Bank of Russia

ИЗМЕРЕНИЕ ИНФЛЯЦИИ В РОССИИ НА ОСНОВЕ ГРАНУЛЯРНЫХ ДАННЫХ ККТ

Алдохин Д.В., Долгов Д.А., Марченко П.В.,
Милютин П.П., Селезнев С.М., Шибитов Д.С.

Департамент исследований и прогнозирования
Банк России

26 марта 2024

Disclaimer

Настоящая презентация отражает личную позицию авторов. Содержание и результаты данного доклада не следует рассматривать, в том числе цитировать в каких-либо изданиях, как официальную позицию Банка России или указание на официальную политику или решения регулятора. Любые ошибки в данном материале являются исключительно авторскими.

Используемые в расчетах данные контрольно-кассовой техники предоставлены ФНС России в обезличенном виде и без обработки (могут содержать аномалии, например, ошибки пользователей ККТ при вводе данных).

Мотивация

Построение ценовых индексов на данных ККТ в перспективе позволит:

- Проводить анализ ценовых процессов до выхода официальной статистики
- Исследовать ценовые процессы в разрезах, недоступных в официальной статистике (например, по другим товарным категориям)

Индексы	Лаг относительно конца отчетного периода и разрезы
Месячные индексы Росстата	Публикуются с задержкой 10 – 15 дней По категориям корзины Росстата
Недельные индексы Росстата	Публикуются с задержкой 2 дня По части категорий корзины Росстата
Дневные индексы на основе данных ККТ	Могут рассчитываться с задержкой 1-2 дня По практически любым категориям, в том числе и по большинству категорий Росстата

Что мы хотим?

Наша цель – разработка методологии построения ценовых индексов на данных ККТ, которая должна покрывать все этапы построения индексов:

- Подготовка данных
- Расчёт (агрегация) дневных цен и объемов для каждого товара
- Классификация товаров по товарным категориям
- Расчёт индексов

Также методология должна:

- Использовать данные как можно более полно
- Быть устойчивой к особенностям данных и их недостаткам

Российские данные контрольно-кассовой техники

Преимущества:

- Практически полное покрытие рынка потребительских товаров и услуг
- Наличие фактических объемов продаж по каждому товару

Недостатки:

- Сложность идентификации товаров и их классификации
- Постоянные выпадения товаров, осложняющие расчет индексов
- Могут содержать операции не только с потребительскими товарами и услугами

Доступные нам атрибуты российских данных ККТ

Основные доступные поля

- Зашифрованный идентификатор кассы
- Зашифрованный идентификатор чека
- Дата
- Регион
- Город/округ
- Тип чека
- Тип операции
- Способ оплаты
- Наименование предмета расчёта
- Количество предмета расчёта
- Стоимость единицы предмета расчёта

Недоступные поля

- Идентификатор продавца
- Точный адрес
- Точное время
- Заводской идентификатор кассы
- Заводской идентификатор чека

Характеристики датасета

- Период: 01.01.2022 – 30.09.2022
- Количество касс: 3,6 млн
- Количество чеков: 53 млрд
- Количество записей: 150 млрд
- Количество уникальных наименований: 3 млрд (из них 1,8 млрд встречались только один раз)

Зачем для данных ККТ нужна другая методология?

Особенности методологии ИПЦ, которые сложно перенести на данные ККТ

- Требуется четкая идентификация каждого товара и его потребительских свойств
- Подразумевается малое количество отсутствующих наблюдений
- Используются цены «на полке» (без учета скидок, списывания бонусов и т.д.)

Преимущества данных ККТ, которые не учитываются в методологии ИПЦ

- Наличие данных о текущей структуре расходов населения

Определение формулы индекса

p_i^t – цена товара i в момент времени t

Модель:

$$p_i^t = PriceTrend_t * ItemQuality_i * IdiosyncraticComponent_{i,t}$$

$$\ln p_i^t = \ln(PriceTrend_t) + \ln(ItemQuality_i) + \ln\left(IdiosyncraticComponent_{i,t}\right)$$

$$\ln p_i^t = \sum_{t=1}^{T} \delta^t D^t + \sum_{i=1}^{N} \gamma_i D_i + \varepsilon_i^t$$

Ln(index)

Определение формулы индекса

$P_{TPD}^{0,t}$ – Time Product Dummy индекс

0 – базовый период, t – отчетный период, T – последний период

$$\ln p_i^t = \sum_{t=1}^{T} \delta^t D^t + \sum_{i=1}^{N} \gamma_i D_i + \varepsilon_i^t$$

$P_{TPD}^{0,t} = \exp(\hat{\delta}^t)$, где $\hat{\delta}^t$ находится с помощью взвешенного МНК, где веса $w_i^t = \frac{p_i^t q_i^t}{\sum_{i=1}^{N} p_i^t q_i^t}$

Преимущества

- Не зависит от базового периода
- Не требует работы с выпадениями товаров

Недостатки

- Пересматривается
- Вычислительно сложнее классических индексов

Определение формулы индекса

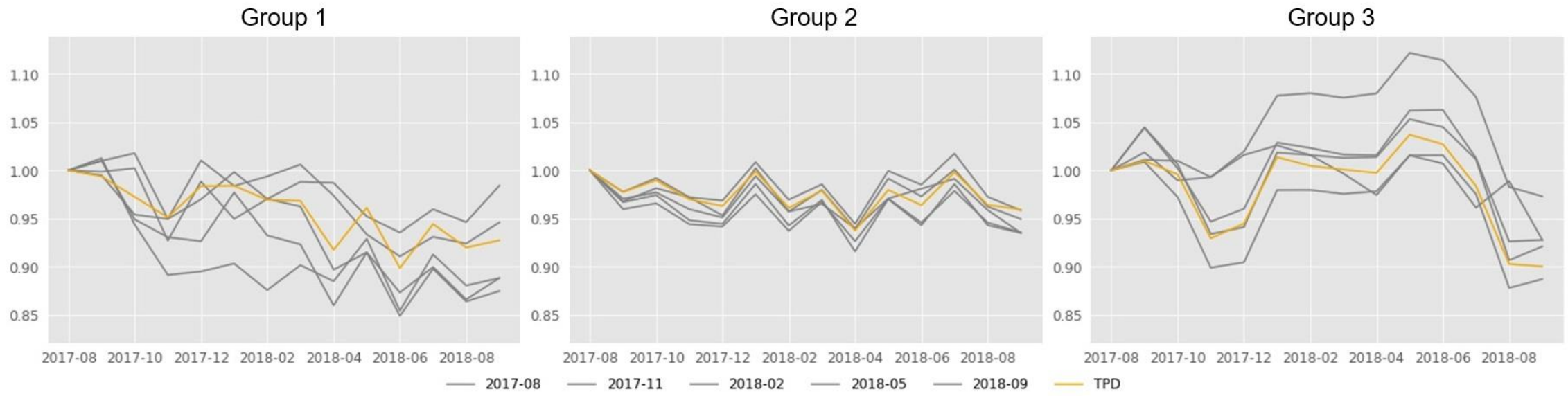
На практике индекс рассчитывается итеративно:

Равносильно расчёту через МНК:

$$\begin{cases} P_{TPD}^{0,t} = \dfrac{\prod_{i \in G_t} \left(\dfrac{p_i^t}{\hat{p}_i} \right)^{w_i^t}}{\prod_{i \in G_0} \left(\dfrac{p_i^0}{\hat{p}_i} \right)^{w_i^0}} \\ \hat{p}_i = \prod\limits_{\tau \in G^i} \left(\dfrac{p_i^\tau}{P_{TPD}^{0,\tau}} \right)^{\frac{w_i^\tau}{\sum_{\tau \in G^i} w_i^\tau}} \end{cases}$$

Свойства формулы индекса

Сравнение TPD индекса с «идеальным» индексом Фишера с разными базовыми периодами*



* Используется размеченная выборка из тестового среза данных (08.2017 – 09.2018)

Определение товара

Товар – всё, что имеет одинаковое наименование и продано в одной кассе.

