



ФЕДЕРАЛЬНАЯ СЛУЖБА
ГОСУДАРСТВЕННОЙ СТАТИСТИКИ

«Разработка концепции использования «больших данных» (Big data) в государственной статистике с учётом международных рекомендаций. Разработка методологических подходов к использованию «больших данных» в отдельных отраслях статистики» (контракт №ST2/2/B.19)

Пестров Никита Николаевич, ведущий специалист по анализу данных ООО «Хабидатум Лаб»



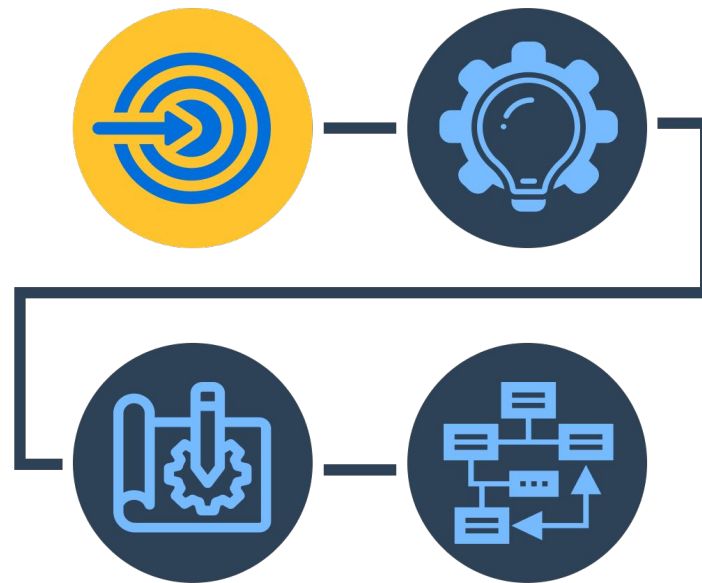
28 июня 2021 года



HABIDATUM

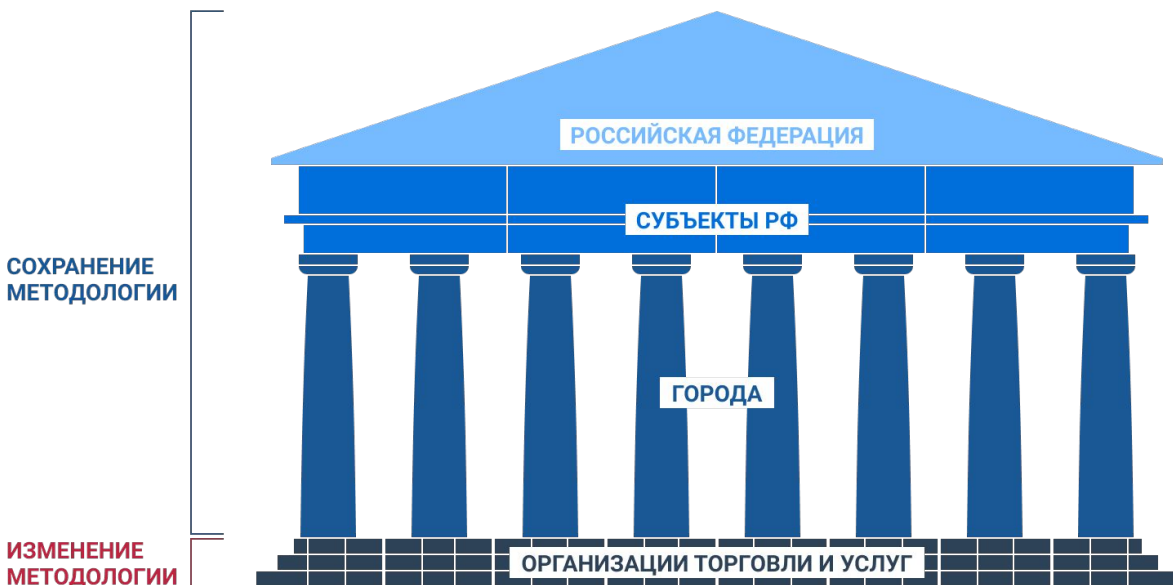
СТАТИСТИКА ПОТРЕБИТЕЛЬСКИХ ЦЕН

1. ОСНОВНЫЕ ЦЕЛИ И ЗАДАЧИ



ЦЕЛЬ РАБОТЫ





Использование Больших данных в рамках существующей методологии расчета индекса потребительских цен (ИПЦ)



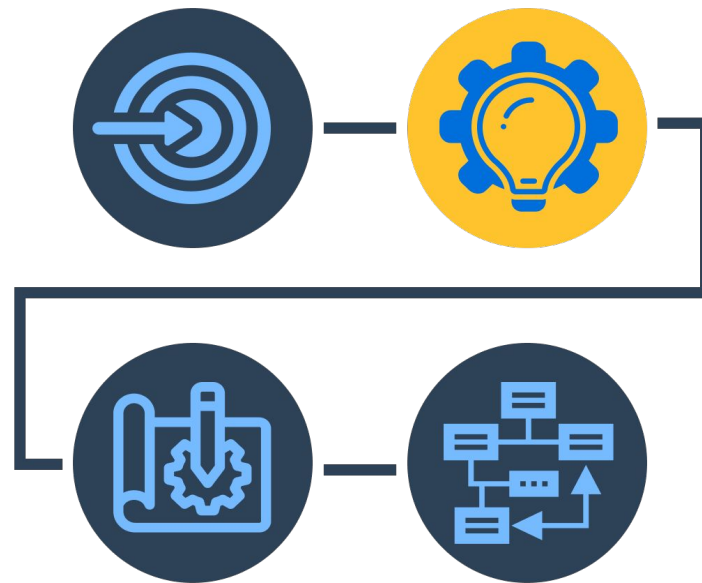
ЗАЧЕМ НУЖНЫ БОЛЬШИЕ ДАННЫЕ

- Расширение возможностей традиционной регистрации цен;
- Увеличение масштабов наблюдения за ценами на товары и услуги в организациях торговли и услуг;
- Детализация средних цен и индексов цен по товарам и услугам и в территориальном разрезе;
- Использование дополнительного источника информации для актуализации выборок товаров и организаций, а также весов для расчета индекса потребительских цен (объемы продаж товаров из чеков);
- Совершенствование существующей методологии расчета индекса потребительских цен.

