Банк России

Июнь 2021

Построение индексов цен на основе данных контрольно-кассовой техники: ключевые вопросы и вызовы

Аналитическая записка

Гришина Т., Долгов Д., Мамедли М., Милютин П., Поршаков А., Селезнев С.

ОГЛАВЛЕНИЕ

Фото на обложке: Shutterstock.com/FOTODOM

Адрес: 107016, Москва, ул. Неглинная, 12
Телефон: +7 (495) 771-99-99 (доб. 694-74)
Официальный сайт Банка России: www.cbr.ru

РЕЗЮМЕ

Использование больших данных (Big Data) открывает перспективы для совершенствования официальных статистических показателей. Большой объем информации о сделках, покупках и различного рода транзакциях, которая раньше была недоступна ввиду непомерных затрат на ее сбор, обработку и хранение, теперь потенциально открыта статистическим ведомствам. В 2020 году Росстат высказал заинтересованность в переходе с традиционного расчета индекса потребительских цен (ИПЦ) на основе ручного сбора информации на расчет с использованием данных контрольно-кассовой техники (данных фискальных чеков), собираемых Федеральной налоговой службой.

Такой подход обладает очевидной перспективностью и огромным потенциалом для использования в статистических и исследовательских целях. Вместе с тем существует ряд прикладных проблем, которые осложняют построение индексов цен на основе чеков, в том числе ИПЦ. Обсуждению и иллюстрации этих проблем посвящена данная аналитическая записка.

В записке показывается, что необходимо учитывать специфику сбора и представления информации в данных контрольно-кассовой техники. Эта специфика обусловлена в том числе наличием скидок и промоакций, нерегламентированными наименованиями товаров и услуг, а также пропусками в данных вследствие отсутствия фактических продаж. В результате в значительной мере ограничивается возможность построения практически всех общепринятых в официальной статистике показателей ценовой динамики.

Мы также обсуждаем возможные пути решения возникающих проблем и демонстрируем один из индексов, построенный с учетом всех рассмотренных ограничений на имеющейся выборке данных. Полученный индекс цен фактически сделанных покупок подтвердил корректность динамики инфляции, отражаемой публикуемым Росстатом ИПЦ. К такому выводу удалось прийти, несмотря на различие в методологии построения и часто отличающуюся динамику, путем сравнения индексов цен в региональном и товарном разрезах.

1. ВВЕДЕНИЕ

Денежно-кредитная политика (ДКП), проводимая Банком России, направлена на поддержание ценовой стабильности. Принимая решения по ДКП, регулятор ориентируется в том числе на широкий спектр макроэкономических показателей и оперативных индикаторов, информация по которым важна для выявления устойчивой и временной составляющих наблюдаемой ценовой динамики, а также определения баланса факторов, которые могут нести риски среднесрочного отклонения инфляции как вверх, так и вниз от цели. Цель Банка России по годовой инфляции (4%) сформулирована в терминах инфляции по ИПЦ на конец года.

Несмотря на то, что совокупный ИПЦ включает в себя ряд традиционно волатильных и регулируемых компонент, на ценовую динамику которых центральный банк не может оказывать значимого влияния своими действиями, выбор такого показателя является наиболее распространенной практикой среди регуляторов в мире, которые проводят ДКП в режиме таргетирования инфляции. Так, динамика ИПЦ является не только наблюдаемым показателем в официальной статистике, но также и показателем изменения цен в экономике, который в наибольшей степени узнаваем широким кругом экономических агентов. Этот показатель в той или иной степени (в зависимости от заякоренности инфляционных ожиданий и доверия к ДКП) служит для субъектов экономики в качестве ориентира для принятия ими финансовых решений.

В настоящее время идет дискуссия о направлениях совершенствования статистического учета цен. Так, в 2020 году Росстат выразил заинтересованность в переходе с расчета ИПЦ на основе ручного сбора данных к расчету с использованием данных контрольно-кассовой техники[1,2] (ККТ). В отличие от существующей методологии, а также быстро развивающихся альтернативных подходов, основанных на сборе данных с использованием веб-скрейпинга[3], новый подход потенциально претендует на расчет ценовых индексов практически по всей генеральной совокупности цен. Технически это способно повысить скорость и полноту сбора информации о ценах и точность публикуемой статистики цен. Это важно для мониторинга и анализа Банком России ценовой динамики на этапе формулирования ДКП.

Однако такой подход порождает ряд прикладных проблем, которые нужно решить уже на этапе разработки методологии. В настоящей записке мы формулируем эти методологические проблемы (возникающие не только при построении ИПЦ, но и при работе с ценовыми индексами на основе больших данных в принципе) и предлагаем возможные пути их решения. Для этого мы используем в том числе пример индекса цен, построенного нами на данных ККТ.

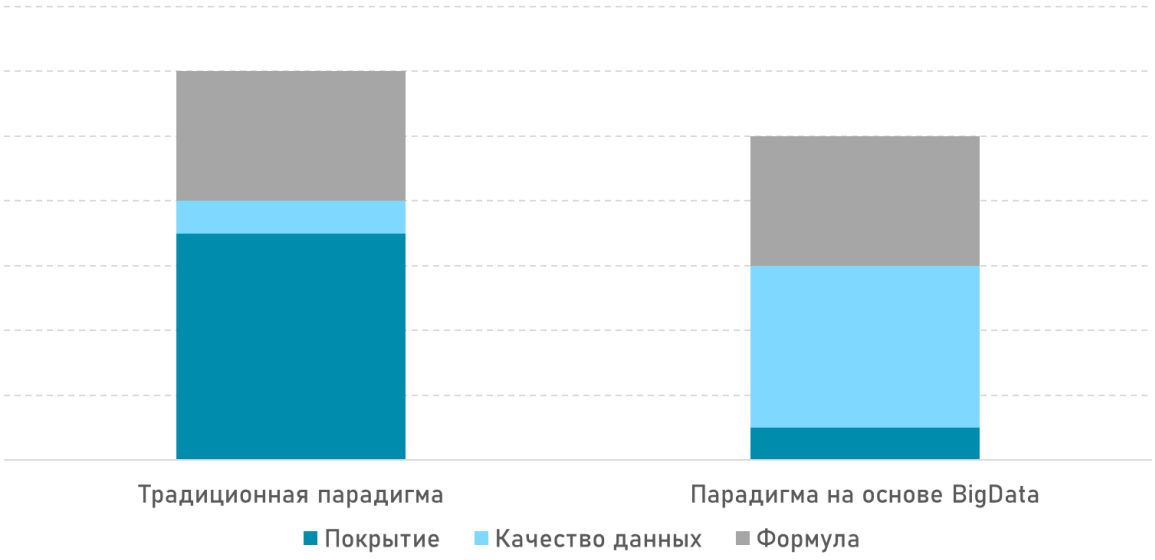
[1] Глава Росстата заявил об идее изменить расчет индекса потребительских цен.
[2] Росстат начнет считать инфляцию на основе «миллионов» цен.
[3] См, например, Хабибуллин и Яковлева (2019) и Исаков и др. (2021).

Важной отправной точкой для нашего анализа является четкое разграничение между двумя парадигмами при построении статистических показателей:

1) традиционная парадигма (которая лежит в основе текущей официальной статистики цен);
2) парадигма больших данных.

С одной стороны, построение статистических показателей на основе больших массивов данных по своей сути похоже на построение традиционных (классических) статистических показателей и содержит все те же этапы: сбор первичных данных, их обработку, выбор формулы расчета и непосредственно расчет. *С другой стороны*, принципиальное различие между парадигмами заключается в том, какие погрешности вносятся на этих этапах в финальный «идеальный» показатель[4]. На Рисунке 1 изображен абстрактный иллюстративный пример декомпозиции погрешности в оценках ИПЦ, которая накапливается в связи с различными источниками ошибок и неточностями.

