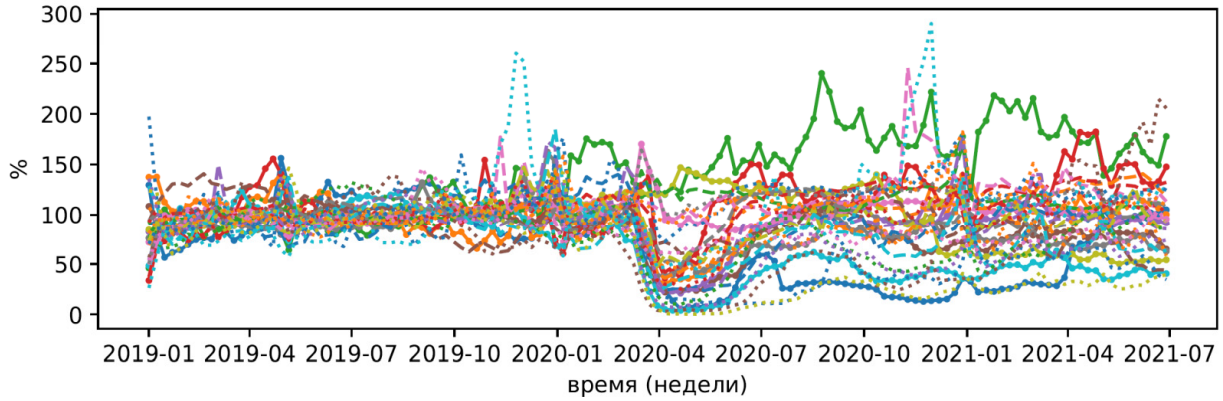
Вклад ЭОФ по категориям потребительской активности



Данные для построения модели



обратное преобразование
в пространство данных



Обучение и оптимизация модели в рамках Байесова подхода

- **Обучение** - подгонка модели к данным на основе выбранной ценовой функции

- **Оптимальная модель** - баланс между слишком простой и слишком сложной моделями

- **Байесов подход** - задача поиска оптимальной эмпирической модели сводится к одной из известных задач математической статистики – определению неизвестных параметров распределения по имеющейся выборке данных

- **В рамках байесова подхода удается определить обоснованность учета в эмпирической модели внешних воздействий**

$$\mathbf{y}_n = \mathbf{f}(\mathbf{y}_{n-L}, \dots, \mathbf{y}_{n-1}) + \hat{\mathbf{g}}\,\xi_n$$

μ_f – параметры детерминированной части

$\hat{\mathbf{g}}$ – параметры стохастической части

$\mathbf{H}_l = \{L, m\}$ – структурные параметры

Ценовая функция на параметры модели:

$$P(\mathbf{y}|\mu_f, \hat{\mathbf{g}}, \mathbf{H}_l) \cdot P_{\mathbf{H}_l}(\mu_f, \hat{\mathbf{g}})$$

Ценовая функция

на структурные параметры модели (обоснованность):

$$P(\mathbf{y}|\mathbf{H}_l) = \int P(\mathbf{y}|\mu_f, \hat{\mathbf{g}}, \mathbf{H}_l) \cdot P_{\mathbf{H}_l}(\mu_f, \hat{\mathbf{g}}) d\,\mu_f d\hat{\mathbf{g}}$$

Байесов критерий оптимальности модели:

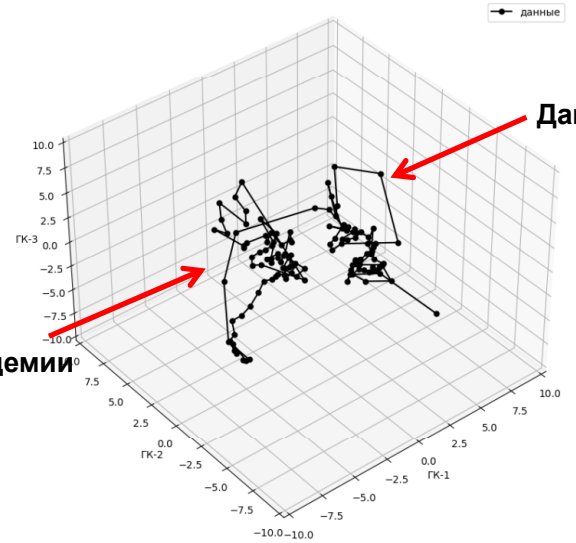
$$\Psi_{\mathbf{H}_l}(\mu_f, \hat{\mathbf{g}}) = -\ln\left[P(\mathbf{y}|\mu_f, \hat{\mathbf{g}}, \mathbf{H}_l) \cdot P_{\mathbf{H}_l}(\mu_f, \hat{\mathbf{g}})\right],$$

$$L = -\ln P(\mathbf{y}|\mathbf{H}_l) = \Psi_{\mathbf{H}_l}(\overline{\mu_f}, \bar{\hat{\mathbf{g}}}) + \frac{1}{2}\ln\left[\left|\frac{1}{2\pi}\nabla\nabla^T\Psi_{\mathbf{H}_l}(\overline{\mu_f}, \bar{\hat{\mathbf{g}}})\right|\right]$$

Штраф за сложность модели

Фазовый портрет эмпирической модели

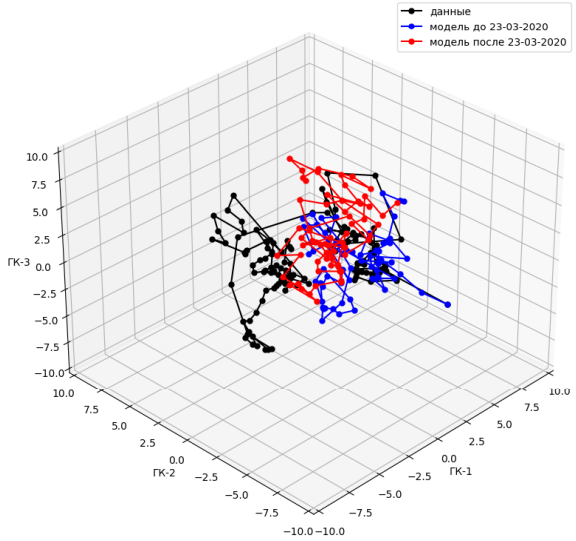
- В фазовом пространстве различимы две разнесенные области
- Момент перехода фазовой траектории между областями совпадает с началом пандемии в РФ
- Модель с учетом индекса ограничений является оптимальной с точки зрения байесова критерия и отражает данный переход