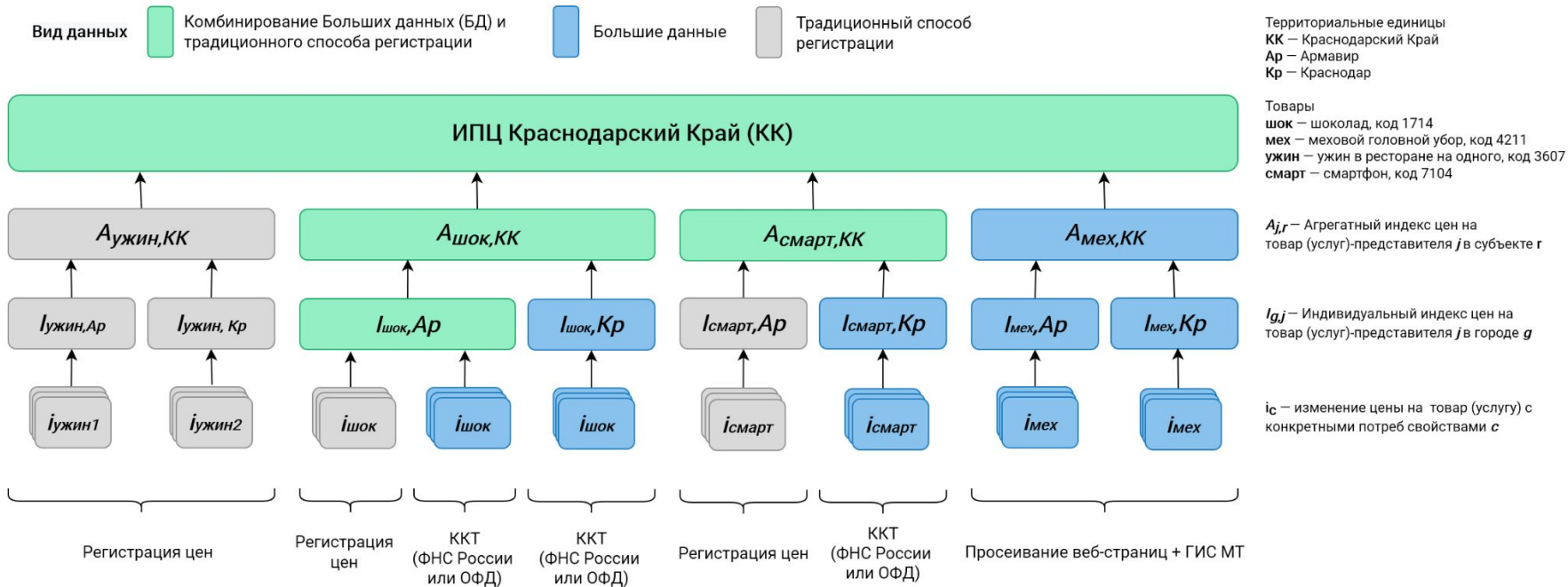
ГДЕ БЕРЕМ БОЛЬШИЕ ДАННЫЕ

	Источник	Формат взаимодействия	Тип данных
	ФНС России – контрольно-кассовая техника (далее — ККТ)	<ul style="list-style-type: none"> - Регулярное получение данных - Совершенствование формата выгрузок 	Данные спроса
	Оператор информации о маркированных товарах (ГИС МТ, ИС МДЛП) – ККТ	<ul style="list-style-type: none"> - Настройка регулярного получения выгрузок из ГИС МТ - Проект соглашения с Минздравом России, Росстатом и ООО “Оператор ЦРПТ” для получения данных о ценах на лекарственные препараты 	Данные спроса
	Онлайн-агрегаторы	<ul style="list-style-type: none"> - Котировки из открытой базы потребительских цен “Твердые цифры” от VTB Capital/РАНХиГС (сегодня география ограничена Москвой и МО) 	Данные предложения
	Операторы фискальных данных	<ul style="list-style-type: none"> - Подписание коммерческого договора для проведения пилота 	Данные спроса

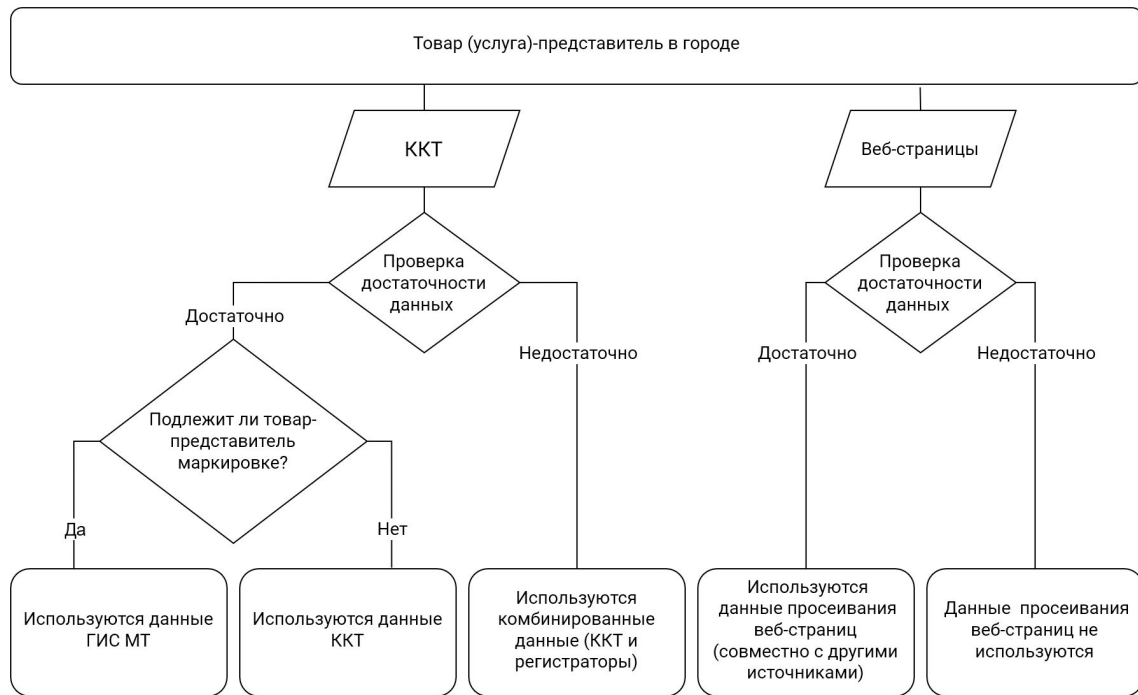
2. МЕТОДОЛОГИЯ



ПРИМЕР КОМБИНИРОВАНИЯ ДАННЫХ



ПРИНЦИПЫ КОМБИНИРОВАНИЯ НЕСКОЛЬКИХ ИСТОЧНИКОВ ДАННЫХ



Принципы:

- Определение типа регистрации происходит для **каждого города и товара (услуги)-представителя**;
- **Преимущественно используются Большие данные**;
- Большие данные используются в случае достаточного покрытия конкретного сегмента организаций в городе (например, сетевые);
- **Данные веб-страниц используются** наряду с данными ККТ и данными регистраторов;
- **Для маркированных товаров** используются данные ГИС МТ.

