



НАЦИОНАЛЬНЫЙ БАНК КАЗАХСТАНА

Система Galymzhan: online-оценка потребительской инфляции в Казахстане

Департамент денежно-кредитной политики
Экономическое исследование №2021-2

Тулеуов О.,
Ержан И.,
Сейдахметов А.

Экономические исследования и аналитические записки Национального Банка Республики Казахстан (далее – НБРК) предназначены для распространения результатов исследований НБРК, а также других научно-исследовательских работ сотрудников НБРК. Экономические исследования распространяются для стимулирования дискуссий. Мнения, высказанные в документе, выражают личную позицию автора и могут не совпадать с официальной позицией НБРК.

Система Galymzhan: online-оценка потребительской инфляции в Казахстане.
NBRK – WP – 2021 – 2

© Национальный Банк Республики Казахстан

Любое воспроизводство представленных материалов допускается только с разрешения авторов

Система Galymzhan: online-оценка потребительской инфляции в Казахстане

Тулесуов Олжас¹
Ержан Ислам²
Сейдахметов Ансар³

Аннотация

Центральным банкам, придерживающимся политики инфляционного таргетирования, крайне важно отслеживать и проводить анализ текущих инфляционных тенденций. Наличие оперативной информации об изменении цен дает возможность действенно реагировать на различные шоки и при необходимости своевременно корректировать денежно-кредитную политику.

В настоящее время сдерживающим фактором оперативного анализа изменения цен на товары потребительской корзины является ограниченная частотность публикаций официальной ценовой статистики. В этой связи, особое значение приобретает вопрос построения косвенных высокочастотных прокси-индикаторов инфляции.

В данной работе описаны методология и результаты построения высокочастотного прокси-показателя инфляции Казахстана в Национальном Банке посредством использования технологии веб-скрепинга, подразумевающей автоматическое получение данных путем их извлечения с веб-страниц, реализованным с помощью программного алгоритма.

Ключевые слова: инфляция, веб-скрепинг, Galymzhan, онлайн магазин, цены, big data, ИПЦ, парсинг, потребительские товары.

JEL-классификация: E31, E37, E39.

¹ Тулеуов Олжас – советник Председателя Национального Банка Республики Казахстан. E-mail: Olzhas.Tuleuov@nationalbank.kz

² Ержан Ислам – главный специалист-аналитик управления макроэкономических исследований и прогнозирования Департамента денежно-кредитной политики Национального Банка Республики Казахстан. E-mail: Islam.Yerzhan@nationalbank.kz

³ Сейдахметов Ансар – ведущий специалист-аналитик управления макроэкономических исследований и прогнозирования Департамента денежно-кредитной политики Национального Банка Республики Казахстан. E-mail: Ansar.Seidakhmetov@nationalbank.kz

Содержание

1. Введение.....	5
2. Обзор литературы и международного опыта.....	6
3. Используемые данные и методология	10
3.1 Galymzhan 1.0	10
3.2 Galymzhan 2.0	11
4. Обсуждение результатов	13
5. Выводы и рекомендации для дальнейших исследований.....	16
Список литературы.....	18

1. Введение

Инфляция – показатель средневзвешенного роста цен на товары потребительской корзины, является основополагающим индикатором при принятии решений для центральных банков, осуществляющих политику инфляционного таргетирования. Устойчиво низкие темпы инфляции способствуют расширению горизонта планирования и росту инвестиций, что приводит к расширению экономической активности и улучшению благосостояния населения.

Сохранению устойчиво низких темпов роста цен на товары потребительской корзины способствует проведение последовательной денежно-кредитной политики.

В периоды различных кризисов в целях сдерживания инфляционных процессов монетарному регулятору необходимо действенно реагировать на происходящие и ожидаемые макропроцессы, а также при необходимости своевременно корректировать денежно-кредитную политику. Наличие оперативной информации об изменении потребительских цен дает возможность принимать более качественные решения.

На текущий момент сдерживающим фактором оперативного анализа изменения цен на товары потребительской корзины в Казахстане является ограниченная частотность публикаций официальной статистики. В этой связи, особое значение приобретает вопрос построения косвенных высокочастотных прокси-индикаторов инфляции.

В последние годы среди статистических органов и центральных банков многих стран особую популярность приобретают системы сбора ценовых изменений, основанные на «сканировании» цен и веб-скрепинге интернет ресурсов, которые позволяют получать высокочастотную информацию в большом объеме.