Рисунок 1. Декомпозиция погрешности при построении статистических показателей



Основное отличие подходов демонстрируют голубая и синяя области столбцов (*«Покрытие»* и *«Качество данных»* соответственно). Так, подход на основе больших данных позволяет работать с ощутимо большим числом наблюдений по сравнению с традиционным подходом, что помогает уменьшить ошибку за счет большего покрытия (*синие столбцы «Покрытие»*). В рамках традиционного подхода сбор статистической информации по ценам во многом ориентирован на человеческие ресурсы и ручной

[4] Под «идеальным» показателем здесь понимается показатель, который рассчитан на генеральной совокупности и построен исходя из конкретной задачи (например, ценовой индекс, таргетирование которого центральным банком помогает максимизировать общественное благосостояние). Важно отметить, что он не обязательно должен совпадать с официальными показателями даже в рамках выбранной математической формулы.

режим работы, в то время как большие данные собираются автоматически и пока что в основном не с целью расчета статистических показателей. Соответственно, можно ожидать, что традиционный подход к сбору данных предполагает сравнительно меньшее количество ошибок и «мусорной» информации, а построение статпоказателя в парадигме больших данных, наоборот, несет риск получения существенных объемов нерелевантной информации (*голубая область столбца «Качество данных»*).

В свою очередь, *серые области столбцов* (*«Формула»)* обозначают вклад математической формулы, выбранной для расчета. На первоначальном этапе анализа логично предположить их по умолчанию одинаковыми для традиционной парадигмы и парадигмы больших данных. Эти столбцы могут быть как больше остальных, если формула далека от теоретически оптимального индекса, который в общем случае не известен и может отличаться от принятых в официальной статистике показателей (см. сноску 4), так и практически нулевыми, если измеряется то же самое, что и рекомендует теория. Важно лишь отметить, что эти ошибки зависят только от расхождения теоретической и практической формул и не зависят от выбранной парадигмы.

В этой записке мы концентрируемся главным образом *на проблеме качества данных*, то есть на голубых областях столбцов, и показываем причины, по которым в текущий момент для данных ККТ связанная с этим аспектом погрешность достаточно велика. Мы также обсуждаем, какие индексы цен возможно построить с использованием данных ККТ.

Важно отметить, что настоящая записка не ставит целью критически проанализировать качество измеряемого Росстатом ИПЦ, а лишь очерчивает круг вопросов, которые нуждаются в решении, для того чтобы статистика цен могла считаться с использованием данных фискальных чеков.

Дальнейшее содержание записки структурировано следующим образом. **Раздел 2** знакомит читателей с данными ККТ. В **Разделе 3** обсуждаются вопросы и проблемы построения индекса. **Раздел 4** посвящен примеру построения индекса на имеющейся выборке из данных ККТ. В **Разделе 5** представлено заключение.

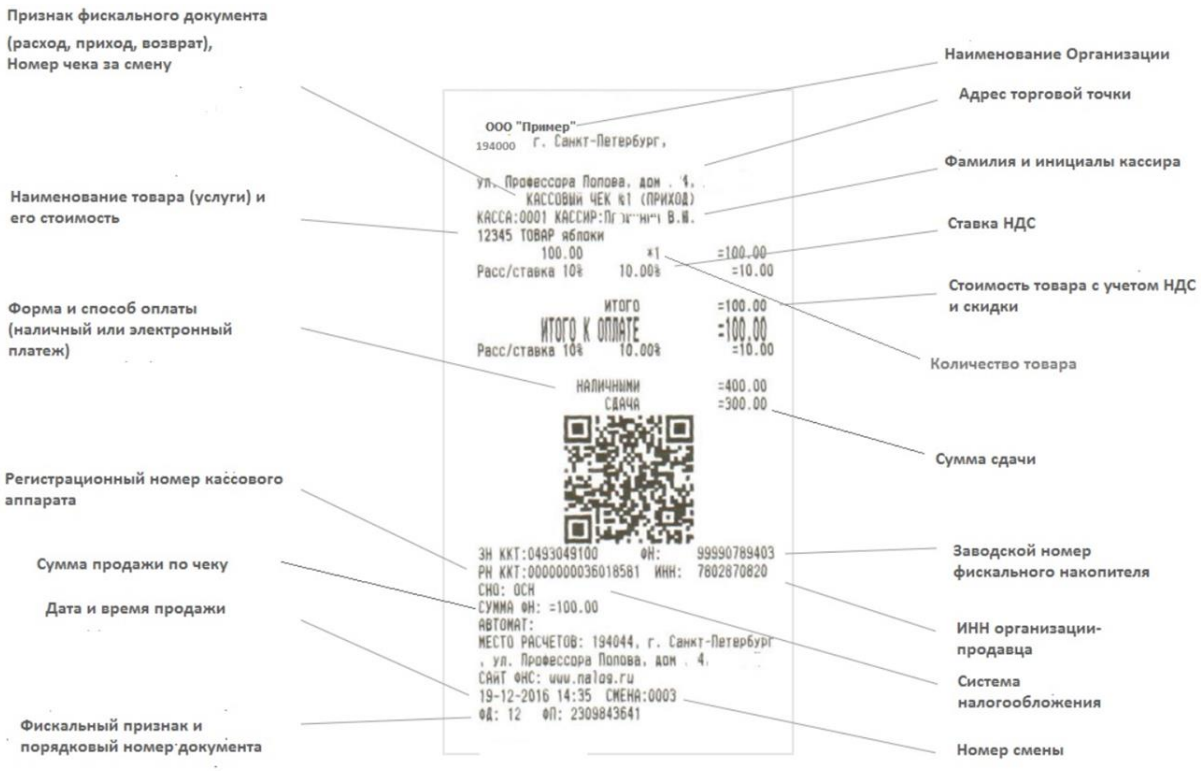
2. ДАННЫЕ ККТ

Данные ККТ собираются Федеральной налоговой службой (ФНС) со второй половины 2016 года и содержат подробную информацию о каждом чеке, прошедшем через систему. Из каждого чека в систему попадает информация о купленных товарах, включая наименование товара, его стоимость и проданное количество, а также ряд дополнительных полей о месте продажи, способе оплаты и так далее (пример чека со списком ключевых полей изображен на Рисунке 2).

К настоящему моменту мы проанализировали данные по чекам за период *с октября 2016 по сентябрь 2018 года*. Эти данные обезличены, то есть в них не содержится никакой информации, позволяющей однозначно идентифицировать ни

торговую точку, в которой была совершена покупка, ни покупателя. Вместе с тем для аналитических целей присутствует идентификатор, с помощью которого можно, например, определить, что две покупки были совершены в одной кассе. Полный список доступных полей приведен в Приложении 1.

Рисунок 2. Пример фискального чека



Источник: Первый ОФД

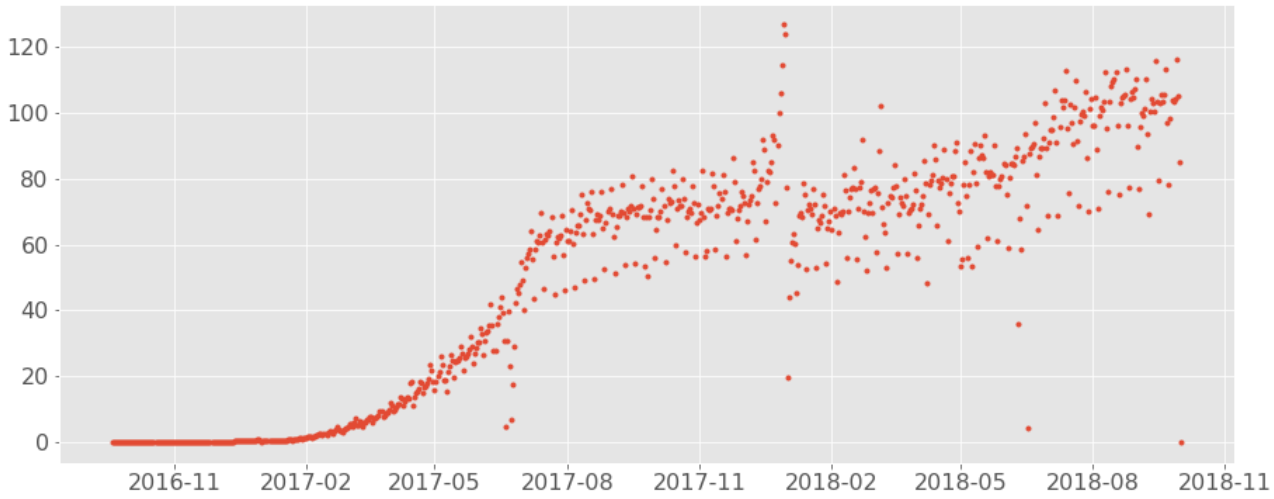
Процесс подключения розницы и услуг к системе ККТ не закончен и к настоящему моменту, а в период с 2016 по 2018 год доля покрытия могла быть совсем невелика. Для того чтобы оценить покрытие чеков, объемы продаж розницы по данным Росстата[5] за 2018 год в региональном разрезе были сопоставлены нами с соответствующими данными за скользящий год (с октября 2017 по сентябрь 2018 года) на основе ККТ. В базе присутствует полный 2017 год, но, как видно из Рисунка 3, из-за активной фазы перехода товарных точек на ККТ до середины 2017 года шел активный рост объемов, поэтому использование 2017 календарного года при оценке покрытия сильно занизит результаты. Рисунок 4 демонстрирует, что для подавляющего большинства регионов покрытие составляет более 60% и примерно для половины – больше 80%[6]. Таким образом, уже на конец 2018 года в большинстве