ОБРАБОТКА ИНФОРМАЦИИ ИЗ АЛЬТЕРНАТИВНЫХ ИСТОЧНИКОВ

Ежегодно

Группировка наименований товаров чеков ККТ со схожими свойствами (кластеринг)

Структурирование данных в соответствии с перечнем товаров(услуг)-представителей для расчета ИПЦ

Идентификация наименований товаров и услуг с конкретными потребительскими свойствами (матчинг)

Верификация сформированных искусственным интеллектом соответствий наименований товаров (кластеринга и матчинга)

Формирование перечней:

- организаций (на основе стабильного наличия в них распознанных наименований товаров и услуг)
- товаров (услуг)-представителей для расчета ИПЦ (на основе выделенных групп наименований и объемов продаж)
- ценовых котировок (основных и резервных)

Отсечение наименований товаров:

- не входящих в перечень товаров(услуг)-представителей для расчета ИПЦ
- без соответствующих единиц измерения
- не соответствующих уровням цен, определенных для товара-представитель

Определение методов регистрации цен для каждого товара/услуги в каждом городе (традиционный/комбинированный/большие данные)

Курсивом выделены задачи, требующие вовлечения регистраторов / ассессоров / экспертов

Регулярно

Отсечение

- неидентифицированных (нераспознанных) наименований товаров
- наименований товаров, не входящих в перечень товаров(услуг)-представителей для расчета ИПЦ
- наименований товаров без соответствующих единиц измерения
- ценовых котировок из организаций, не входящих в перечень объектов наблюдения
- скидок на товары
- котировок на основе ценовых трендов

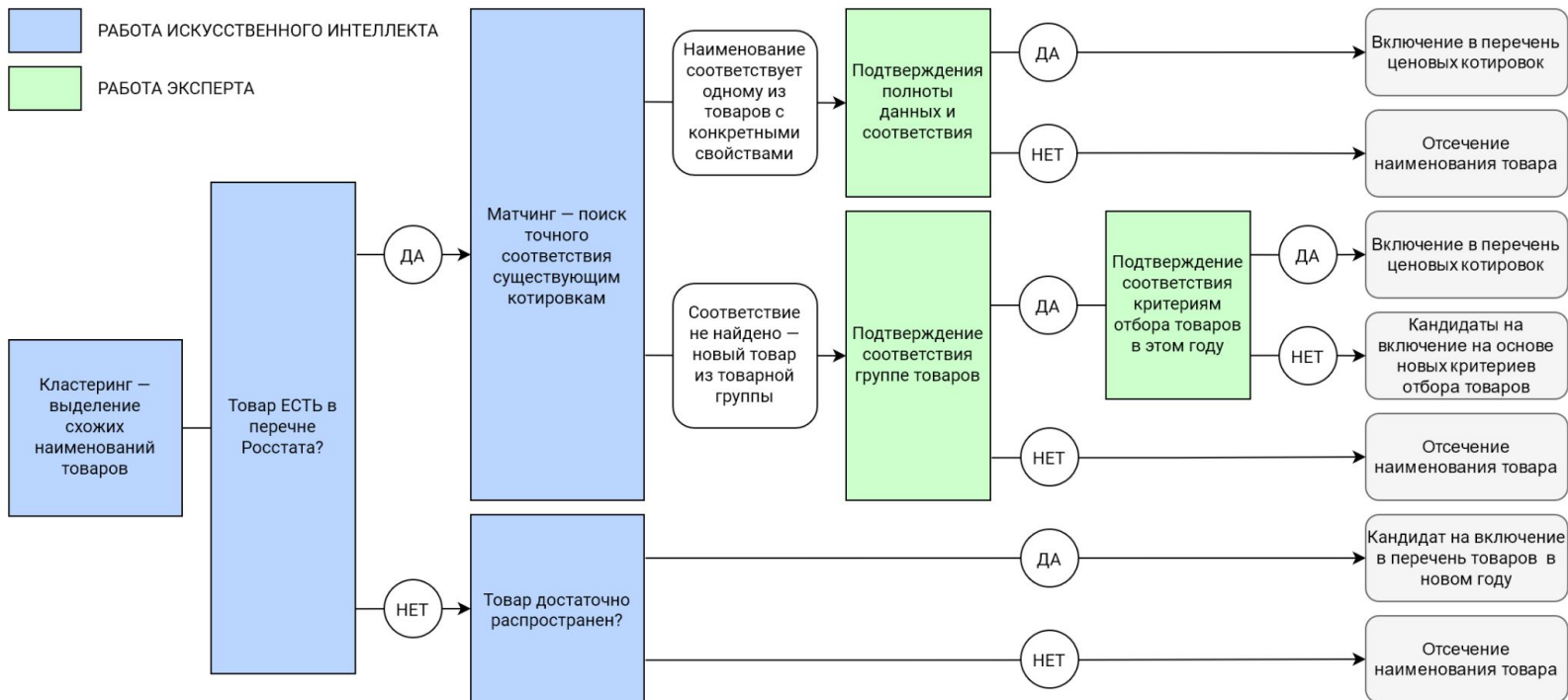
Накопление новых наименований товаров и услуг, объемов продаж

Соотнесение новых наименований товаров и услуг со списком товаров и услуг, идентификация товаров и услуг с конкретными потребительскими свойствами

Верификация сформированных искусственным интеллектом соответствий наименований товаров

Подтверждение замены товаров на основе резервных ценовых котировок, новых наименований товаров, смены упаковок