В широком смысле «сканирование» цен подразумевает использование данных ретейлеров о транзакциях, в которых содержится информация о цене и обороте товаров, идентифицируемым по бар-кодам. Использование данной технологии подразумевает тесное сотрудничество с магазинами, которые предоставляют данную информацию (G. Chessa, 2016).

Технология веб-скрепинга различных сайтов позволяет провести сбор ценовых данных в структурированном виде для дальнейшей обработки и получения ценной высокочастотной информации об изменении потребительских цен. Технология получила свое развитие ввиду того, что интернет-покупки становятся все более популярными, и многие вендоры переходят к онлайн-продажам, открывая свои официальные сайты с ценами на всю товарную продукцию.

Данный вид сбора ценовых изменений набирает большую популярность как в государственных учреждениях (статистические органы, центральные банки), так и в частных организациях различных стран, ввиду экономии времени, человеческих ресурсов и минимизации расходов. Получение данных об изменении уровня цен посредством веб-скрепинга стало особенно актуальным в

2020 году, когда на фоне пандемии коронавируса COVID-19 многие торговые точки были вынуждены приостановить свою деятельность в условиях карантинных ограничений, что усложняло сбор ценовых изменений классическим методом.

Национальный Банк Казахстана приступил к изучению вопроса сбора данных с интернет ресурсов в 2018 году, когда была разработана система сбора и анализа динамики потребительских цен, названная Galymzhan.

С технической точки зрения Galymzhan представляет собой алгоритм, написанный на языке программирования R. Интеграция данного кода в систему Windows позволила добиться автономности данной программы. Так, ежедневно, включая праздничные и выходные дни в одно и тоже время система запускает алгоритм сбора и обработки данных. Результатом работы системы является построение ежедневного прокси-индикатора инфляции.

Данная работа состоит из нескольких частей. Первый раздел представлен обзором литературы, в котором рассмотрены аналогичные работы других авторов. Во втором разделе описывается методология сбора ценовых изменений и построения конечного индекса. Далее представлен раздел обсуждения результатов, в котором авторы работы сравнивают полученный индекс с официальными статистическими данными. Выводы данного исследования являются заключительным разделом.

2. Обзор литературы и международного опыта

Увеличение количества различных интернет ресурсов вкупе с развитием технологий, позволяющих извлекать нужную информацию с открытых источников в структурированном виде, дало сильный импульс к развитию многих сфер жизни. Наличие ранее недоступной информации позволило автоматизировать и оптимизировать бизнес-процессы.

Процесс получения статистической информации, в частности прокси-показателя инфляции, с применением технологии веб-скрепинга в последнее время набирает большую популярность. Данный метод в сравнении с классическим обследованием цен, проводимым официальными статистическими органами, позволяет фиксировать ценовые изменения с более высокой частотностью и высвободить человеческий ресурс, что предоставляет статистическим органам возможность одновременно повышать качество и частотность публикуемых данных.

В настоящее время статистические органы таких стран как Италия, Австралия, Бельгия, Германия, Нидерланды, Венгрия, Великобритания, Канада начали осваивать и внедрять системы веб-скрепинга в своей операционной деятельности (Boshoff, 2020). Так, Национальный институт статистики Италии проводит активную работу по внедрению системы автоматического фиксирования цен (Polidoro, 2015). В частности, группа статистиков и IT-специалистов провели тестирование возможности использования методов извлечения данных с веб-страниц в обследовании потребительских цен,

сосредоточив внимание на двух группах товаров: бытовая техника и авиабилеты. Результаты тестирования подтвердили наличие большого потенциала использования автоматического сбора цен в интернете с точки зрения повышения эффективности и качества статистических данных.

Прежде всего, использование методов парсинга в качестве инструмента получения данных для измерения инфляции непосредственно связано с одной из трех составляющих больших данных (Big Data), а именно со скоростью (Velocity)⁴. Следует отметить, что сотрудники института статистики и ранее использовали цены, полученные с веб-ресурсов, однако до этого цены фиксировались работниками вручную (в режиме «copy and paste»).