[5] В рассматриваемый период услуги в базе почти не представлены.

[6] Наличие регионов с оценкой более 100% может быть связано с несколькими причинами: разницей в сопоставляемых периодах времени, попаданием услуг в данные ККТ, попаданием транзакций, которые не относятся к рознице или услугам, в базу данных по ККТ, различием в региональной принадлежности в данных Росстата и ККТ или другими особенностями методологи Росстата. Так как упражнение по оценке покрытия

регионов данных было достаточно для того, чтобы использовать их для усовершенствования текущей методологии построения индексов цен. Поэтому проблема совершенствования традиционной методологии по большей части находится именно в плоскости качества и обработки больших данных.

Рисунок 3. Подневные объемы продаж, попавших в базу данных ККТ (млрд руб.)



3. СЛОЖНОСТИ ПРИ ПОСТРОЕНИИ ИНДЕКСА ЦЕН

Для того чтобы выявить сложности и ограничения при построении индекса цен с использованием данных ККТ, мы попытались пройти все необходимые после сбора данных шаги. Построение индекса цен можно разбить на следующие этапы:

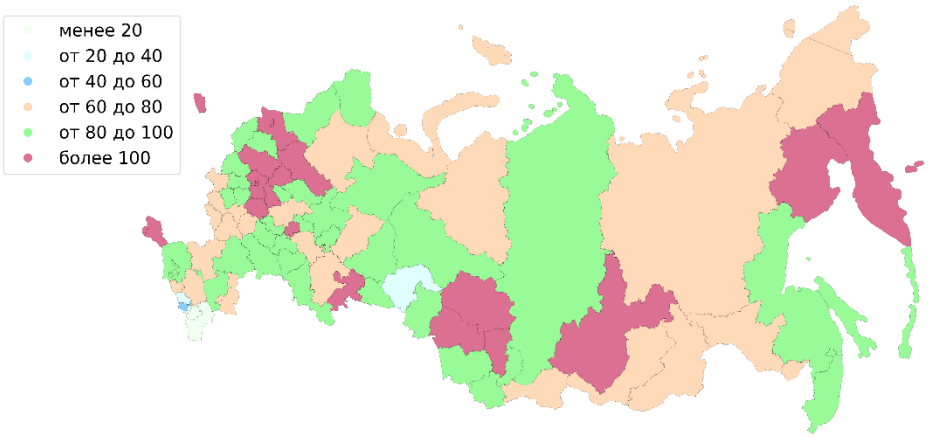
1. Выбор формулы.
2. Расчет индексов цен по индивидуальным товарам[7].
3. Объединение товаров в товарные группы и расчет агрегированных индексов, субиндексов и так далее.

Как упоминалось во введении, выбор формулы должен зависеть от задачи построения индекса. Поэтому мы не будем выбирать конкретную формулу в этом разделе, а сначала обсудим трудности, возникающие на 2 и 3 этапах, и потенциальные пути их решения, а затем рассмотрим, какие индексы возможно построить при условии таких ограничений. В следующем разделе мы продемонстрируем пример одного из таких индексов.

является лишь демонстрацией полноты базы (не конечной целью), а покрытие обычно не более чем на 20% превосходит 1, мы оставили выяснение причин такого поведения данных за рамками этой записки.

[7] Далее везде говорится о товарах, т.к. объем услуг в базе в период с 2016 по 2018 год невелик. Однако все, что обсуждается, также применимо и к услугам.

Рисунок 4. Отношение объемов скользящего года по данным ККТ к 2018 году Росстата (%)



3.1. Расчет индексов цен по индивидуальным товарам

Для того чтобы рассчитать индекс цен по каждому индивидуальному товару, необходимо сначала определиться с тем, *что мы понимаем под термином «товар»* в принципе. Если следовать классической экономической теории, то товаром называется продукт определенной марки с определенным весом/литражом, вкусом/цветом/другими характеристиками, продаваемый в конкретном месте (то есть два абсолютно одинаковых пакета молока, проданных в разных магазинах, нужно считать разными товарами).

Приведенное выше определение понятия «товар» является весьма удачным с точки зрения классической парадигмы, поскольку оно позволяет сначала снимать цены одних и тех же товаров в одних и тех же магазинах, а затем агрегировать эти данные практически в любые индексы, для которых возможно посчитать или измерить веса.

Однако при переходе к данным ККТ такой подход сталкивается со следующей проблемой: в чеках в настоящий момент единственным обязательным полем, идентифицирующим продукт в конкретной кассе, является его название. Заполнение этого поля никак не регламентируется, то есть два разных продукта могут называться одинаково (может быть не указана марка, вкус, цвет, литраж и так далее) либо один и тот же товар может называться по-разному (например, в разные периоды времени).

Другой возникающий вопрос – *что считать ценой товара.* Рассчитываемый в настоящий момент ИПЦ, например, измеряет цену товара на полке (цену товара, которая написана на ценнике)[8,9]. В чеках же указана фактическая цена, по которой совершена продажа, причем с учетом различного рода скидок (по картам лояльности,

[8] С учетом промоакций, длящихся не менее 1 недели.

[9] Отметим, что в теории построения ИПЦ нет требования к построению именно индекса цен на полке, однако ввиду того, что цены сделок в обычной методологии недоступны, традиционно используются цены на полке.

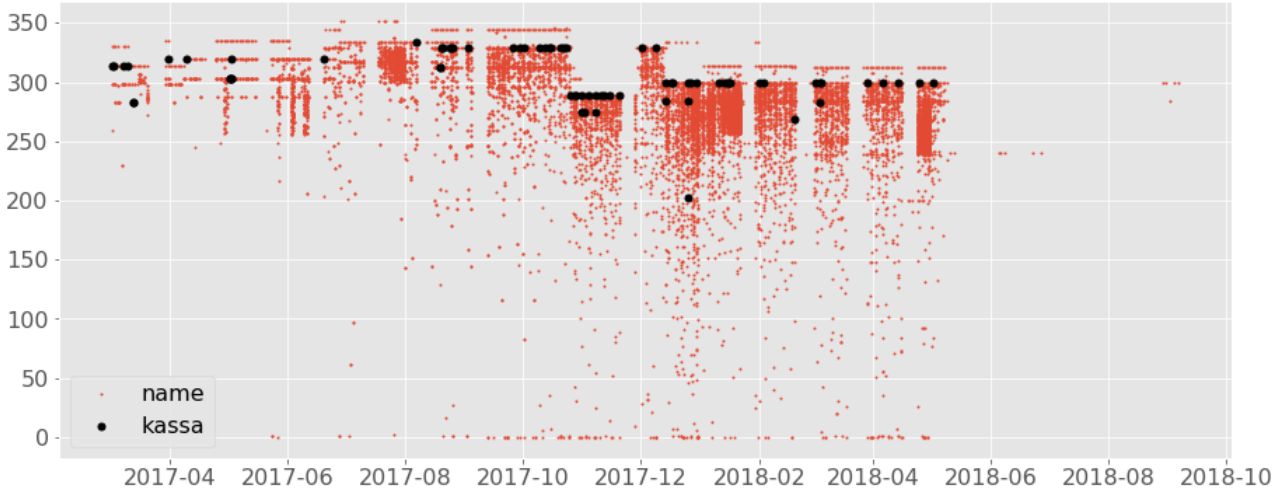
определенным группам населения и так далее) и промоакций. Таким образом, при необходимости построения индекса цен на полке, подобного рассчитываемому в настоящий момент ИПЦ, на данных ККТ возникает проблема перевода цены сделки в цену на полке.