- Позволяет разделять фактически разные товары, продающиеся под одинаковыми названиями в разных магазинах

Классификация товаров

Категории:

- Основаны на категориях Росстата с учетом особенностей номенклатуры ККТ

Обучающая выборка:

- Срез данных ККТ, размеченный сотрудниками Банка России
- Собранные методом веб-скрейпинга каталоги товаров торговых агрегаторов

Модель:

- Обученная на данных ККТ языковая модель BERT + NN classifier + OOD detector

Поскольку в данных содержится информация о цене и объеме продаж **каждого** отдельного товара, для построения **общего** индекса цен классификатор **не нужен.**

Фильтры

Чек не участвует в расчёте индекса, если:

- Сумма отдельных позиций превышает сумму чека более чем на 1 рубль
- Сумма чека меньше 5 рублей
- Сумма чека превышает 10*90-й процентиль максимальных дневных чеков в кассе за год

Товар не участвует в расчёте индекса, если:

- Он встречается менее в 5 днях
- Значение метрики вклада в волатильность индекса слишком высоко (см. далее)

При этом также **удаляются возвраты**

Проблема экстремальной волатильности индекса

Предположение: основной вклад в волатильность индекса вносят неинформативные наименования и наименования «заплатки».

Фильтр на товары с наибольшим вкладом в волатильность

Формула для расчёта меры вклада товара в суммарную волатильность индекса:

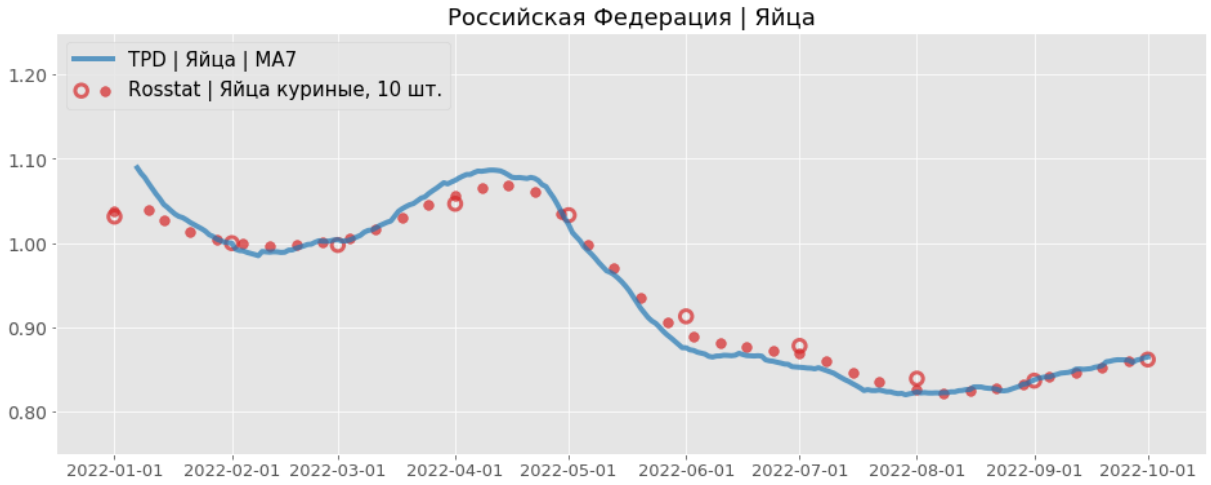
$$v_i = \sum_{t \in T} {w_i^t}^2 * \left(log\left(p_i^t\right) - \overline{log\left(p_i^t\right)} \right)^4$$

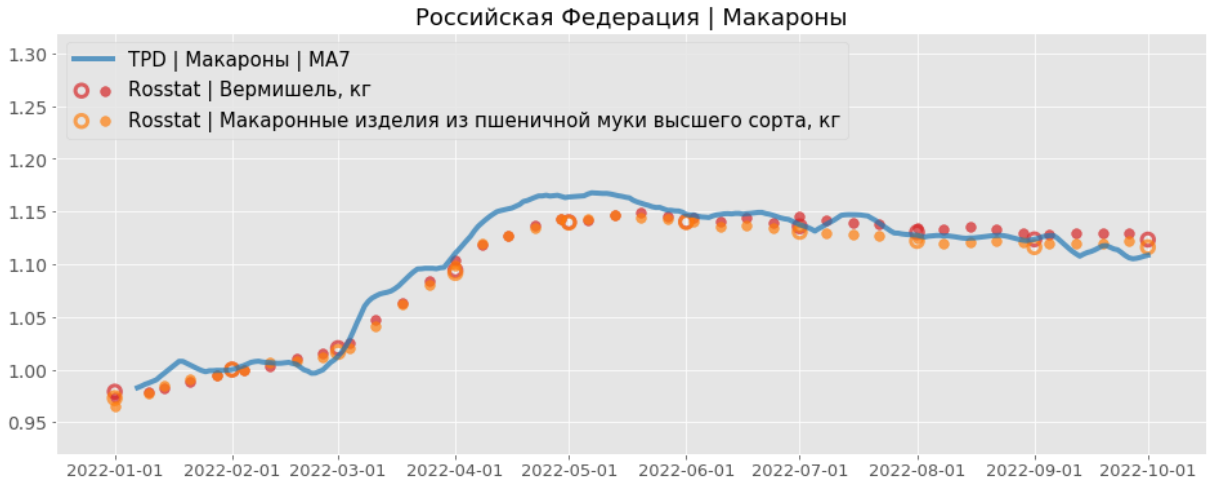
Перед расчётом индекса исключается часть товаров с максимальными значениями v_i.

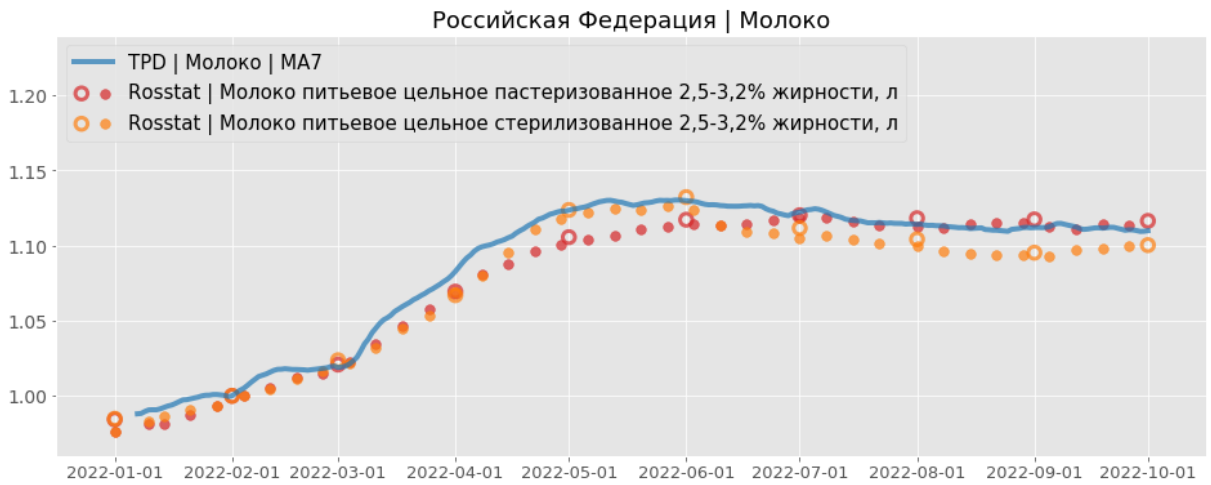
В настоящий момент доля исключаемых товаров определяется вручную для каждого индекса. Ведется работа по автоматизации этого процесса.