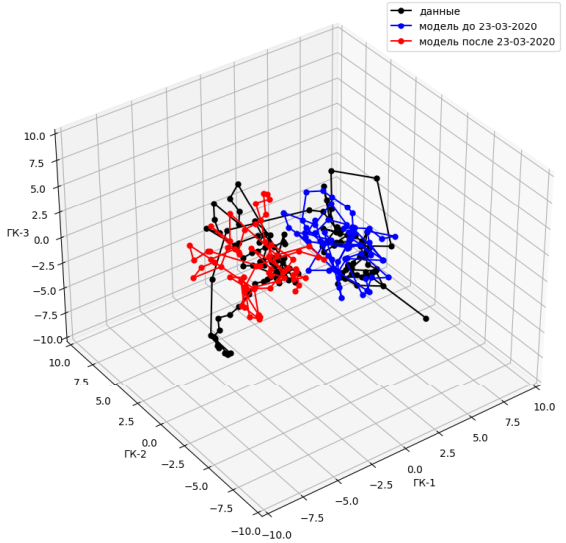
Данные до начала пандемии

Данные после начала пандемии

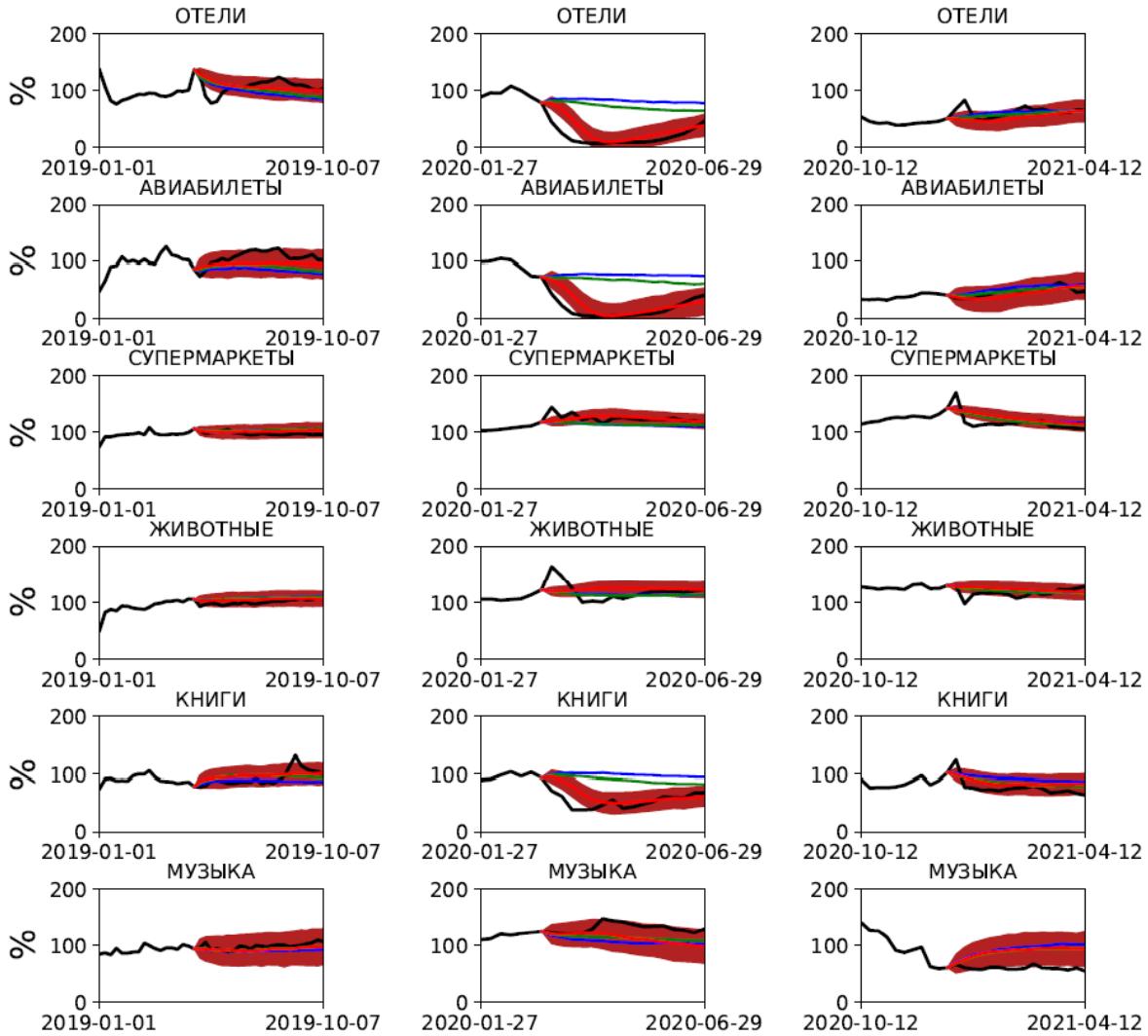
Модель без учета ограничений



Модель с учетом ограничений



Примеры ретроспективного прогноза конкретных категорий потребительской активности



Меры предсказательной способности эмпирической модели

Среднеквадратичная ошибка прогноза:

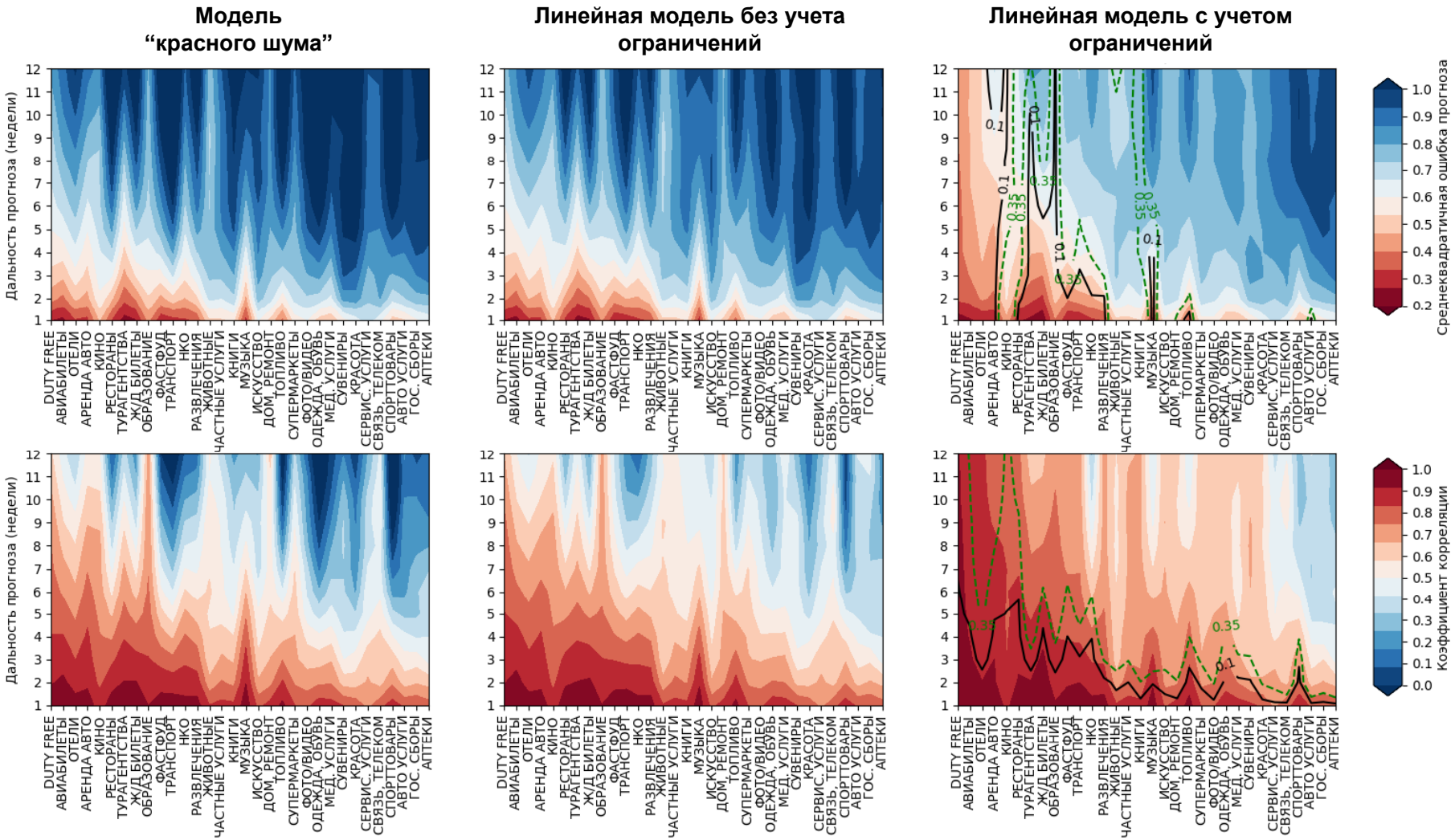
$$e_j = \sqrt{\frac{\sum_{n=l}^{N-j} (x_{n+j} - \bar{x}_{n,j})^2}{N - j - l + 1}}$$