Рисунок 5 иллюстрирует обе описанных выше проблемы. На нем показаны ценовые уровни (цены, по которым совершены сделки внутри одного дня) для двух различных определений товара одного и того же продукта[10] (которые по сути близки к определению, данному выше, но учитывают специфику данных ККТ):

- **kassa.** Товаром называется все, что имеет одинаковое (без учета регистра) название и продано в одной и той же кассе. Ввиду того, что нам априори неизвестны адреса магазинов, мы берем именно кассы, а не магазины.

- **name.** Товаром называется все, что имеет одинаковое (без учета регистра) название и продано в одном и том же регионе.

Рисунок 5. Ценовые уровни для разных определений товара, Москва (руб.)



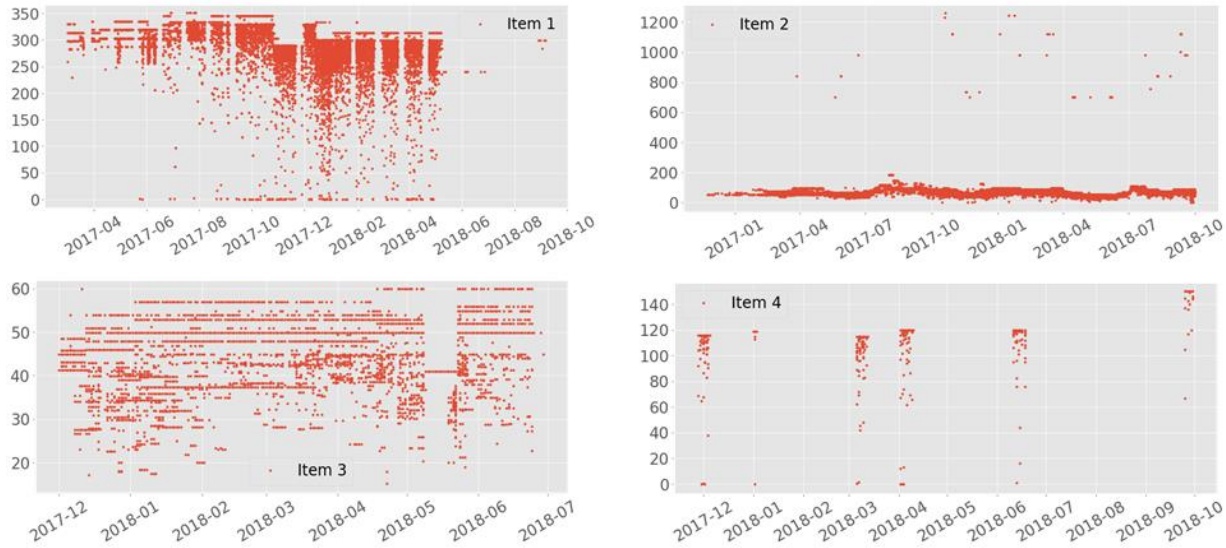
Из Рисунка 5 видно, что даже внутри группы касс с одинаковым названием продукта (определение «name», далее – группы касс) в одном городе один и тот же товар может быть продан по различным ценам. При этом, если рассматривать отдельную кассу, то для данного товара ценовые уровни также не уникальны.

Важными являются и несколько других наблюдений. *Во-первых*, максимальные цены внутри одной кассы не всегда совпадают с максимальной ценой внутри группы касс. *Во-вторых*, максимальная внутридневная цена внутри группы касс, которая может быть рассмотрена в качестве кандидата для оценки цены на полке, может значительно колебаться день ото дня. Как следствие, она является плохой аппроксимацией, особенно при расчете индекса в режиме, приближенном к реальному времени. Более того, проблемы разнятся по своей специфике от товара к

[10] Несмотря на то, что выше говорилось о потенциальных проблемах с идентификацией товаров в чеках, Рисунок 5 был построен непосредственно на примере товара, который нам удалось однозначно идентифицировать в чеках (то есть по названию мы абсолютно точно смогли определить товар).

товару (см. Рисунок 6), что делает для большинства из них решение задачи по автоматизации определения цены на полке практически невозможным. *В-третьих*, из рисунка 6 также видно, что для данных ККТ актуальна проблема, связанная с отсутствием наблюдений в некоторые дни. Это вызвано тем, что данные о цене появляются в базе только при наличии покупки. Если же товар стоит на полке, но не покупается, то, в отличие от классического подхода, где человек может подойти и переписать цену, наблюдение будет пропущено.

Рисунок 6. Ценовые уровни для случайно выбранных товаров, Москва (руб.)



Проблема отсутствия наблюдений для конкретного товара также возникает и при классическом подходе, но это происходит только если товар пропал с полок, например в случаях снятия с производства, прекращения продаж в конкретном магазине или проблем с поставками. В данных ККТ потенциальных причин отсутствия наблюдений гораздо больше. Это может быть как отсутствие товара в магазине, так и отсутствие продаж или изменения названия товара в чеке ККТ. Чтобы продемонстрировать масштаб проблемы, на Рисунке 7 была построена доля «выпадающих» товаров (доли товаров, которые не появлялись с даты на графике до сентября 2018 года) по кассам с большим объемом продаж продовольственных товаров по категории «Мясопродукты»[11,12]. Если строить индекс по названиям (которые доступны для всех чеков), то к середине рассматриваемого периода из корзины выпадает около 5% товаров, а к концу периода – около 15% (если не брать в расчет последние несколько недель, резкий рост в которые обусловлен краевыми эффектами выборки)[13]. Мы не рассчитали аналогичный показатель по кассам, но по построению доля ушедших товаров в этом случае должна быть еще больше. Иная ситуация показана на Рисунке 8, где изображена динамика доли «выпадающих»

[11] Для сокращения времени расчетов была использована только одна товарная группа, а не все наименования.

[12] Описания того, как выделяются товарные группы, представлено в следующем подразделе. Принцип формирования подвыборки касс подробнее описан в разделе 4.

[13] Резкий рост в конце рассматриваемого периода связан с тем, что многие наименования просто ни разу не продавались в последние несколько недель.

товаров для тех касс из подвыборки, где за короткий период в конце марта – начале апреля выпадает наибольшее количество товаров (большой скачок на Рисунке 7). В начале апреля 2018 года множество товаров в этой подгруппе были переименованы, что привело к резкому исчезновению более 30% наименований в чеках этой подвыборки. Все это показывает, что проблема отсутствия наблюдений и соотнесения названия товара в чеке и реального продукта также остро стоят при использовании данных ККТ.