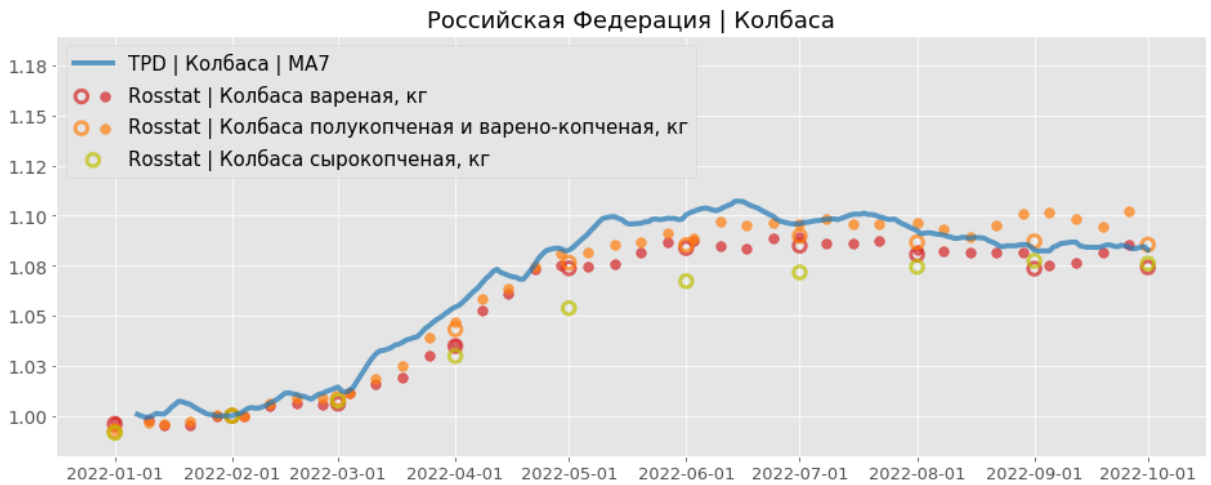
Индексы

Индексы для продовольственных товаров

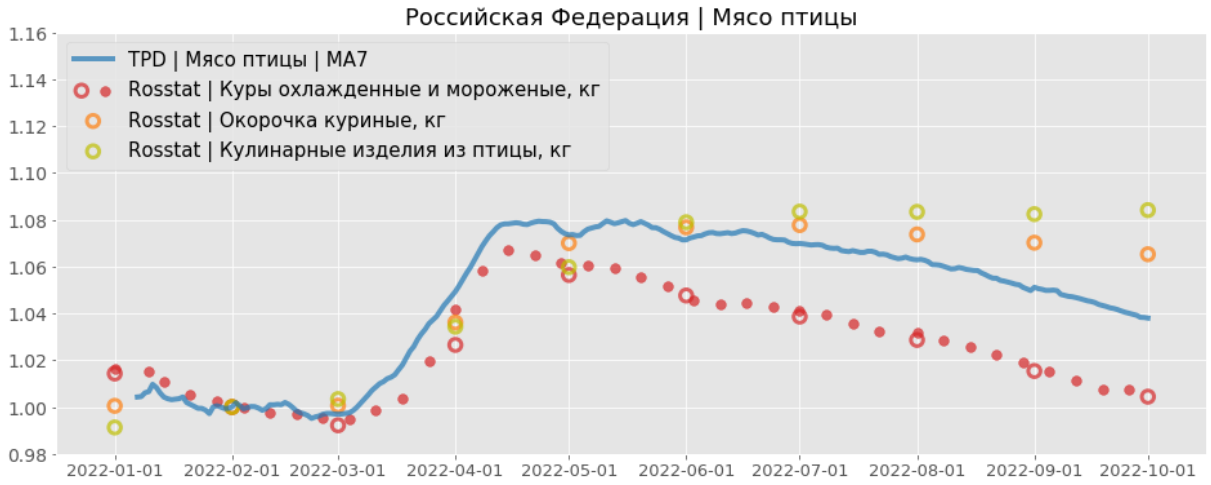


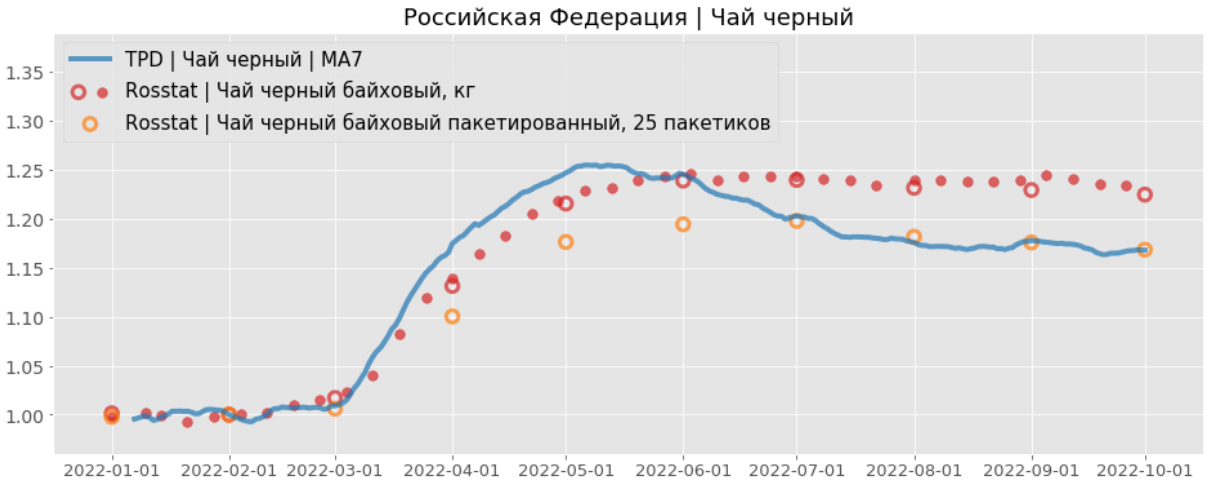


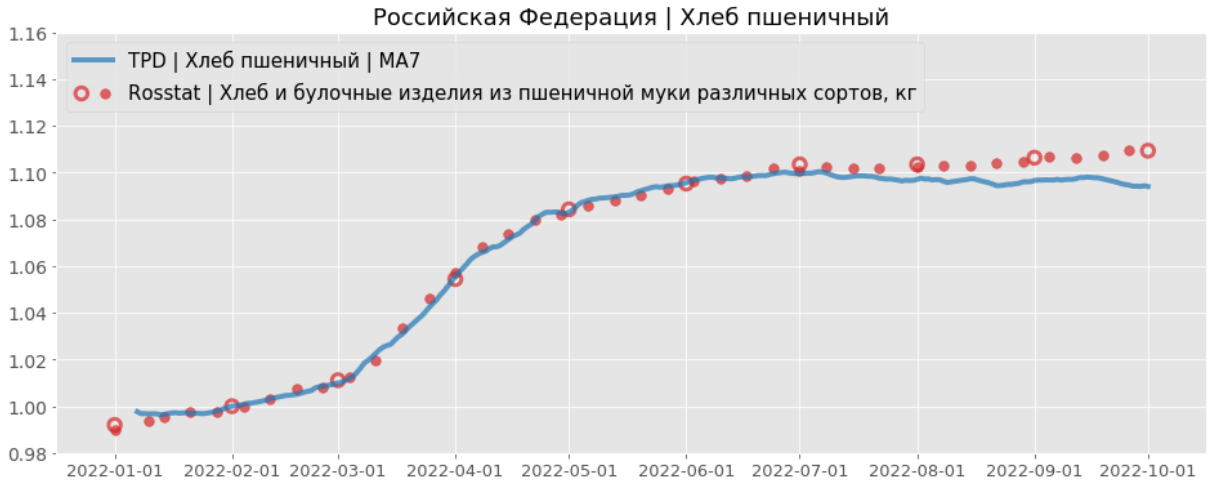


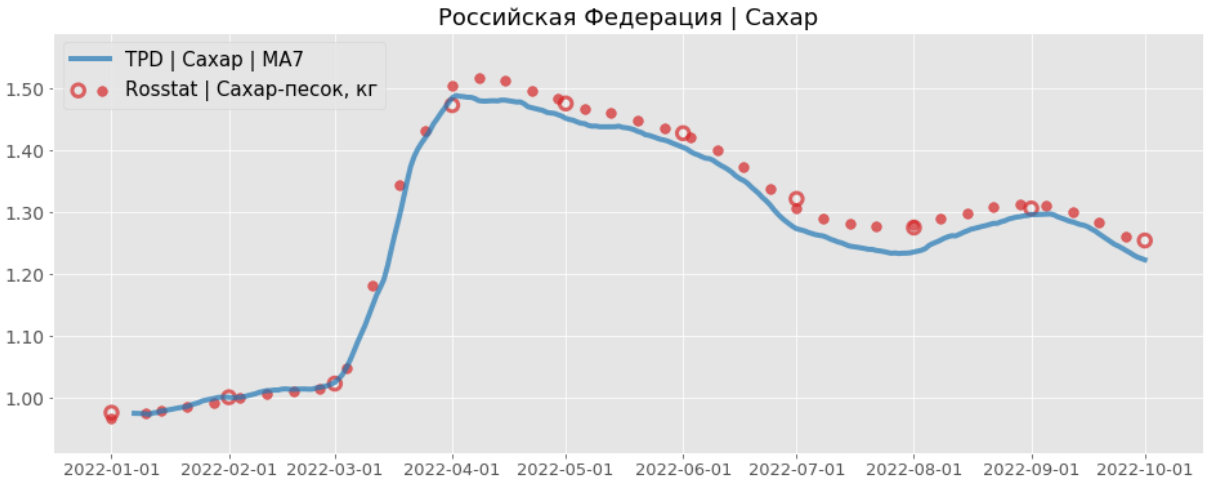


Индексы для продовольственных товаров