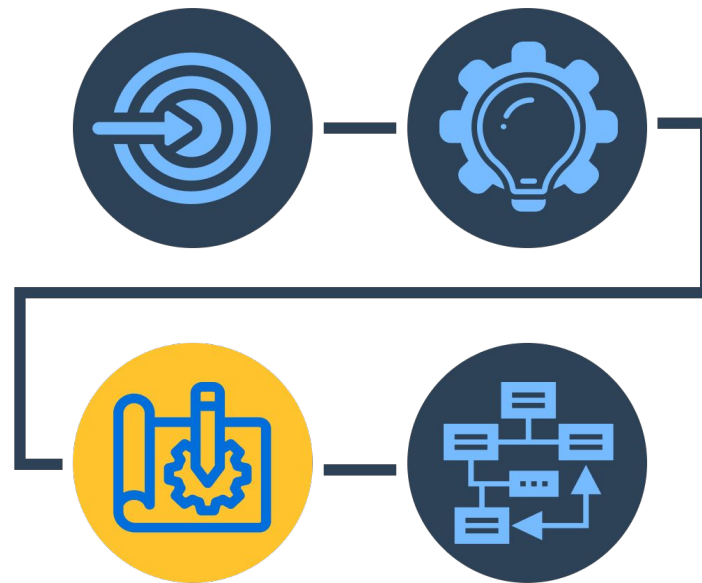
РАСПОЗНАВАНИЕ НАИМЕНОВАНИЯ ТОВАРОВ В ЧЕКАХ



СЛОЖНОСТИ И ОГРАНИЧЕНИЯ ИСПОЛЬЗОВАНИЯ БОЛЬШИХ ДАННЫХ

- Зависимость от устойчивости источников данных;
- Неравномерное покрытие данными в разных субъектах РФ, городах и товарных группах;
- Сопоставимость перечней организаций торговли и сферы услуг, а также ценовых котировок во времени;
- Данные о фактической продаже, а не предложении в большинстве источников;
- Отсутствие ИНН, точного адреса и наименования организации в существующем формате выгрузки ККТ.
- Определение временных параметров сбора данных, пропуски и выбор сопоставимых цен;
- Определение критериев отсеечения нерелевантных данных о ценах: пределы цен, скидочные товары;
- Методы импутации отсутствующих цен.

3. РЕЗУЛЬТАТЫ ПИЛОТНЫХ РАСЧЕТОВ



ПИЛОТНЫЕ РАСЧЕТЫ НА ОСНОВЕ ККТ

Источник: Данные ОФД с долей более 10% рынка ККТ в регионе.

Сопоставление по 33 товарам-представителям.

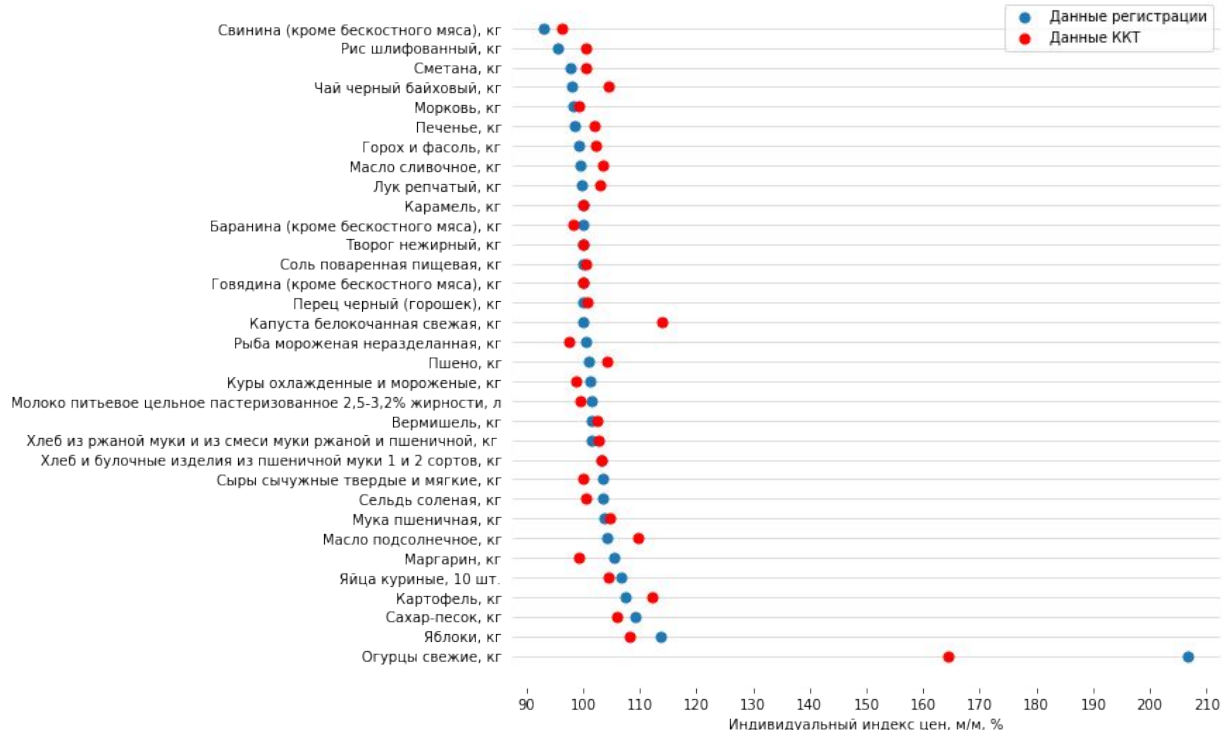
Периоды: IV кварталы 2019 и 2020 гг.

Регионы: Краснодарский край и Алтайский край.

Город	Число организаций из ККТ	Число котировок (наименования в различных организациях)
Краснодар (Краснодарский край)	70% организаций присутствовали в обоих годах	43% котировок были сопоставимы в обоих годах
Барнаул (Алтайский край)	75% организаций присутствовали в обоих годах	50% котировок были сопоставимы в обоих годах

РЕЗУЛЬТАТЫ СОПОСТАВЛЕНИЯ ИНДИВИДУАЛЬНЫХ ИНДЕКСОВ ЦЕН

Краснодар, ноябрь 2020



Исходные условия:

- Индивидуальные индексы цен по данным ККТ и данным регистрации;
- 33 товара в Краснодаре;
- Ноябрь 2020 г. к октябрю 2020 г.