Коэффициент корреляции:

$$r_j = \frac{\sum_{n=l}^{N-j} \Delta x_{n+j} \bullet \Delta \bar{x}_{n,j}}{\sqrt{\sum_{n=l}^{N-j} (\Delta x_{n+j})^2 \bullet \sum_{n=l}^{N-j} (\Delta \bar{x}_{n,j})^2}}$$

- x_{n+j} – истинное значение предсказываемой величины
- $\bar{x}_{n,j}$ – значение, спрогнозированное эмпирической моделью
- n-номер стартовой точки
- j-дальность прогноза в неделях
- N-длина наблюдаемой выборки
- l-длина памяти модели в неделях

Предсказательная способность эмпирической модели для различных категорий потребительской активности



Оценка статистической значимости на основе суррогатных данных

Байесов критерий позволяет формально определить оптимальную модель по имеющейся выборке данных

Имеет ли оптимальная модель практическую ценность?

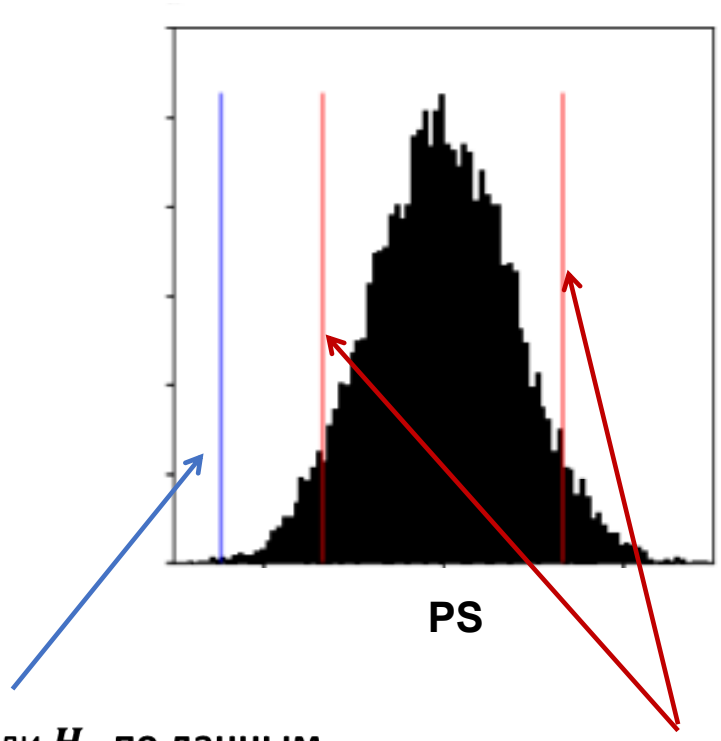
Пример:

H_1 – модель без учета внешнего воздействия

H_2 –модель с учетом внешнего воздействия

Требуется сравнить прогностическую способность моделей: **PS**

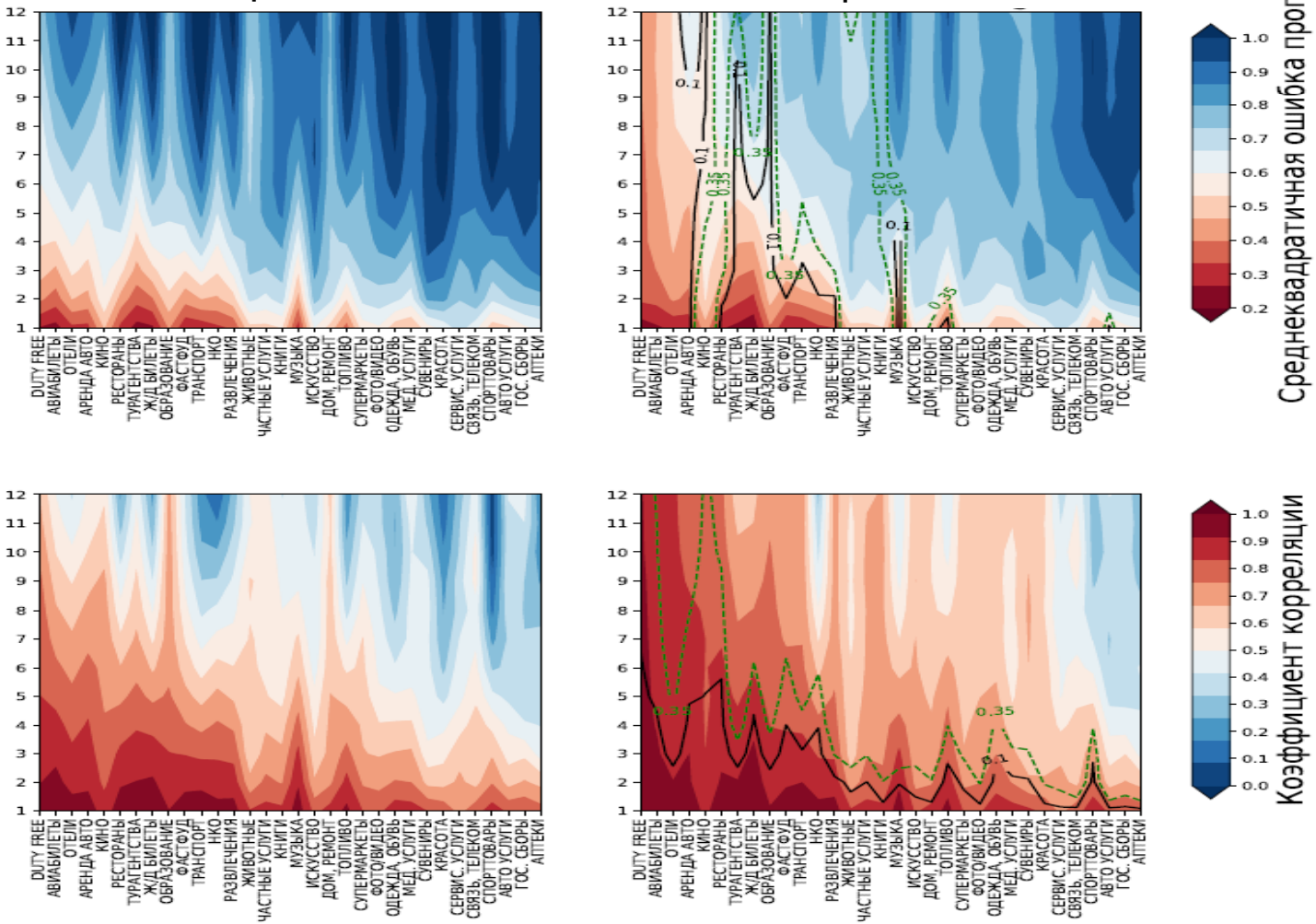
Распределение величины **PS модели H_2 , рассчитанное по суррогатам модели H_1 (специально обеспечено отсутствие информации о внешнем воздействии)**



Предсказательная способность эмпирической модели для различных категорий потребительской активности