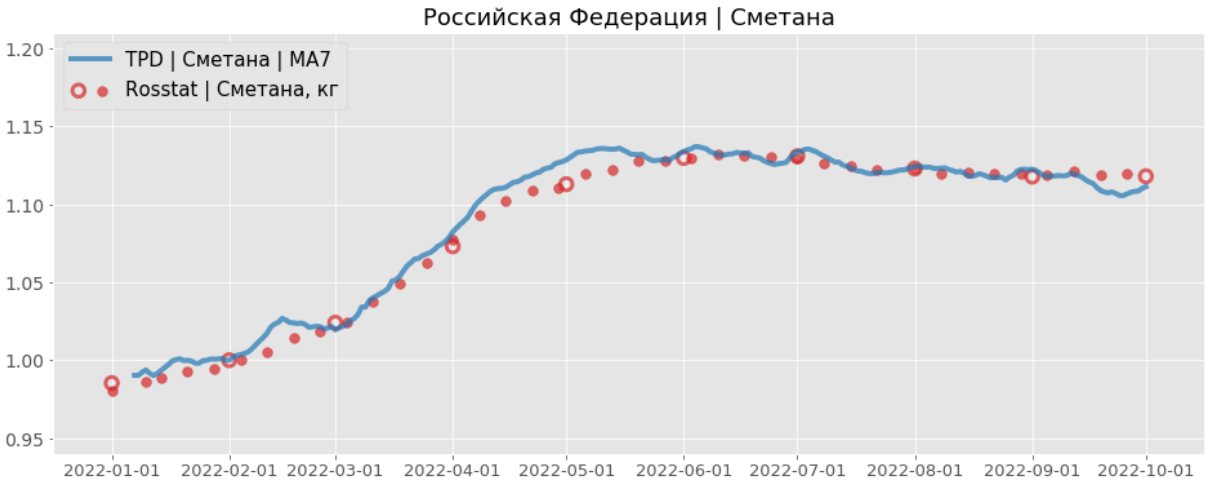


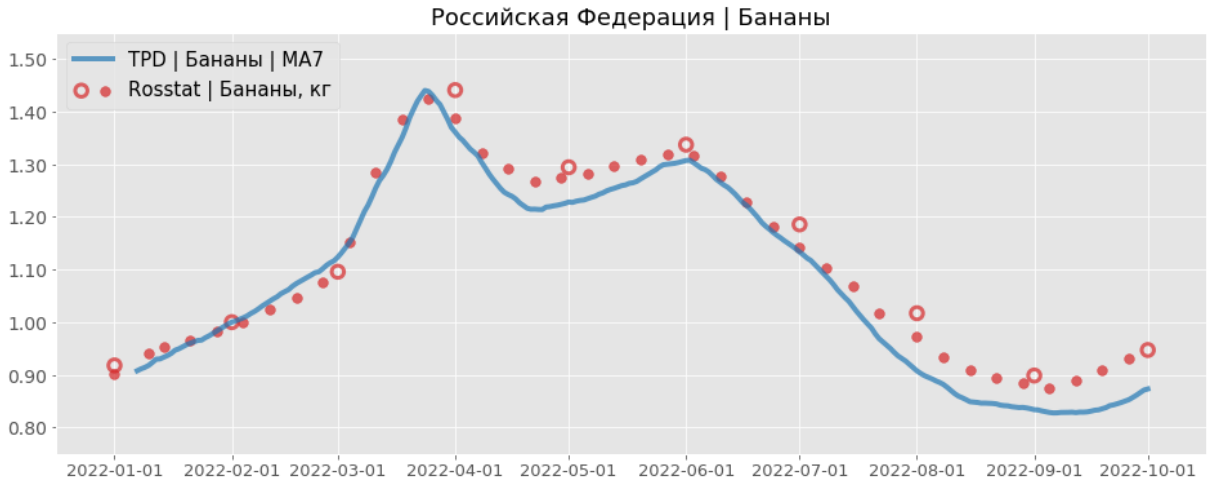


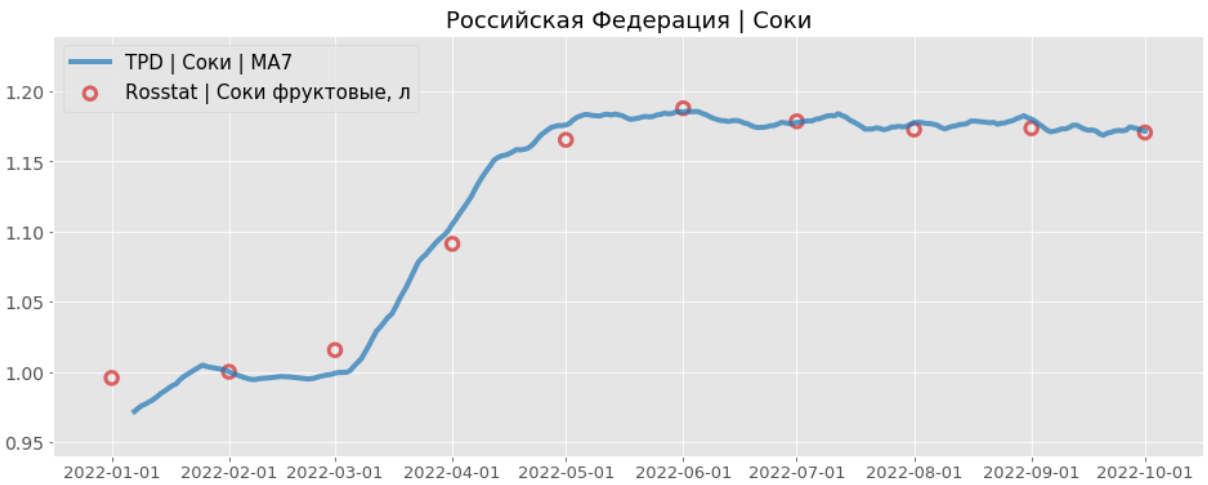


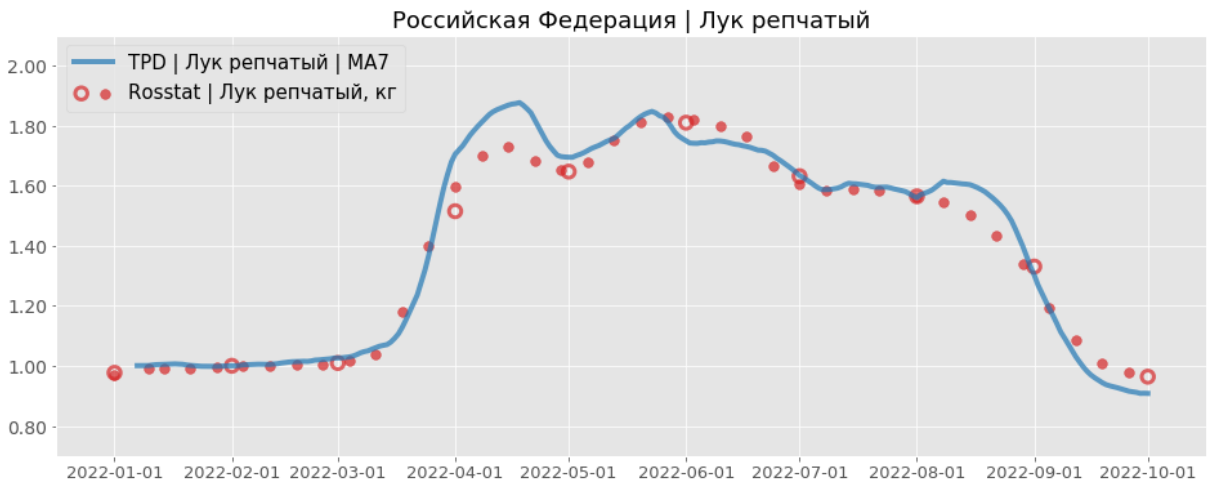


Индексы для продовольственных товаров

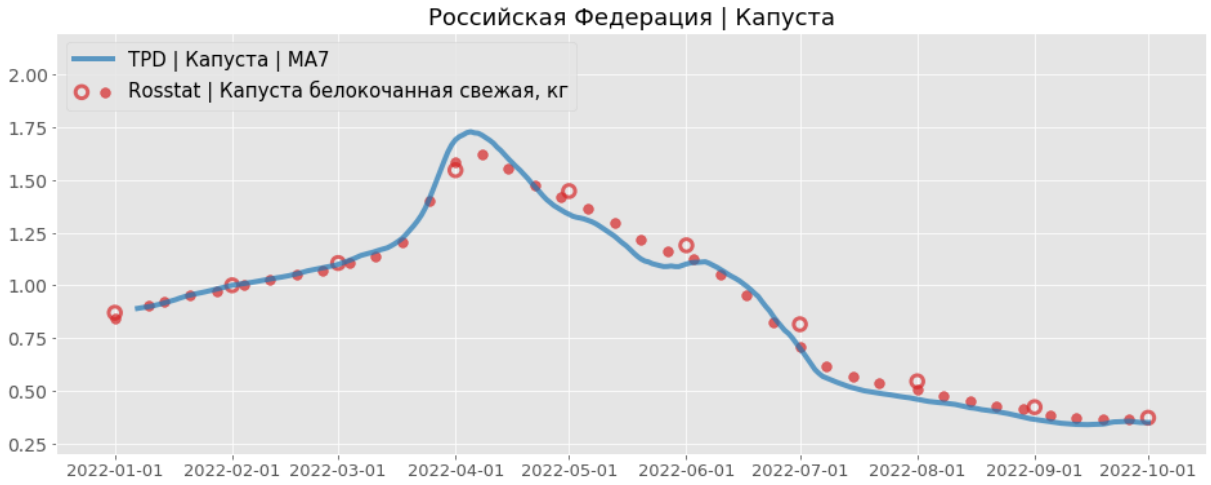


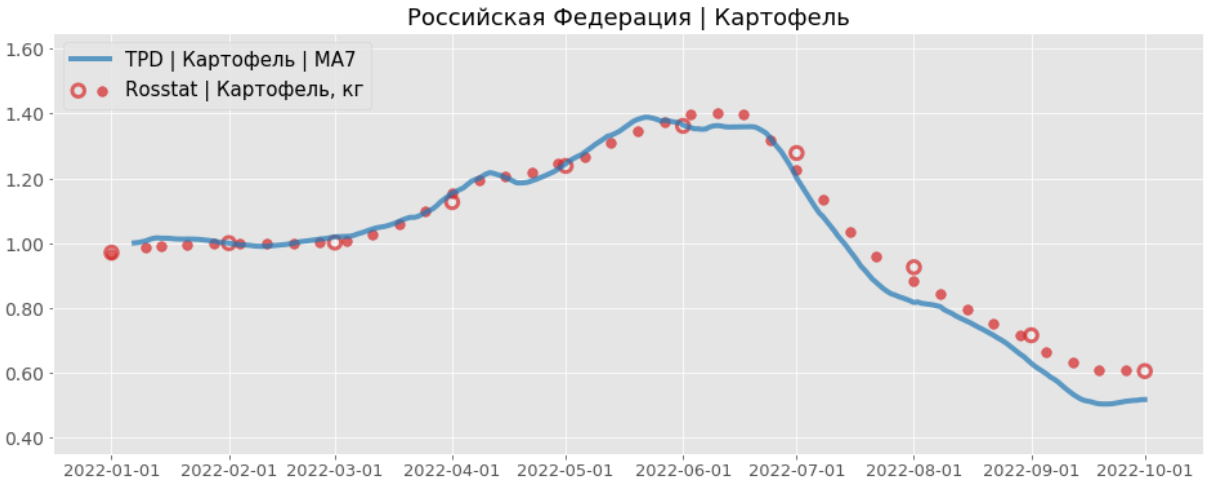


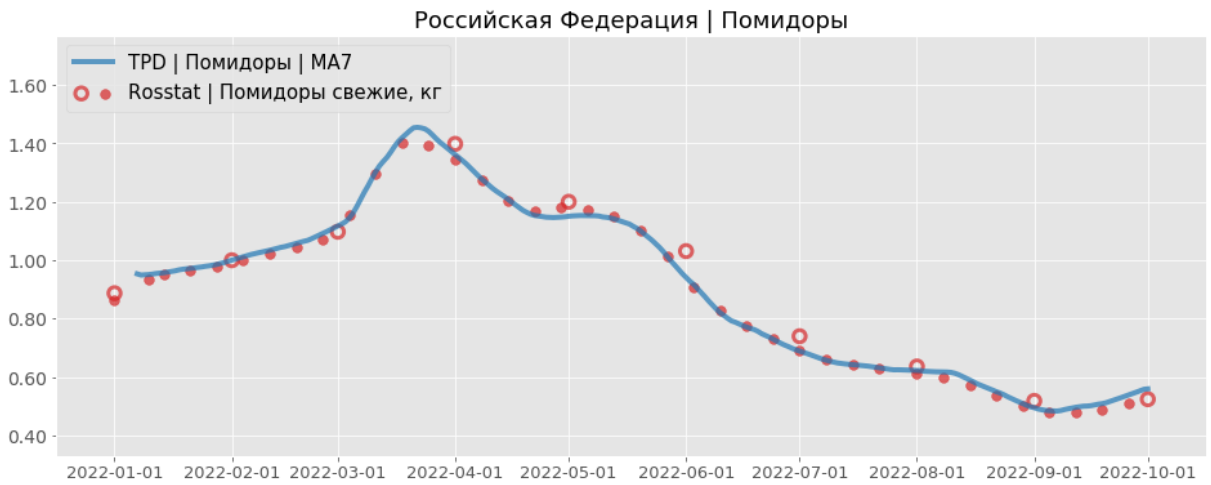