Рисунок 7. Доля выпадающих товаров по кассам с большим объемом продаж продовольственных товаров , Мясопродукты, Москва

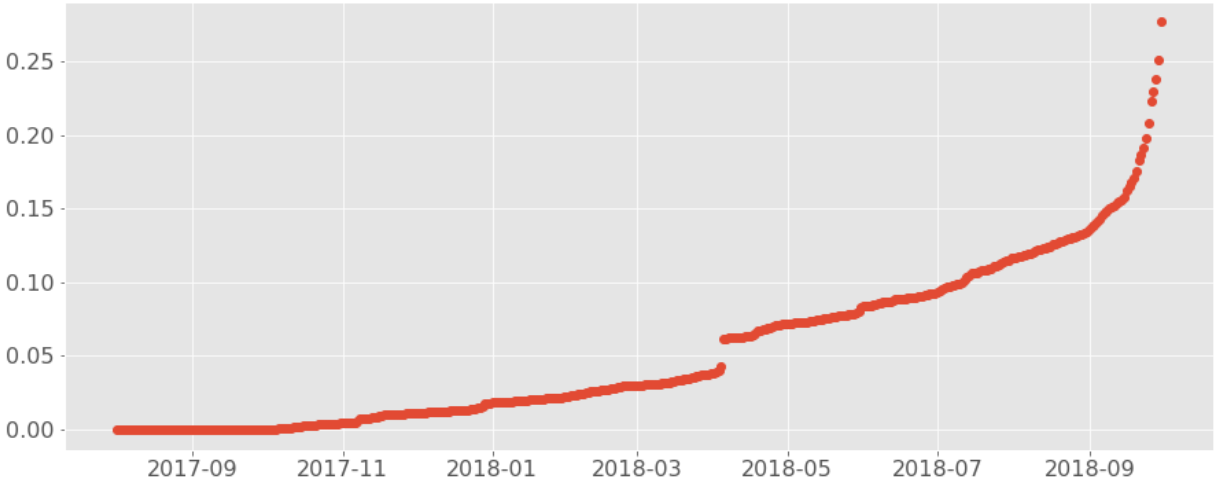
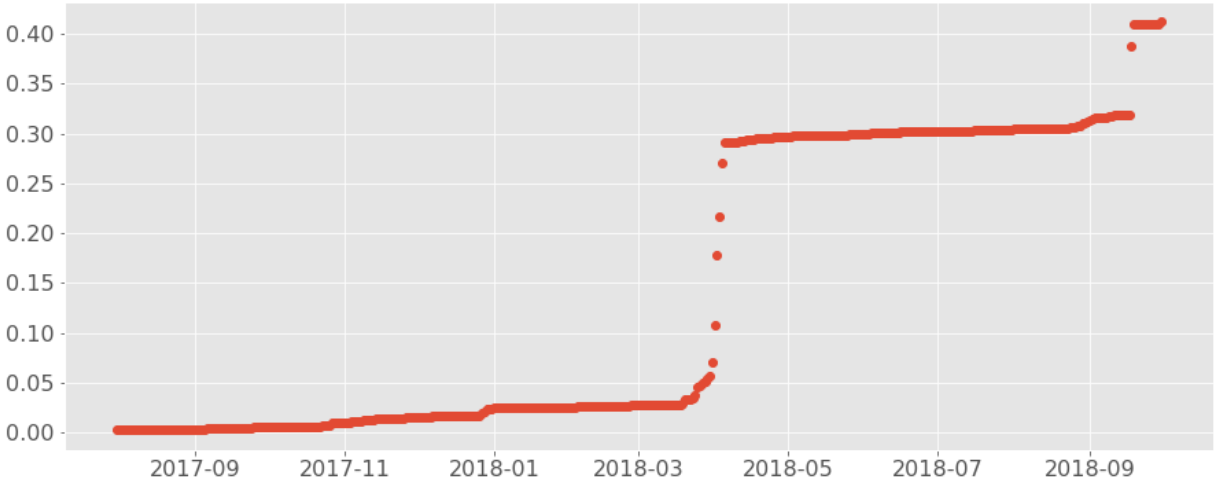


Рисунок 8. Доля выпадающих товаров для подвыборки касс, где выпадает наибольшее количество товаров, категория «Мясопродукты», Москва



Ввиду отсутствия в доступной выборке ряда ключевых полей, таких как идентификатор торговой точки, проведенный выше анализ основан на неполной информации, представленной в чеках. Однако даже при гипотетическом наличии возможности группировки касс внутри одного магазина описанные выше методологические проблемы останутся актуальными по целому ряду причин:

- Сохранится проблема, возникающая из-за невозможности точно идентифицировать товар по названию и разных названий одного и того же товара.

- Сложность с различными ценовыми уровнями и определением цены на полке возникает даже в случае наличия одной кассы. При увеличении количества касс, и, как следствие, количества ценовых уровней, данная проблема, очевидно, будет стоять лишь острее.
- На Рисунках 6–8 можно увидеть, что пропуски в данных возникают даже тогда, когда объединяются все товары с одинаковыми во всех магазинах региона названиями. Если объединять товары с одним и тем же названием только внутри одного магазина, то таких пропусков станет еще больше из-за меньшего числа наблюдений.

3.2. Объединение товаров в товарные группы и агрегация

Объединение товаров в товарные группы требует алгоритма отнесения товара к той или иной категории. Когда человек собирает информацию, он, видя товар, может распределить его по описанию товарных групп. С чеками же дело обстоит сложнее. *Во-первых*, количество товаров в чеках велико[14], чтобы распределить их вручную, а *во-вторых*, по наименованию в чеке не всегда понятно, о каком товаре фактически идет речь. Таким образом, работа с данными ККТ, в отличие от традиционного построения индексов цен, требует разработки своего подхода для объединения товаров.

После объединения индивидуальные индексы/цены нужно агрегировать внутри товарных групп, а затем и в индексы более высокого уровня. Здесь существует два ключевых отличия от классического подхода. *С одной стороны*, данные ККТ содержат подробную информацию о количестве и сумме проданных товаров, что позволяет рассчитывать веса практически любым образом (в том числе и игнорируя эту информацию, просто используя средние арифметические/геометрические на отдельных этапах, как это делает Росстат при построении ИПЦ ввиду недоступности подробной статистики). *С другой стороны*, как показано в Таблице 1, в чеках далеко не всегда присутствует информация по объему, весу или количеству штук в упаковке, что не позволяет рассчитывать средние цены за литр/килограмм/штуку.

Таблица 1. Случайно выбранные названия майонеза «провансаль» одного бренда

Название товара
Майонез КАЛЬВЕ Провансаль 200 гр.
Майонез КАЛЬВЕ Прованс
Майонез КАЛЬВЕ м\у 800гр Провансаль
КАЛЬВЕ майонез Провансаль Легкий 390г/24
КАЛЬВЕ Соус майонез Прован Легкий 20
Майонез КАЛЬВЕ соус провансаль 0,2
КАЛЬВЕ Майонез Провансаль Лёгкий 2
КАЛЬВЕ майонез Легкий Провансаль 2

[14] В базе более 600 млн уникальных названий товаров.

3.3. Пути решения описанных сложностей

Пути решения для всех описанных выше сложностей можно разделить на два больших класса: 1) на этапе *сбора данных* и 2) на этапе *построения индекса*.

Первый путь предполагает включение в чек дополнительного набора обязательных полей. Чтобы определять цену товара на полке, можно дополнить чек информацией о цене до промоакций и различных программ лояльности, чтобы однозначно определять товар – ввести единый классификатор товаров и вносить в чек в том числе информацию из этого классификатора[15]. Аналогично можно внести в чек вес, объем и количество штук в упаковке. Таким образом, можно решить все проблемы, за исключением расчета цены в случае отсутствия продаж. Однако вероятно, что при наличии множества дополнительной информации влияние этого фактора также может быть практически полностью нивелировано.

Второй путь основан на разработке автоматизированных алгоритмов и правил для решения каждой из обозначенных выше проблем. Так, например, для названий товаров могут быть построены правила отнесения их в ту или иную товарную группу или разработаны алгоритмы классификации на основе современных методов машинного обучения.

Главным достоинством второго пути является потенциальная возможность решения практически любой задачи, но остается вопрос о полноте и точности. С учетом того, что месячный рост цен традиционно варьируется в пределах нескольких десятых процентного пункта, даже небольшие ошибки алгоритмов, применяемых на разных этапах, могут накапливаться и приводить к более значимым итоговым погрешностям в расчетах. Как следствие, при выборе этого пути необходимо уделять огромное внимание анализу ошибок и погрешностей. Решение же проблем на этапе сбора информации сильно повышает точность, сводя погрешность практически к нулю. Однако этот способ видится труднореализуемым в ближайшей перспективе ввиду того, что он фактически требует перекладывания непосредственно на магазины трудозатрат, никак не связанных с предоставлением налоговой отчетности.