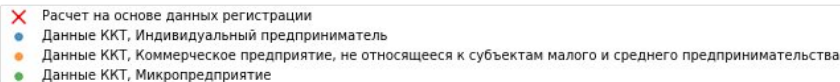
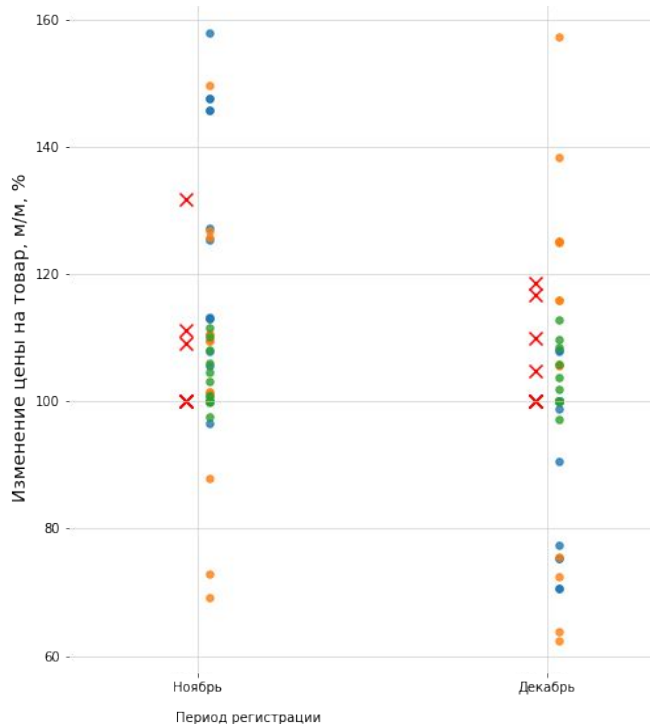
Результаты:

- Среднее расхождение индексов: 3,05 п.п.

Причины расхождений:

- Точность распознавания наименований товаров;
- Неполнота перечня организаций ККТ (выборка ОФД).

СОПОСТАВЛЕНИЕ ИЗМЕНЕНИЯ ЦЕН НА МАСЛО ПОДСОЛНЕЧНОЕ



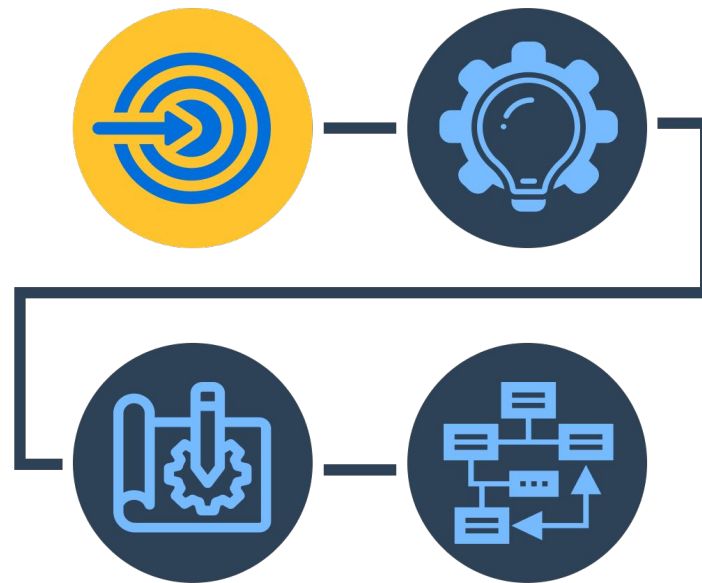
- Изменения цен по всем котировкам на Масло подсолнечное из всех организаций, в которых наблюдались цены на этот товар 21-25 числа месяца (с учетом отсеечения по уровню цен);
- Фильтрация наличных котировок в этот период регистрации во всех месяцах;
- Расчет изменения цены на основе котировки, полученной 23 числа, или средней цены в магазине в соседние даты;
- Разными цветами отражены котировки, полученные из разных типов организаций.

ОСНОВНЫЕ ВЫВОДЫ

- Использование Больших данных возможно с сохранением существующей методологии расчета ИПЦ;
- Изменение технологии сбора информации происходит на уровне “организация-город”;
- Для максимального покрытия необходимо комбинирование традиционных и альтернативных источников;
- Структурирование данных требует работы искусственного интеллекта и контроля экспертов;
- Анализ информации об объемах продаж отдельных товаров позволяет оценить потребительский спрос на конкретные товары-представители и использовать его в качестве весов;
- Ежегодное составление выборки перечней обследуемых товаров, котировок и организаций из альтернативных источников основано на существующих методологических принципах;
- Для формирования выборок на основе Больших данных необходимо использование алгоритмов отсеечения ценовых котировок и наименований товаров;
- Результаты пилотных расчетов демонстрируют незначительные отклонения цен из альтернативных и традиционных источников.

СТАТИСТИКА РОЗНИЧНОЙ ТОРГОВЛИ

1. ВВЕДЕНИЕ



РАССМАТРИВАЕМЫЕ ПОКАЗАТЕЛИ

В работе **разработаны и описаны алгоритмы анализа данных и расчета** основных показателей статистики торговли: оборота розничной торговли и общественного питания, распределение оборота по категориям товаров, по типам хозяйствующих субъектов, а также объема платных услуг.

Основная задача при распределении оборота по показателям: **корректное** отнесение оборотов по **типу продаж**

В рамках пилота **рассчитаны** следующие показатели:

- Оборот розничной торговли
- Оборот розничной торговли продовольственными товарами
- Оборот розничной торговли непродовольственными товарами
- Оборот общественного питания
- Объем платных услуг

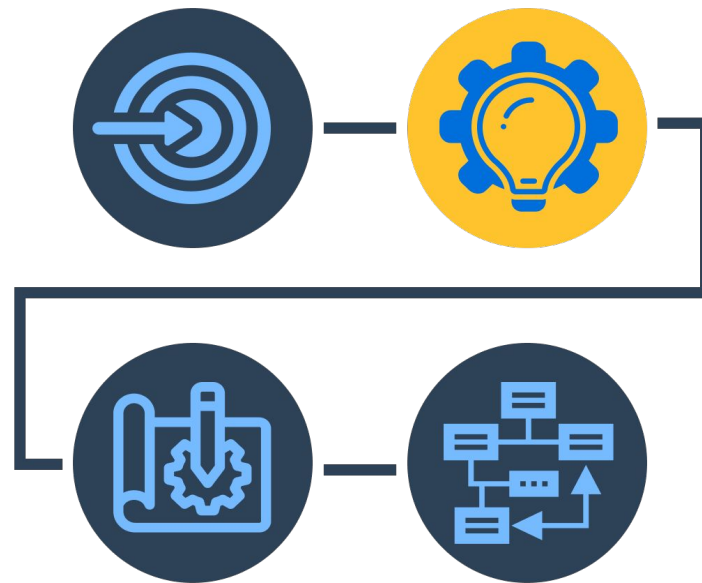
БОЛЬШИЕ ДАННЫЕ В СТАТИСТИКЕ ТОРГОВЛИ

- Расширение информационной базы, используемой для расчета показателей в соответствии с действующей методологией;
- Расширение перечня наблюдаемых товаров и услуг в общем товарообороте, использование детальных категорий продаж;
- Повышение оперативности формирования показателей;
- Переход к сплошному обследованию предприятий торговли, относящихся к субъектам малого предпринимательства;
- Освобождение некоторых типов хозяйствующих субъектов от статистической отчетности;
- Увеличение пространственной детализации наблюдения до уровня муниципальных образований.