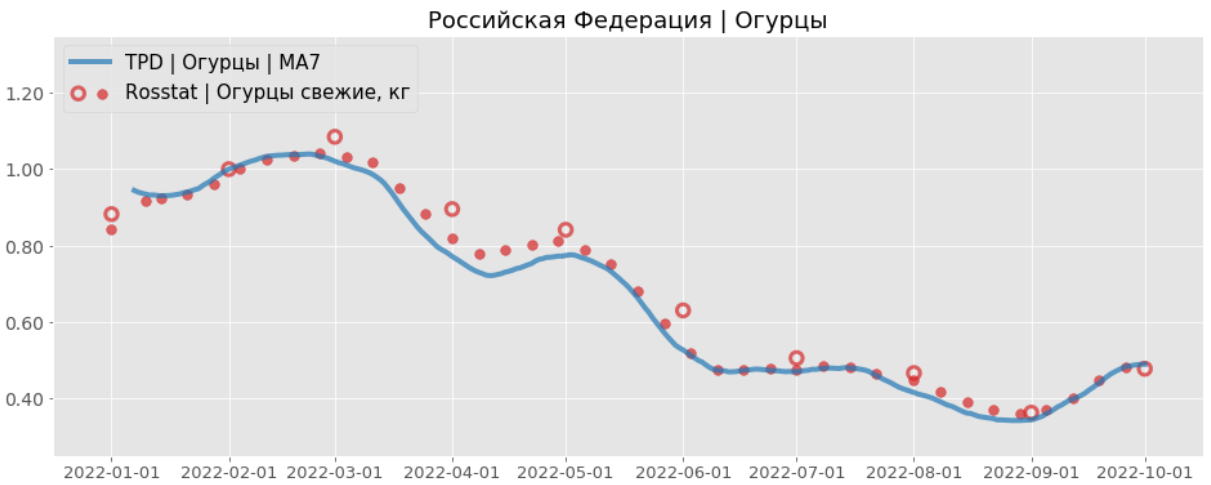


Индексы для продовольственных товаров

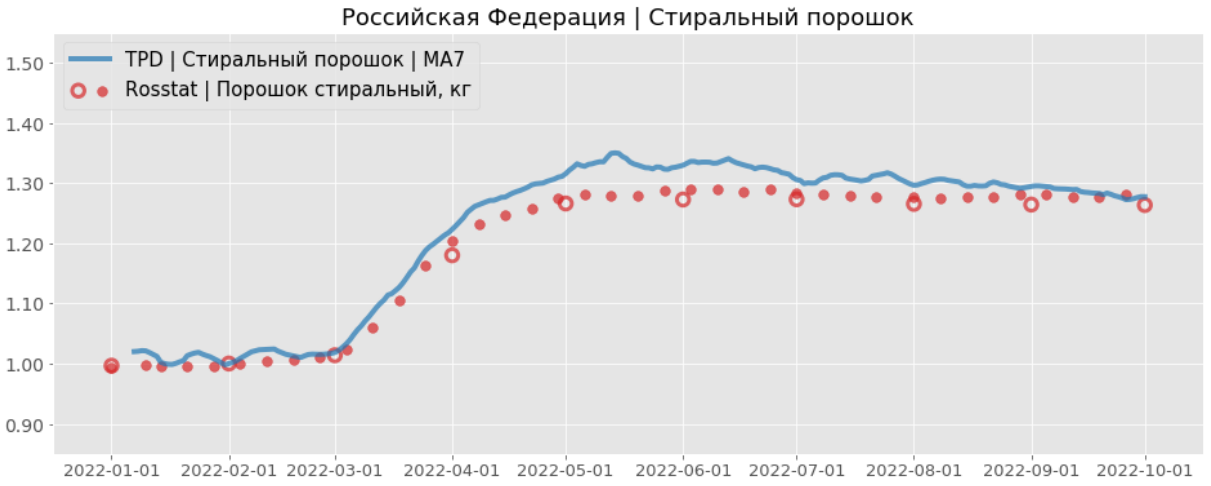


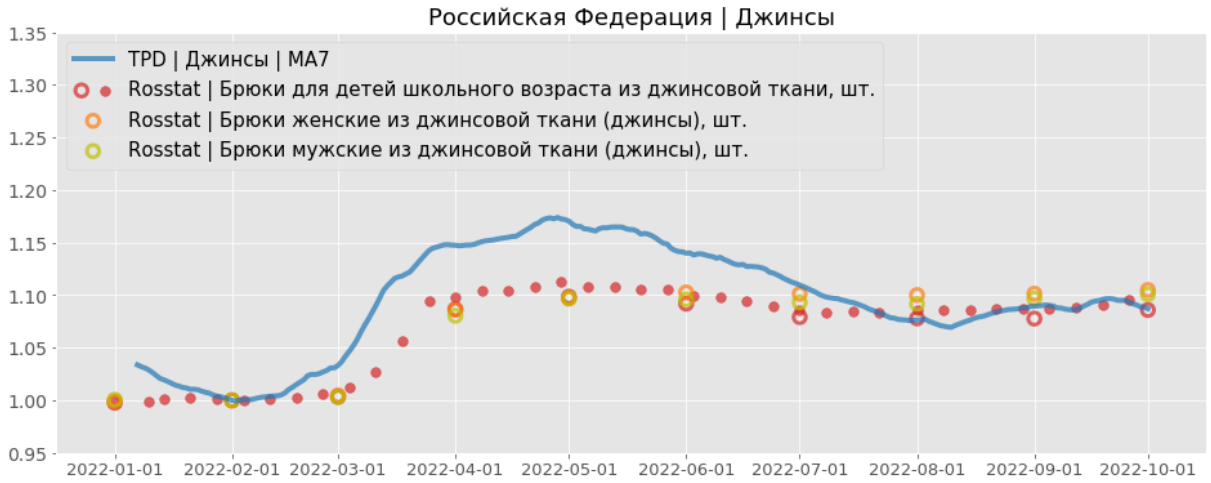


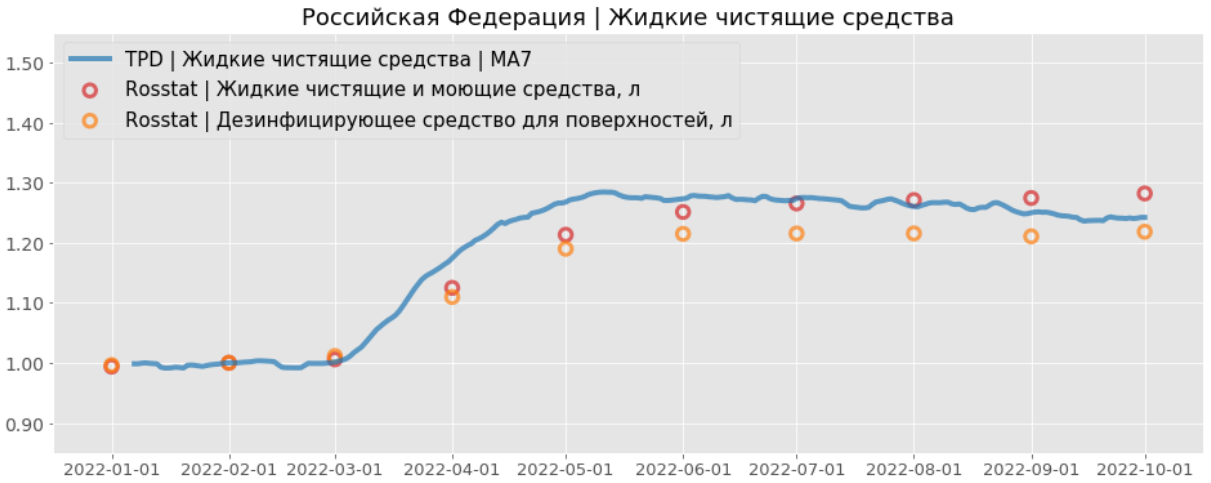




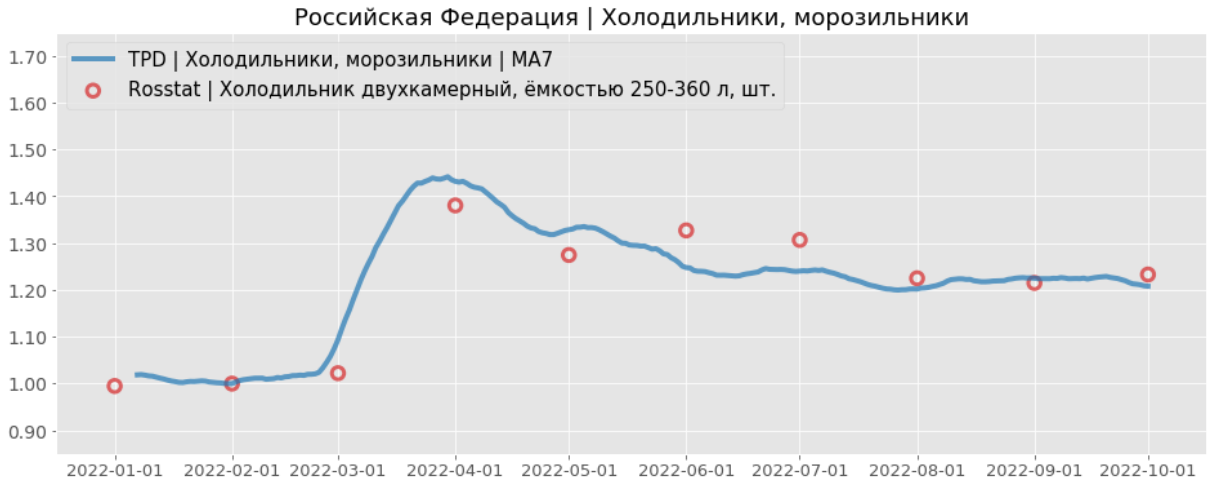
Индексы для непродовольственных товаров