3.4. Выбор формулы

Поскольку на текущий момент доступны только алгоритмические решения, и для большинства задач алгоритмы решения еще не придуманы либо непонятны их точность и вклад в финальную погрешность, круг потенциальных формул для расчета индекса цен достаточно узок.

Для расчета могут быть использованы только доступные в чеках данные. Например, при выделении отдельных товаров и классификации их по группам это могут быть название товара, географическая и юридическая принадлежность

[15] Частично это уже реализуется в рамках проекта по маркировке товаров (честныйзнак.рф, национальный-каталог.рф).

продавца, цены и суммы продаж[16]. Однако формула не должна содержать цены на полке, поскольку они не присутствуют в данных чеков – вместо этого, в зависимости от конкретной задачи, связанной с построением ценового индекса, могут использоваться, например, либо средние, либо максимальные цены продажи.

В свою очередь, выбор весов сопряжен с гораздо большей свободой. Пока не представляется возможным лишь взвешивать цены за литр или килограмм, т.к. во множестве чеков такие данные отсутствуют (таким образом, единственная альтернатива – частично пожертвовать покрытием). Отдельным вопросом остается задача дозаполнения цен товаров, которые не продаются. Это может быть сделано несколькими способами, например, дозаполнением последними наблюдаемыми значениями цены либо заменой похожим товаром в случае, если произошла смена названий или товар не появляется в течение продолжительного времени (но в обоих случаях нужно оценивать масштаб искажений, которые данный подход несет).

4. ИНДЕКС ЦЕН

В этом разделе мы продемонстрируем пример индекса, который может быть построен с использованием данных ККТ. Важно отметить, что этот индекс не ставит своей целью репликацию ИПЦ или повторение методологии построения ИПЦ, а лишь является иллюстрацией и может быть позиционирован как самостоятельный индикатор ценовой динамики.

Для разбиения товаров на товарные группы был построен алгоритм, который по названию распределяет товары в одну из заранее заданных групп. На текущий момент это укрупненные товарные группы Росстата по продовольственным товарам, а также отдельная группа для непродовольственных товаров (см. Приложение 2). Для этого около 100 000 названий товаров из данных ККТ были вручную разделены на группы, а затем на них[17] была обучена модель, в которой деревья решений «голосуют» за одну из групп, и побеждает та, которая набрала наибольшее количество голосов. Такой метод позволяет находить баланс между точностью процедуры классификации (которая влияет на погрешность обработки данных) и долей классифицированных товаров (которая влияет за счет ошибки покрытия), варьируя минимальный процент деревьев (уровень уверенности[18]), необходимый для отнесения товара к той или иной группе. В Таблице 2 показано, как количество классифицированных товаров меняется с увеличением уровня уверенности.

[16] Цены и суммы продаж могут быть полезны, например, при дифференцировании между разными по объему, но одинаковыми по названию товарами.

[17] В качестве дополнительных данных использовались данные интернет-магазина «Ашан» и сайта GoodsMatrix, однако основной вклад в обучение внесли размеченные данные.

[18] Не то же самое, что и уровень доверия в статистике.

Таблица 2. Зависимость количества классифицированных товарных позиций от уровня доверия

Уровень уверенности	50%	65%	80%	95%
Точность	93%	96%	98%	99%
Доля классифицированных (обучение)	88%	79%	71%	51%
Доля классифицированных (выборка для индекса)	88%	80%	71%	51%

При построении индекса мы используем определение товара, которое соответствует названиям (определение «name» из раздела «Сложности при построении индекса цен» выше), то есть все продажи с одним названием в каждом регионе считаются одним товаром. Ввиду доступных нам вычислительных мощностей берутся только данные по подвыборке касс. На первом этапе выбираются кассы с большой долей продовольственных товаров[19]. Затем из них выбираются только те кассы, в которых присутствуют наиболее продаваемые (по названиям) товары. Получившаяся выборка покрывает около 15% розницы.

На Рисунке 3 видно, что наполнение базы сильно увеличилось с июля 2017 года. Поэтому для того, чтобы построить индекс хотя бы за один полный год и приблизиться по крайней мере частично к построению индекса в режиме реального времени, мы фиксируем набор товаров, которые продавались с июля по сентябрь 2017 года, строим индивидуальные индексы *с октября 2017 по сентябрь 2018 года* и нормируем их к единице на 1 октября 2017 года. В качестве цены используется средневзвешенная (внутри дня) по сумме продаж цена товара[20]. В период до июля 2017 года происходило активное подключение пользователей к системе ККТ. Чтобы учесть это и избежать сезонности в весах, мы жертвуем полным повторением построения индекса в реальном времени и в качестве весов используем средние объемы продаж с октября 2017 по сентябрь 2018 года при условии, что товар изначально попал в выборку на прошлых этапах.

Индекс по группе товаров получается усреднением индивидуальных индексов с весами, описанными выше[21]:

$$I_t^g = \sum_{i=1}^{N} \frac{Q^i}{Q^g} I_t^i,$$

где I_t^g – индекс по товарной группе g в момент времени t; I_t^i – индивидуальный индекс товара i в момент времени t; Q^i – сумма продаж товара i; $Q^g = \sum_{i=1}^{N} Q^i$ – сумма продаж по товарной группе g.

Пропуски в данных на текущей стадии расчетов заполнялись последним из известных значений. Это, безусловно, не оптимальный метод, его доработка является предметом дальнейших исследований[22]. Так, товары, которые переименовались или пропали из продажи, будут учитываться по последней цене и не будут вносить вклад в рост цен. В классическом подходе такие товары были бы заменены товарами-заменителями, которые в среднем растут в цене (из-за положительной инфляции).

[19] Также в этих кассах продаются и непродовольственные товары.

[20] Если в какой-то из дней объемы продаж были аномально маленькими или большими, то цена в такой день считалась нерепрезентативной и удалялась.

[21] В формуле используется среднее арифметическое. Для иллюстративного примера и одного года это нормально, но в дальнейшем, когда будет доступно больше данных, планируется переход на средние геометрические, т.к. они обладают чуть лучшими математическими свойствами.

[22] Нами ведется работа над алгоритмами замены товара в случае его отсутствия в продаже длительное время.

Таким образом, используемый метод заполнения пропусков потенциально имеет тенденцию к занижению совокупного роста цен по сравнению с классическим подходом. Товары, которые не продаются, но стоят на полках, также вносят занижающий вклад в рост цен в случае повышения цен на них, но это не настолько критично по сравнению с предыдущим примером, т.к. такое занижение должно носить временный характер – изменение цены будет зафиксировано после первой же продажи по новой цене. Получить верхнюю границу доли занижения можно из Рисунка 7, который является типичным для большинства товарных групп. Если исключить последние полтора месяца, т.к. они обусловлены окончанием выборки, а не исчезновением товаров[23], и аппроксимировать оставшуюся часть трендом, то получается, что к концу рассматриваемого года доля месячного роста цен занижается не более чем на 20%, а скорее всего, гораздо меньше (то есть если бы месячный рост цен составлял один процент, то в октябре 2017 года он был бы оценен практически полностью, а в сентябре 2018 года – в диапазоне от 0,8 до 1).

Гранулярность данных ККТ позволяет рассчитывать субиндексы с достаточной степенью дезагрегации[24]. Для демонстрации этого на Рисунках 9 и 10 показаны индексы цен[25], начиная с отдельных товарных групп и заканчивая региональными индексами на продовольственные товары. На Рисунках 11–13 изображены агрегированные по стране индексы на продовольственные, непродовольственные и все товары.

Несмотря на то, что построенный индекс – лишь иллюстрация, он позволяет сделать *несколько общих выводов*.