ТЕКУЩАЯ СИТУАЦИЯ: ВАРИАНТЫ ИСПОЛЬЗОВАНИЯ ККТ

<i>Основа формирования выборки</i>	<i>Возможности</i>	<i>Ограничения</i>
Данные ФНС России (текущая ситуация)	Использование выгрузки возможно только для <u>обучения</u> системы распознавания н	Деперсонифицированные данные; невозможно проводить расчеты: нет ID торговой организации, атрибутов ККТ
Данные ОФД (пилотный проект)	Сопоставление списков ИНН от Росстат с ИНН от ОФД, обогащение выборки новыми атрибутами из данных ККТ	Деперсонифицированные данные, нет возможности сопоставления предприятий на уровне ИНН; информация не передаётся, если в сегменте менее 3-х ККТ
Данные ФНС России без ИНН (альтернативное решение, в проработке)	Получение <u>дополнительных атрибутов</u> из данных ККТ, возможность работы с данными индивидуальных объектов	Решение находится в проработке, нет четких сроков перехода на новый формат данных ККТ. <u>Нет возможности пообъектного сопоставления.</u>
Данные ФНС России с ИНН (будущее решение, в проработке)	Персонифицированные данные, получение всех дополнительных атрибутов из данных ККТ, возможность сопоставления предприятий на уровне ИНН	Решение находится в проработке, <u>нет четких сроков перехода</u> на данный формат работы с данными ККТ, необходимость внесения изменений в 102 статью НК РФ

2. МЕТОДОЛОГИЯ



СЦЕНАРИИ ИСПОЛЬЗОВАНИЯ ДАННЫХ ККТ

1. **Полная замена** — переход на данные ККТ для выбранных показателей, отказ от использования федеральных форм. Реализуема в долгосрочной перспективе;
2. **Смешанное наблюдение** — смешанное наблюдение, когда полноценный перевод расчета показателя на данные ККТ невозможен, но возможно заменить ККТ расчеты по части хозяйствующих субъектов;



АЛГОРИТМ ИСПОЛЬЗОВАНИЯ ДАННЫХ ККТ

1. **Обработка данных ККТ** ФНС России (преобразование в таблицы, распознавание наименований товаров, отсечение опта и выделение основного типа продаж хоз. субъектов);
2. Обработка данных, полученных **традиционными методами** обследования (приведение данных Росстата из форм обследований к единой структуре);
3. **Сопоставление данных** ККТ и данных официальной статистической отчетности (обогащение перечня атрибутами из данных ККТ, распределение объема продаж ККТ по показателям розничной торговли, общественного питания и платных услуг, корреляция полученных значений);
4. **Расчет** показателей розничной торговли, общественного питания и платных услуг (с использованием коэффициентов досчета).

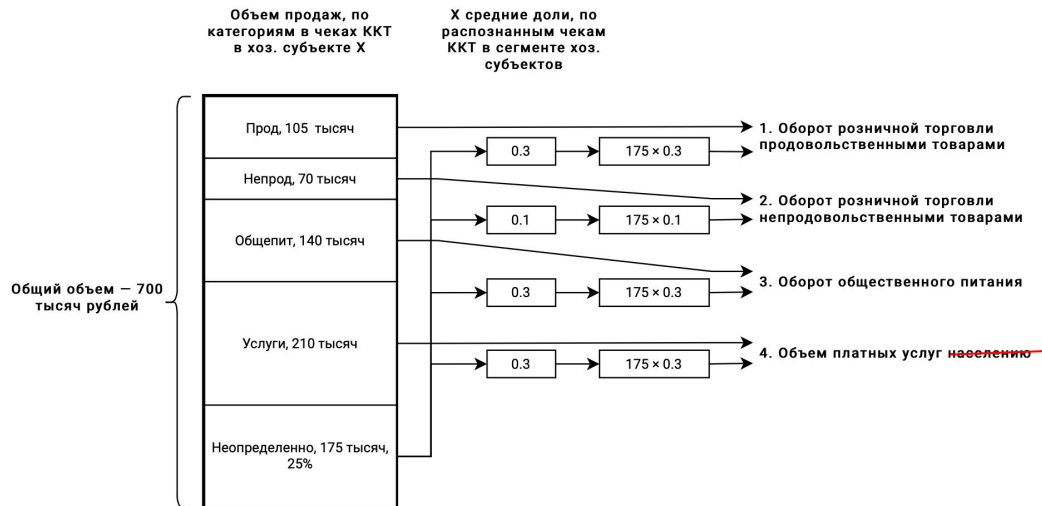


2. РАСПРЕДЕЛЕНИЕ ОБЪЕМА ПРОДАЖ ККТ ПО ПОКАЗАТЕЛЯМ ТОРГОВЛИ

Распределение на основе распознавания части наименований товаров и услуг в чеках ККТ

Нераспознанные наименования и услуги распределяются в соответствии со средними соотношениями между показателями в выбранном сегменте.

Сегмент — перечень хоз. субъектов, с одним типом (например, малые) основным кодом ОКВЭД (например, 47.11), расположенных в одном субъекте РФ.