Другим статистическим органом, использующим технологии извлечения данных об изменении цен с веб-ресурсов, является Бюро статистики Австралии. С середины 2017 года Бюро статистики постепенно вносит в базу данных, используемую при расчете индекса потребительских цен (ИПЦ), показатели, полученные в результате веб-скрепинга⁵. Перед этим цены, полученные из сети Интернет, сравнивали с ценами, собранными классическим методом за последние 12 месяцев, чтобы убедиться, что они совпадают и подходят для измерения динамики цен. После этого собранные вручную цены на определенные продукты были заменены данными, полученными из интернета, и ручной сбор для этих продуктов прекратился. Собранные таким образом данные в основном используются для построения индекса цен на алкоголь, небольшого количества товаров одежды и автомобильных запчастей, что составляет около 5% ИПЦ. Такой подход имеет ряд преимуществ. Средняя цена на основе данных, собранных дважды в неделю, более репрезентативна, чем цена, собираемая один раз в месяц. Также это позволяет включать больше продуктов в корзину ИПЦ и снижать затраты на сбор. В настоящее время еженедельно ведется сбор цен на 500 000 товаров-представителей, тогда как при традиционных способах сбора собиралось всего около 1000 позиций⁶.

Федеральное статистическое управление Германии в 2017 году начало проводить работу по онлайн-сбору цен. На экспериментальном этапе сбор проводился по 242 товарам потребительской корзины, не включая цены на услуги. Всего в потребительской корзине Германии насчитывается около 600 товаров и услуг. Технология веб-скрепинга позволила собирать 3050 цен каждый час. Столь высокая частотность объясняется тем, что данная работа на начальном этапе помимо прикладного характера имела и исследовательскую направленность (Blaudow and Burg, 2018). В результате высокочастотного сбора цен авторам удалось выявить подробную картину динамики цен в интернет-

⁴ Принято считать, что большие данные можно объяснить с помощью трех V: Velocity (скорость), Variety (разнообразие) и Volume (объем). В исследовательском отчете 2001 года аналитик META Group (ныне Gartner) Дуг Лэйни определил большие данные как трехмерные, то есть увеличивающийся объем (объем данных), скорость (скорость поступления и передачи данных) и разнообразие (диапазон типов данных и источников). Позже в 2012 году Gartner обновила определение больших данных как большого объема, высокой скорости и большого разнообразия.

⁵ Официальный сайт Австралийского бюро статистики, Режим доступа: <https://www.abs.gov.au/articles/web-scraping-australian-cpi>.

⁶ Australian Bureau of Statistics (ABS), Maximizing the use of Web Scraped Data in the Australian CPI.

магазинах в течение суток. Так, большая часть онлайн-магазинов вносит ценовые корректировки в начале суток (00:00).

Система автоматического сбора цен статистического управления Германии написана на языке программирования Java. Парсинг веб-страниц выполнялся с помощью инструмента Selenium, который управляет веб-браузером через воспроизведение заранее указанных в коде действий в браузере. Для исключения блокирования интернет ресурсами в проекте используются динамические IP-адреса.

В настоящий момент при расчете инфляции используются онлайн данные об изменении цен на 10 000 товаров-представителей по таким группам как продукты питания, безалкогольные напитки, алкогольные напитки и табачные изделия (Boshoff, 2020).

Бельгийский офис статистики на протяжении последних лет также ведет сбор данных с интернет магазинов. Система написана на языке программирования R и использует такие пакеты как *rvest*⁷ и *Rselenium*⁸. Полученные данные уже используются при расчете инфляции на международные поездки и видеоигры.

Вдобавок около 70 парсеров ежедневно или несколько раз в неделю проводят сбор цен на одежду, обувь, авиабилеты, гостиничные номера, подержанные автомобили, лекарства, книги и бытовую технику и т.д. Статистическое управление Бельгии в ближайшем будущем планирует включить данные по вышеперечисленным товарам, полученные путем веб-скрепинга, в расчет официально публикуемого индекса потребительских цен (Loon, Roels, 2018).

Статистическое управление Нидерландов на текущий момент проводит сбор цен с интернет-ресурсов трех крупных ретейлеров. Ежедневно фиксируя изменения цен на 100 000 товаров-представителей. При этом, фиксирование ценовых изменений проводится только для предметов одежды. Это объясняется двумя основными причинами. Во-первых, с 2010 года статистическое управление Нидерландов фиксирует изменение цен на часть продовольственных непродовольственных товаров через технологию сканирования цен, и на текущий момент большая часть сбора данных в офлайн магазинах приходится на предметы одежды (Grient, Naan, 2010). Во-вторых, одежда – это товар, который не требует частой корректировки выборки товаров, что упрощает автоматизацию процесса. На текущий момент в тестовом режиме 16 роботов собирают данные в 19 розничных магазинах. После валидации эти данные будут добавлены в основную выборку, что позволит увеличить ее до 500 000 ежедневных наблюдений (Griffioen, Bosch, 2016).