Во-первых, процедура классификации товаров и подвыборка, на которой строится индекс, важны. Как видно из Рисунка 13, рост цен при 50%-ном уровне уверенности практически совпадает с ростом цен, построенным с использованием всех товаров, но значительно отличается от роста цен на уровне уверенности в 95%. Это происходит из-за того, что на уровне 50% классифицируются практически все товары, а на уровне 95% – только половина. Более того, в неклассифицированные зачастую попадают сезонные товары. Яблочный сок, например, не всегда возможно отличить от яблок или яблочного шампуня в названии чеков (во всех названиях может присутствовать «яблок»). Это приводит к тому, что на уровне товарных групп, содержащих эти товары, индексы могут быть достаточно зашумленными и не показывать реальной картины. В нашем случае это не так важно, так как построенный агрегированный индекс не зависит от разбиения на группы, а только от включенных в него товаров, поэтому более точным будет индекс на уровне 50% (чего нельзя однозначно сказать о субиндексах). Но в случае, если используется другая формула,

[23] Товары с определенными названиями продаются раз в несколько недель/месяцев. С момента последней продажи они будут считаться пропавшими, что приводит к увеличению доли пропавших наименований в конце рассматриваемого периода (последние полтора месяца на Рисунке 7). Однако если бы у нас в распоряжении был более длительный период для наблюдений, то в последующие моменты времени товары с такими наименованиями были бы снова проданы и не считались бы пропавшими из продажи.

[24] Можно спуститься даже на уровень отдельных касс.

[25] Для наглядности на графиках исключена внутринедельная сезонность.

как, например, в ИПЦ, ввиду математических особенностей разбиение на группы будет играть роль даже для агрегированного индекса[26].

Во-вторых, вопреки сложившимся стереотипам мы не нашли никаких признаков систематического завышения или занижения инфляции Росстатом. Также несмотря на разницу в уровнях индексов, динамика на разных уровнях уверенности зачастую оказывается схожей. При этом сравнение с ИПЦ Росстата показывает, что во многих случаях ИПЦ по укрупненным категориям не совпадают с нашим индексом. Это может быть связано как с разницей в методологии, так и с разницей в покрытии[27].

Рисунок 9. Индексы на уровне товарных групп в разрезе регионов в зависимости от уровня уверенности

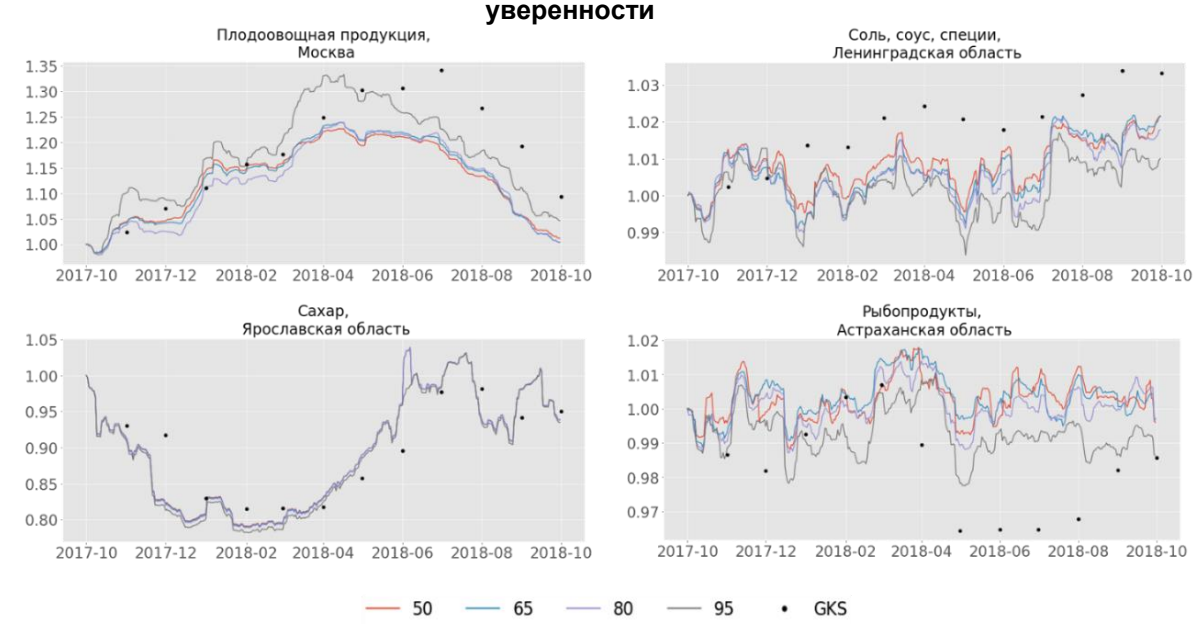
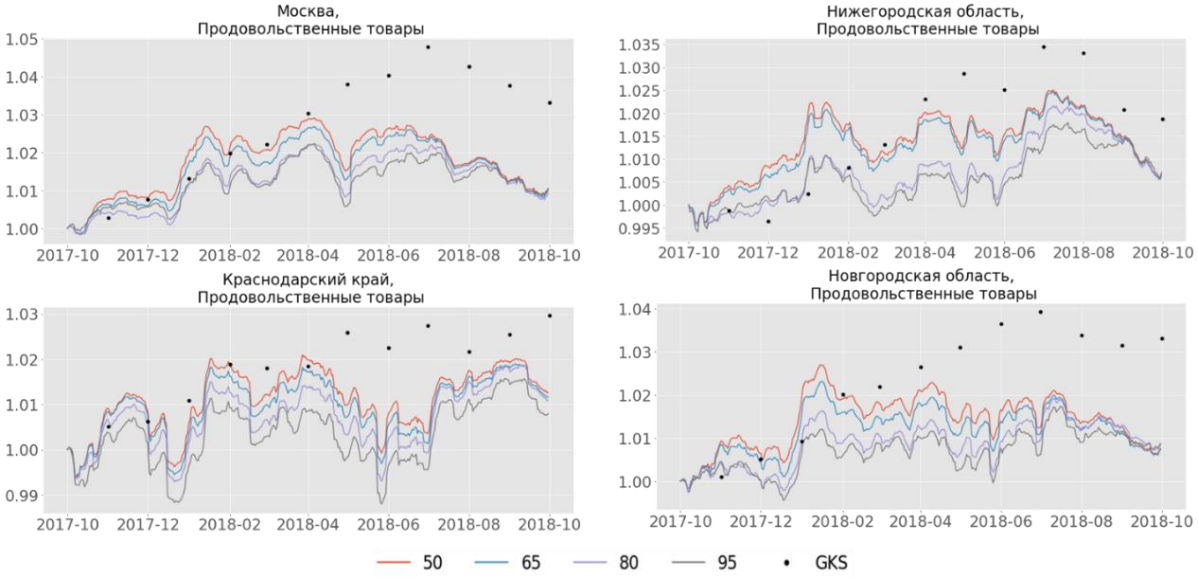


Рисунок 10. Индексы продовольственных товаров в разрезе регионов в зависимости от уровня уверенности



[26] Разбиение будет играть роль, если финальный индекс нелинейно зависит от отдельных субиндексов и их весов. В ИПЦ, например, на некоторых этапах используется среднее геометрическое, что приводит к подобного рода нелинейностям.

[27] Выяснение подробных причин расхождения требует чуть больше исторических данных, поэтому мы оставляем эту задачу для будущих исследований.