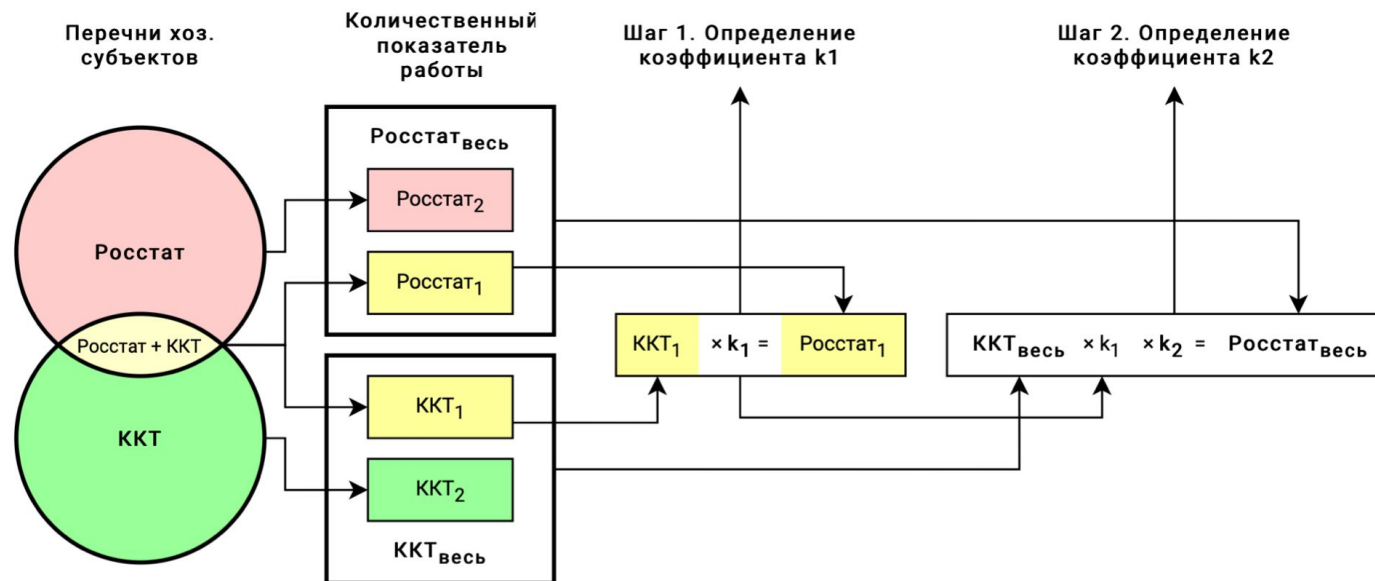
3. КОЭФФИЦИЕНТЫ ДОСЧЕТА (1)

Необходимость использования коэффициентов досчета вызвана двумя факторами:

1. Наличие нераспределенных по типам продаж позиций в чеках и объемов, не учтенных в ККТ (льготы для налогоплательщиков)
2. Различия в перечнях хозяйствующих субъектов, для которых есть данные ККТ и данные форм статистического наблюдения.



3. КОЭФФИЦИЕНТЫ ДОСЧЕТА (2)



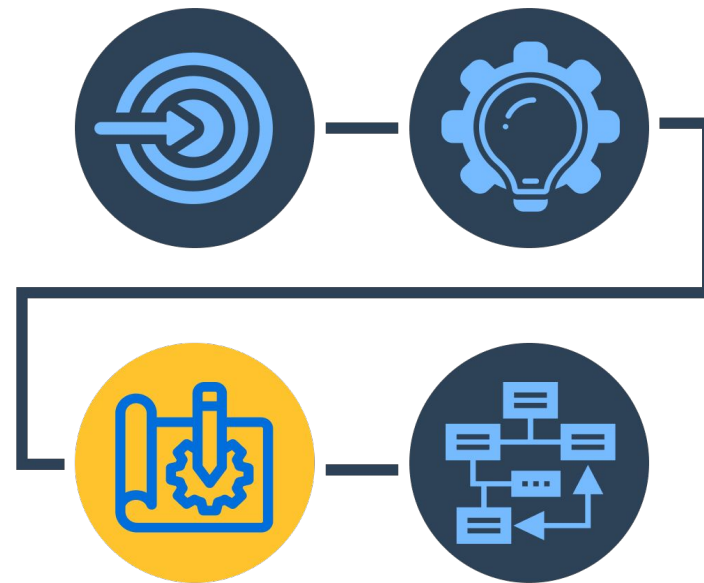
Решение:

Последовательное применение двух типов коэффициентов досчета — до сопоставимого перечня субъектов (K_1) и до генеральной совокупности (K_2).

СЛОЖНОСТИ ИСПОЛЬЗОВАНИЯ БОЛЬШИХ ДАННЫХ

- Отсутствие необходимых реквизитов в существующей выгрузке ККТ от ФНС России — ИНН хозяйствующего субъекта и другие дополнительные атрибуты (код ОКЭД, тип хозяйствующего субъекта, форма собственности);
- Выборочный характер обследования некоторых типов хозяйствующих субъектов существующими формами федерального статистического наблюдения;
- Алгоритмическое отсеечение оптовых продаж и продаж недвижимости без наличия явного обязательного реквизита;
- Невозможность распределения объемов Интернет-торговли по субъектам РФ;
- Использование коэффициентов досчета.

3. РЕЗУЛЬТАТЫ ПИЛОТНЫХ РАСЧЕТОВ



ПИЛОТНЫЕ РАСЧЕТЫ НА ОСНОВЕ ККТ (ОФД)

Источник: Данные ОФД с долей более 10% рынка ККТ в регионе.

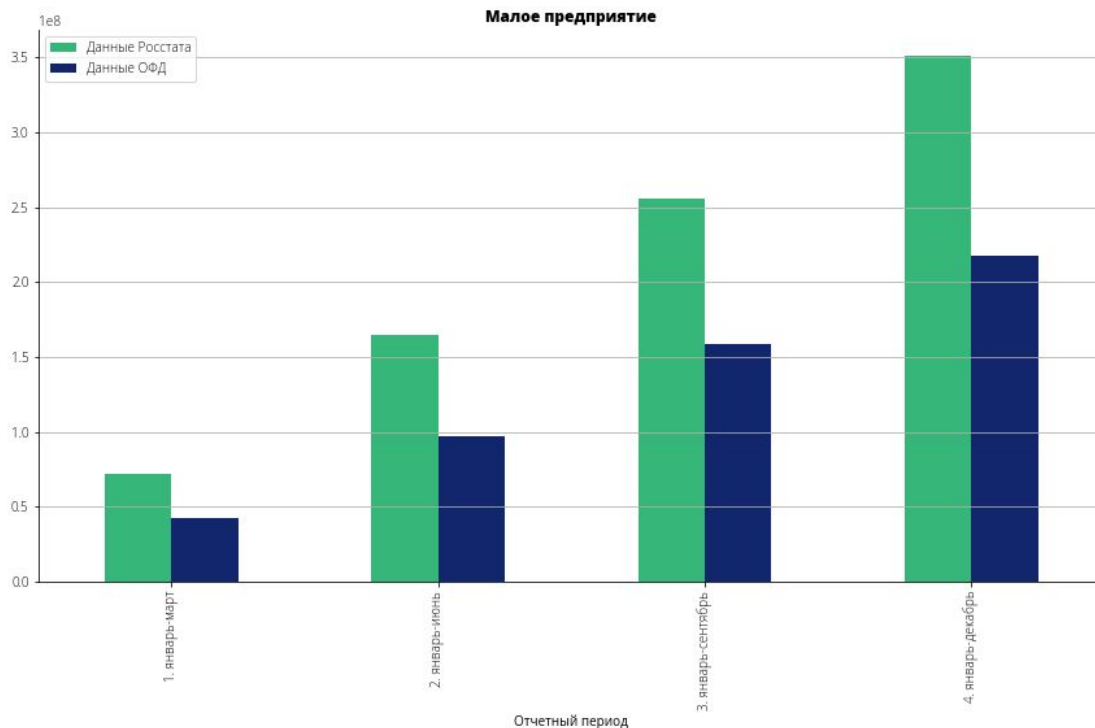
Сопоставление по ИНН.