Линейная модель без учета ограничений

Линейная модель с учетом ограничений

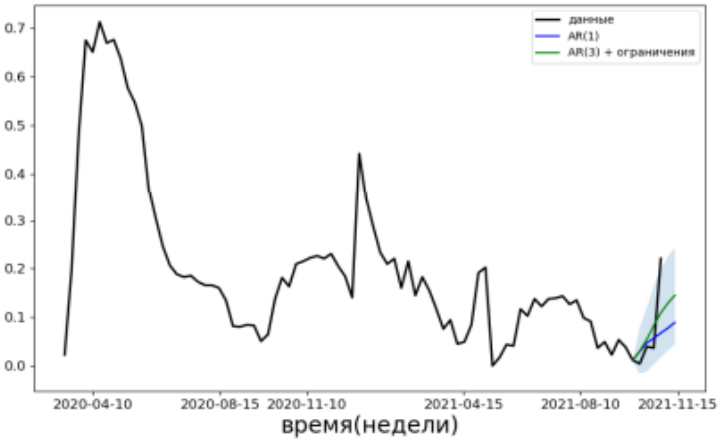


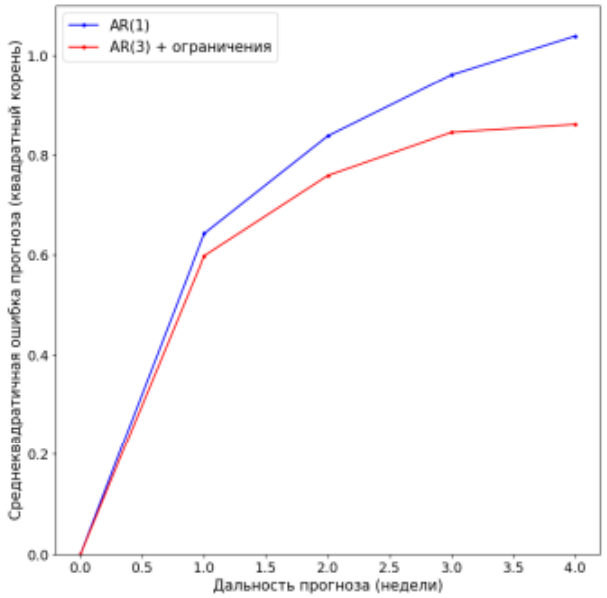
Прогноз активности населения

комбинированная эмпирическая модель:
прогноз параметра агентной модели “*активность населения*” с учетом форсингов в виде главных компонент данных *потребительской активности*

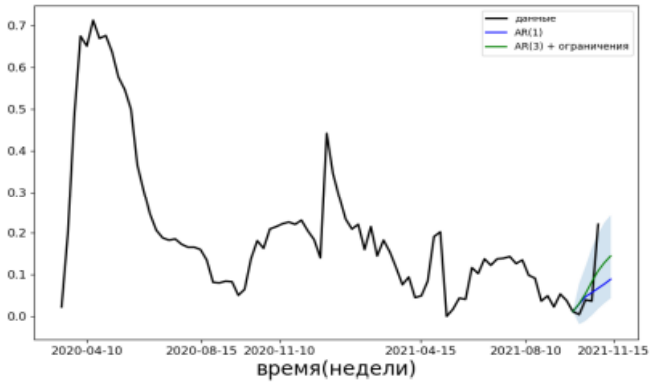
$$\begin{cases} q_{n+1} = f(q_n, \dots, q_{n-l+1}, \mathbf{y}_n) + \sigma\xi_n. \\ \mathbf{x}_{n+1} = F(\mathbf{x}_n, \dots, \mathbf{x}_{n-L+1}, h_n) + \hat{\mathbf{g}}\boldsymbol{\xi}_n. \end{cases}$$

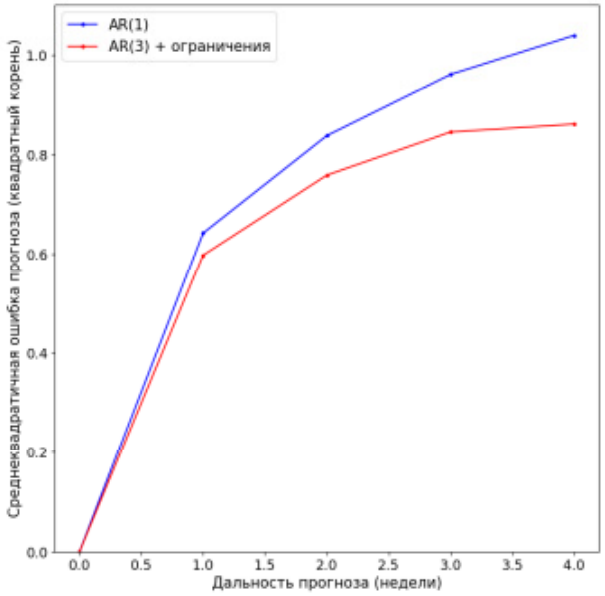
q_n — значения индекса “активность населения” в момент времени n,
$\mathbf{x}_n$ — значения ГК потребительской активности,
h_n — обобщенный индекс ограничений,
l, L — порядки моделей (глубина памяти)