Рисунок 11. Индекс продовольственных товаров в зависимости от уровня уверенности

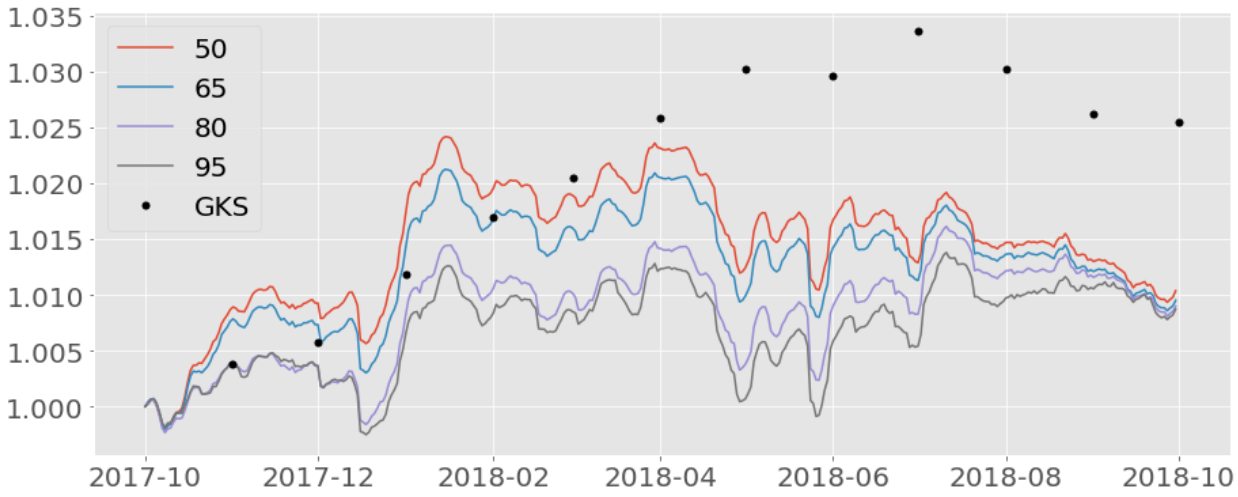


Рисунок 12. Индекс непродовольственных товаров в зависимости от уровня уверенности

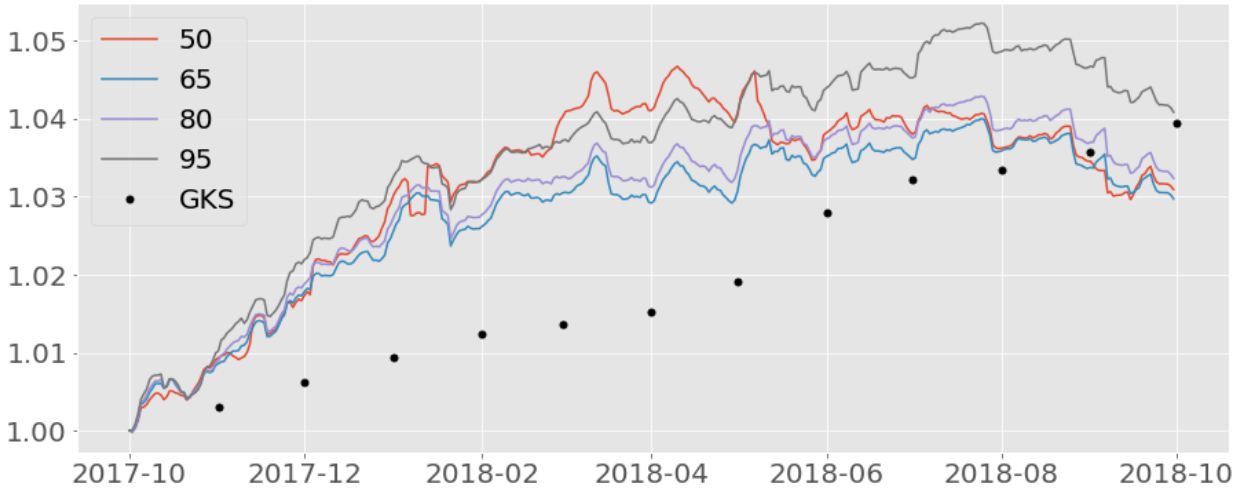
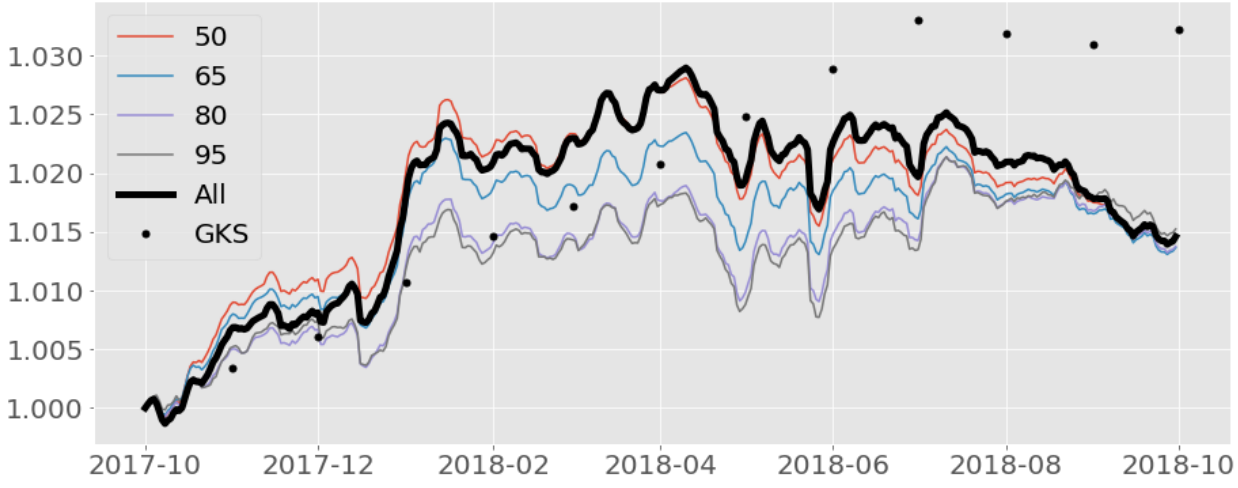


Рисунок 13. Сводный индекс цен в зависимости от уровня уверенности



5. ЗАКЛЮЧЕНИЕ И БУДУЩИЕ НАПРАВЛЕНИЯ РАБОТЫ

В настоящей записке обсуждаются возможности использования данных ККТ для усовершенствования текущей статистики по ценам. Проходя последовательно все шаги построения индекса цен, мы показываем, какие ограничения возникают при попытке построить индексы цен на основе данных ККТ, а также обсуждаем возможные пути их решения.

Подводя итог, можно сказать, что на текущем этапе особенности данных ККТ таковы, что в ближайшее время они вряд ли смогут быть использованы в качестве основного источника для построения общепринятых в официальной статистике показателей ценовой динамики, в частности ИПЦ. При этом данные ККТ могут служить полезным источником для построения альтернативных ценовых индексов, а также гранулярного анализа ценовых процессов.

Кроме того, в записке был показан пример расчета иллюстративного индекса потребительских цен, по результатам которого не было найдено никаких признаков систематического занижения/завышения инфляции, измеряемой Росстатом, относительно динамики цен из данных ККТ на рассматриваемом периоде.

Дальнейшие направления работы с данными ККТ включают в себя:

• Переход к построению индексов на всех данных, а не только по подвыборке касс;

• Построение индекса цен по методологии, максимально приближенной к методологии Росстата, и его анализ.

ПРИЛОЖЕНИЕ 1. РЕКВИЗИТНЫЙ СОСТАВ ДАННЫХ

Информация о ККТ

- регистрационный номер ККТ;
- заводской номер фискального накопителя (ФН);
- регион ККТ;
- населенный пункт ККТ.

Информация о чеке

- номер фискального документа (ФД);
- дата, время расчета;
- сумма расчета, указанного в чеке;
- сумма по чеку наличными;
- признак расчета (наличный или безналичный расчет);
- применяемая система налогообложения.

Информация о предмете расчета

- наименование предмета расчета;
- количество предмета расчета;
- цена за единицу предмета расчета с учетом скидок и наценок;
- стоимость предмета расчета с учетом скидок и наценок.

ПРИЛОЖЕНИЕ 2. КАТЕГОРИИ ДЛЯ КЛАССИФИКАЦИИ

Категории для классификации

- Мясопродукты;
- Рыбопродукты;
- Масло и жиры;
- Молоко и молочная продукция;
- Сыр;
- Консервы овощные;
- Консервы фруктово-ягодные;
- Яйца;
- Сахар;
- Кондитерские изделия;
- Варенье, джем, повидло, мед;
- Чай, кофе, какао;
- Соль, соус, специи, концентраты;
- Мука;
- Хлеб и хлебобулочные изделия;
- Макаронные и крупяные изделия;
- Плодоовощная продукция, включая картофель;
- Алкогольные напитки;
- Напитки безалкогольные;
- Мороженое;
- Общественное питание;
- Непродовольственные товары.