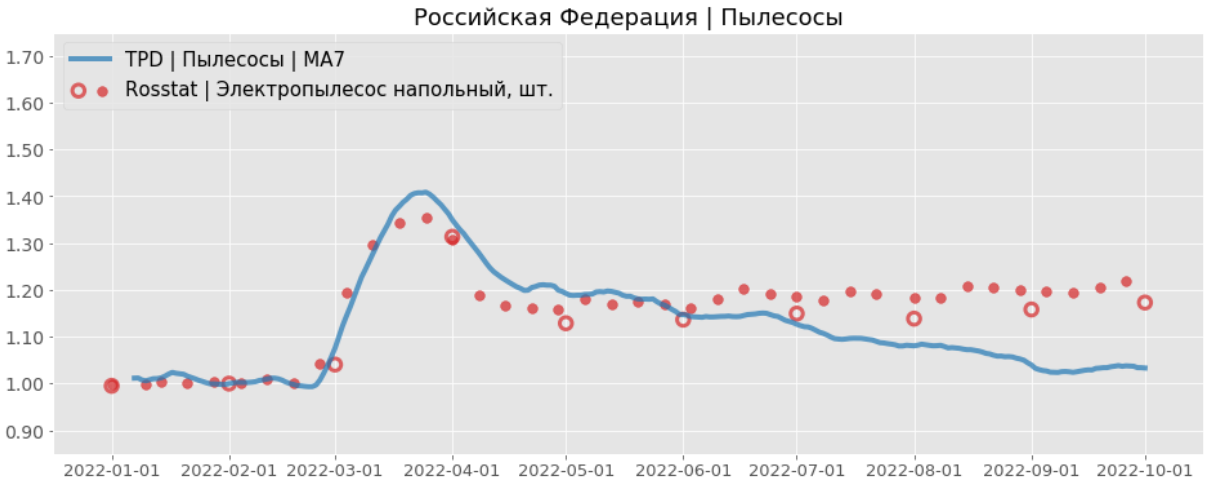




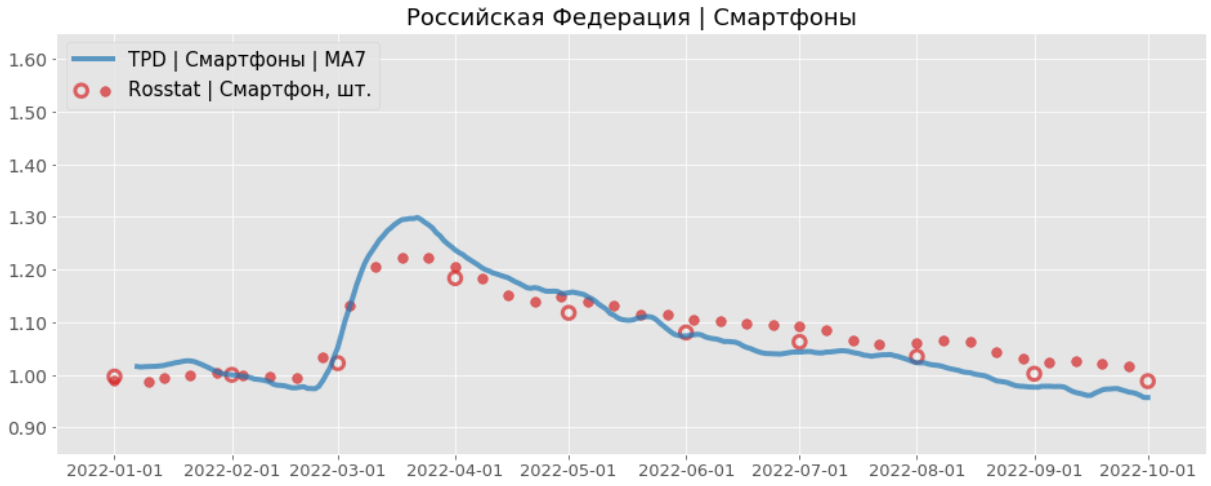


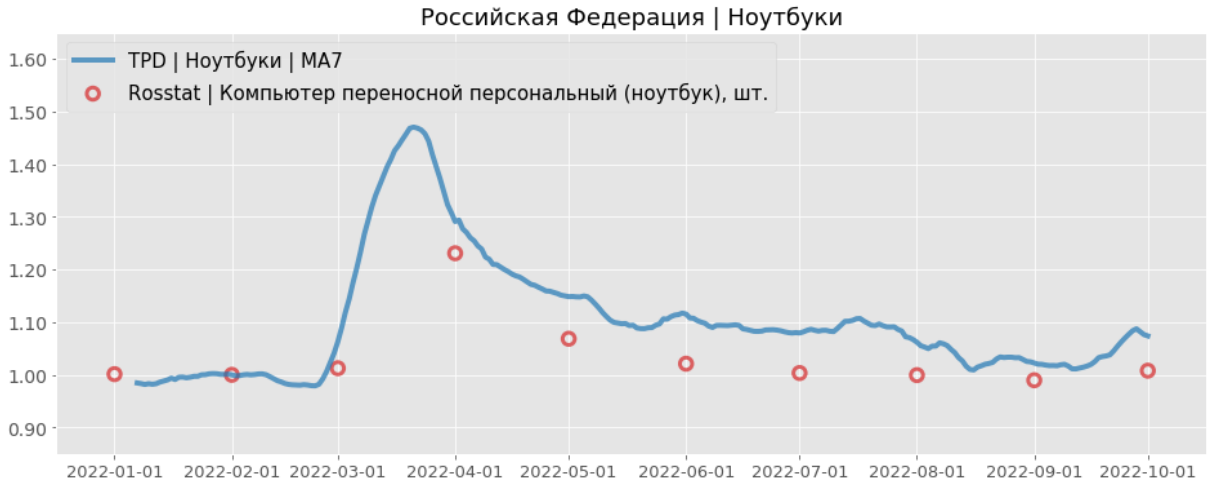
Индексы для непродовольственных товаров

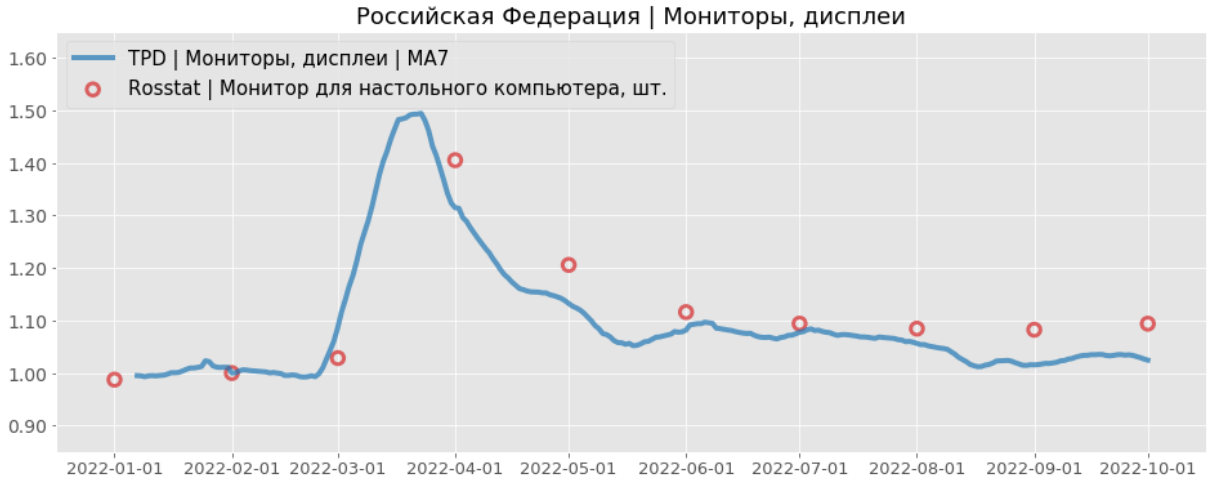


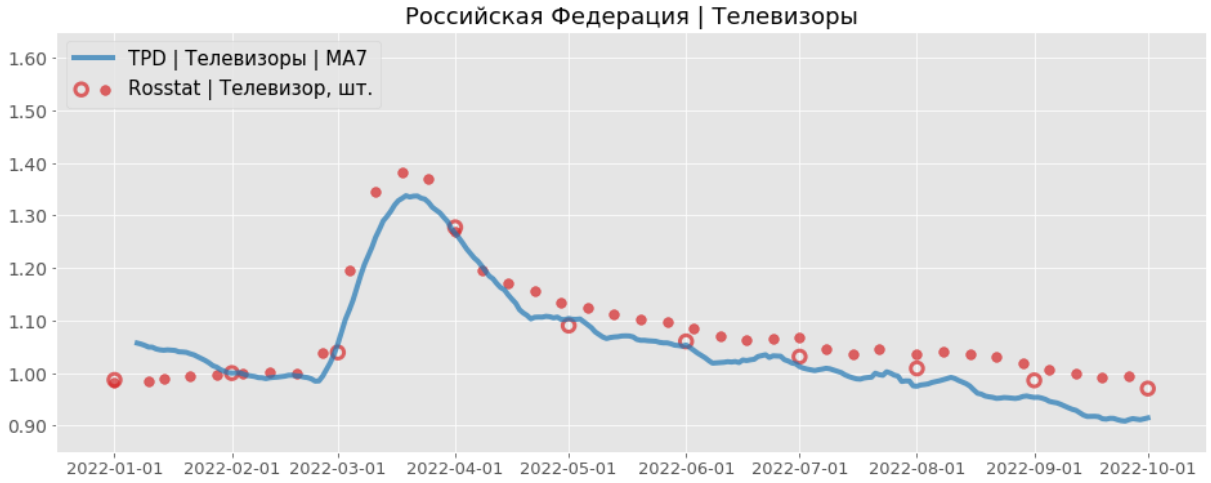


Индексы для непродовольственных товаров