Результат прогноза активности населения

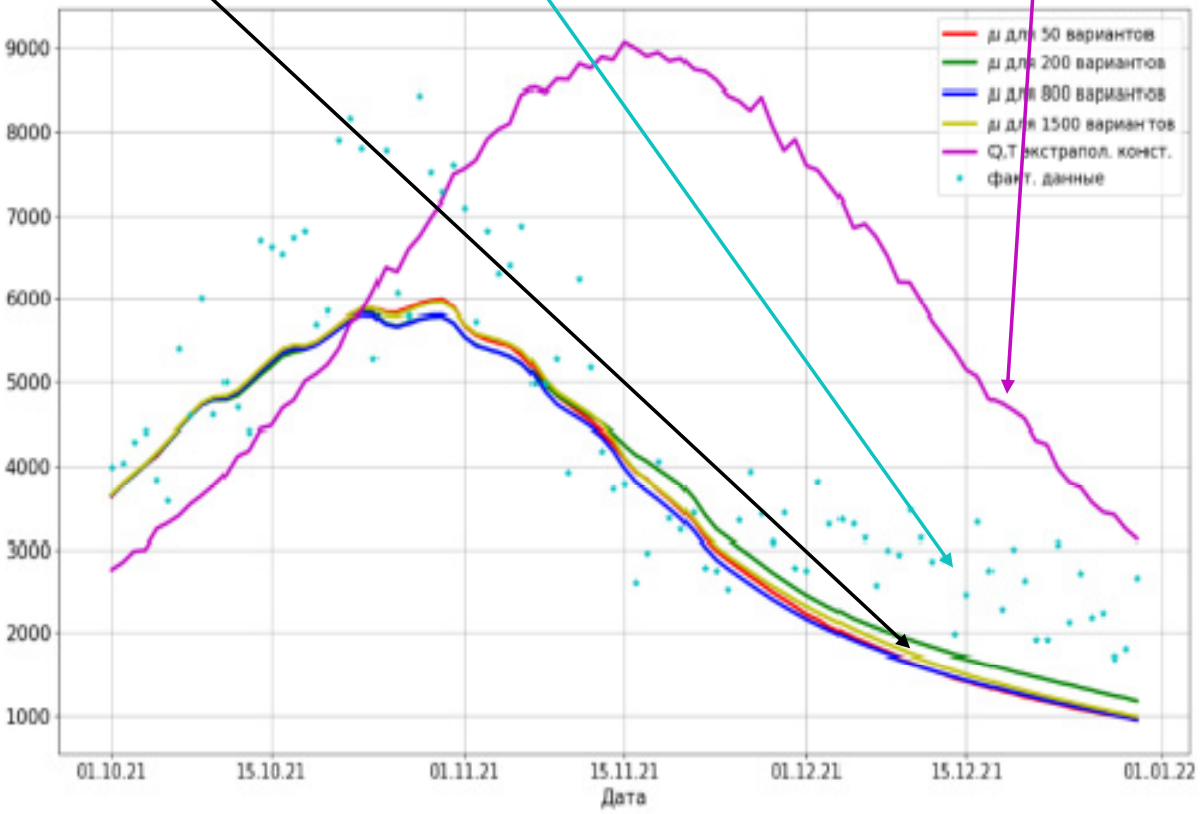




с использованием комбинированной эмпирической модели

экстраполяция параметра активности населения константой

фактические данные



Спасибо за внимание!