Периоды: 2020 год ежемесячно.

Регионы: Краснодарский край и Алтайский край.

Город	ИНН Росстат	ИНН выделенные ОФД
Краснодарский край	79 413	7 584 (~ 10%)
Алтайский край	34 213	868 (~ 3%)

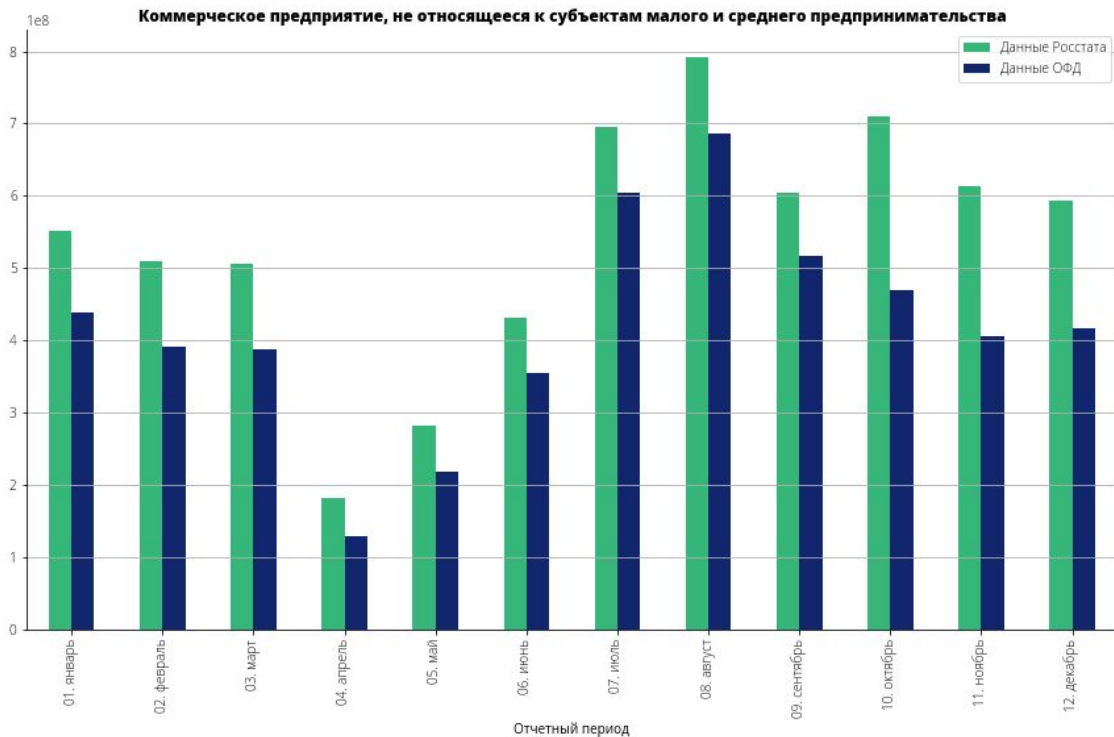
ОБОРОТ РОЗНИЧНОЙ ТОРГОВЛИ, ПРИМЕР ДЛЯ МАЛЫХ ПРЕДПРИЯТИЙ (АЛТАЙСКИЙ КРАЙ)



Для малых предприятий наблюдается стабильное отклонение в течение всего 2020 года по кварталам – 38-41% превышение в сторону Росстата, и при этом минимальное отклонение (1.7%).

Показатели рассчитаны по сопоставимому перечню хозяйствующих субъектов.

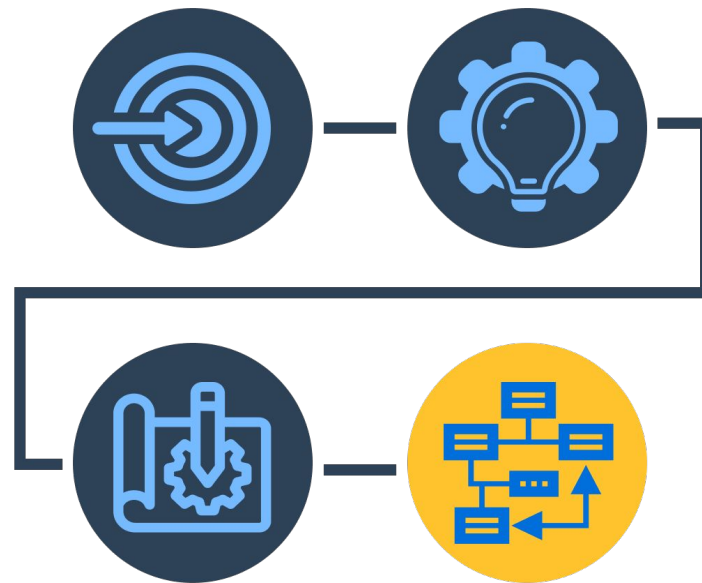
ОБОРОТ ОБЩЕСТВЕННОГО ПИТАНИЯ, ПРИМЕР ДЛЯ КРУПНЫХ ПРЕДПРИЯТИЙ (КРАСНОДАРСКИЙ КРАЙ)



Крупные предприятия показывают стабильные отклонения по месяцам в сторону данных Росстат; переоценка отличается от квартала к кварталу – 23% в 1 квартале, 13% в 3 квартале 13%, рост до 29-33% в четвертом.

Показатели рассчитаны по сопоставимому перечню хозяйствующих субъектов.

РАЗВИТИЕ ПРОЕКТА В СТАТИСТИКЕ ПОТРЕБИТЕЛЬСКИХ ЦЕН И РОЗНИЧНОЙ ТОРГОВЛИ



ЭТАПНОСТЬ ПРИМЕНЕНИЯ БОЛЬШИХ ДАННЫХ