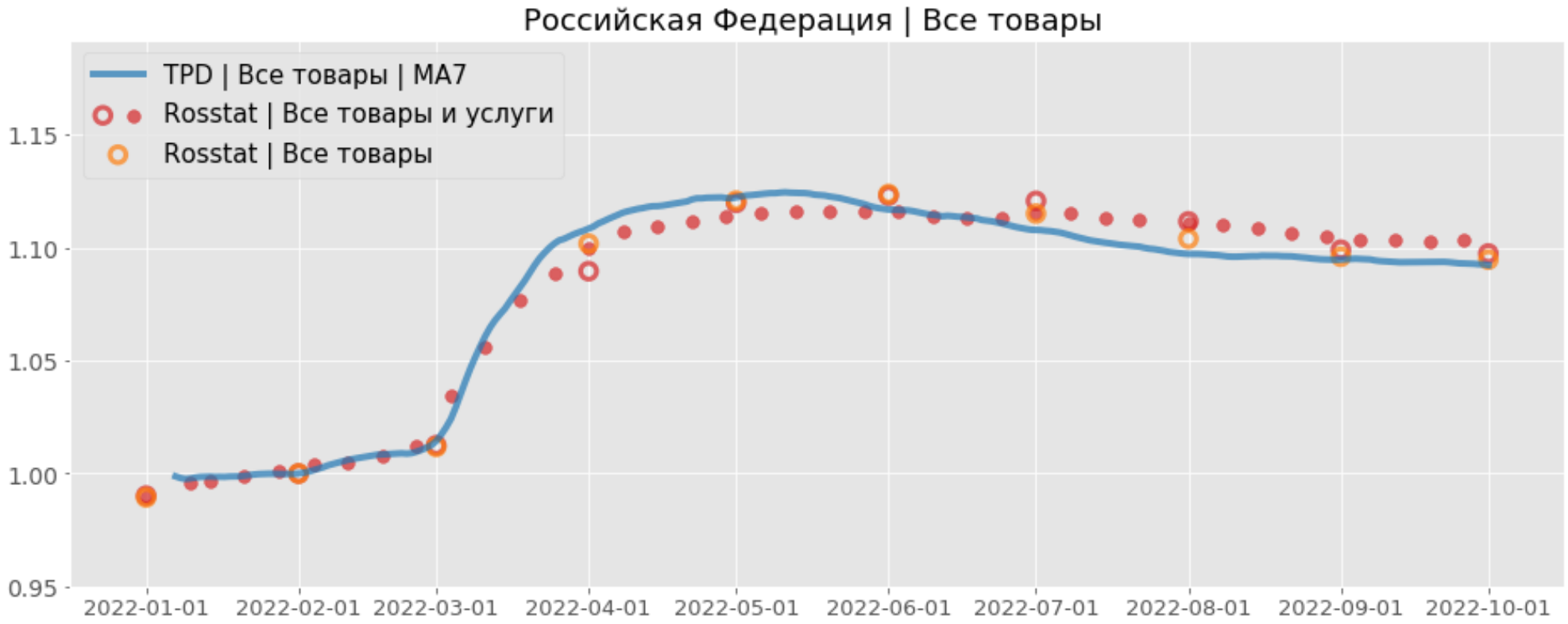








Индекс для всех товаров



Направления для доработки индекса

Определение товара

- Обработка наименований, встречавшихся лишь раз

Фильтрация аномалий

- Совершенствование существующих фильтров и добавление новых
- Автоматизация фильтрации наиболее волатильных товаров

Классификация

- Расширение количества категорий, для которых готовы классификационные модели
- Совершенствование классификационных алгоритмов

Формула индекса

- Проработка проблемы чувствительности индекса к смене ассортимента
- Тестирование индекса в условиях стабильной ценовой динамики

Спасибо за внимание!