Помимо официальных статистических органов использование данных, полученных с использованием технологии веб-скрепинга для построения прокси-ИПЦ, получило широкое распространение и среди негосударственных учреждений. При наличии ресурсов и соответствующих навыков построение

⁷ Библиотека, позволяющая проводить сбор информации с веб-страниц.

⁸ Пакет, предоставляющий возможность использовать «Selenium 2.0 WebDriver» в R для автоматизации веб-браузера.

косвенного ИПЦ становится посильной задачей и для индивидуальных исследователей. Polidoro в своей работе пишет, что в будущем на фоне широкого доступа инструментов и информации, статистические институты могут потерять свою конкурентоспособность, и они, вероятно, могут утратить свою монополию, которая сейчас поддерживается их эксклюзивным правом сбора и распространения статистической информации на официальном уровне (Polidoro, 2015).

Примерами распространения статистики цен неофициальными органами является проекты «The Billion Prices Project» и «PriceStats». Проект «The Billion Prices Project» – это академическая инициатива американских экономистов Альберто Кавалло и Роберто Ригобона, в рамках которой с 2008 года ведется ежедневный сбор цен у сотен интернет-магазинов по всему миру с дальнейшим предоставлением оценки инфляции в реальном времени. Толчком к созданию данной системы явилось манипулирование статистикой инфляции в Аргентине с 2007 по 2015 года. В этот период многим стало очевидно, что официальный уровень инфляции, рассчитываемый Национальным статистическим управлением Аргентины, не отражает реальное изменение цен. По официальным данным института статистики среднегодовой уровень инфляции с 2007 по 2015 годы составил 8%, в то время как среднее значение согласно онлайн-данным составляло более 20%, что соответствовало оценкам в отдельных регионах и инфляционным ожиданиям домохозяйств. Возможность проводить онлайн-сбор цен вне пределов страны оказалось особенно полезными в 2011 году, когда правительство Аргентины начало налагать штрафы и принуждать местных экспертов прекратить сбор данных. В конце 2015 года после смены правительства манипуляции с официальной статистикой прекратились.

В настоящее время «The Billion Prices Project» ежедневно фиксирует цены на около 5 миллионов товаров, продаваемых в 300 онлайн-магазинах, более чем в 50 странах (Cavallo, 2016).

В 2011 году на основе «The Billion Prices Project» был создан коммерческий проект «PriceStats», который собирает данные и, используя их, строит высокочастотные индексы для центральных банков и финансового сектора. В «PriceStats» было значительно увеличено количество и качество используемых данных. На текущий момент компания фиксирует цены на около 15 миллионов товаров от более чем 1200 розничных продавцов для дальнейшего построения ежедневного индекса инфляции в 23 странах и индекса цен производителей в 10 странах. Всего сбор цен осуществляется для 60 стран.

Анализ международного опыта использования технологии веб-скрепинга для построения косвенного индекса потребительских цен продемонстрировал жизнеспособность и возрастающий интерес к данному методу сбора цен как среди государственных учреждений, так и частного сектора. Увеличение количества интернет-магазинов и постоянное совершенствование доступных инструментов извлечения информации, в частности цен с онлайн-магазинов, в ближайшем будущем могут стать основным источником статистической информации.

3. Используемые данные и методология

Процесс развития и становления в Национальном Банке Казахстана системы Galymzhan как инструмента, используемого для анализа текущей ситуации на рынке потребительских цен посредством построения прокси-индикатора инфляции, можно разделить на два периода. С начала 2018 года по октябрь 2020 года использовалась первая версия системы – Galymzhan 1.0. В октябре 2020 года система была существенно усовершенствована. Так, во второй версии увеличилось количество отслеживаемых товаров, расширилась географическая представленность и повысилась автономность системы.

Открытие большого количества новых интернет-площадок не только в крупных городах, но и в региональных центрах Казахстана, катализатором чего выступила пандемия коронавируса COVID-19, а также улучшение и накопление навыков сбора и обработки данных аналитиков, работающих над системой Galymzhan, позволили провести существенное развитие в данном направлении деятельности.

Помимо того, что веб-скрепинг является довольно сложным с технической точки зрения процессом, который требует определенного перечня навыков в программировании, выбор подхода к расчету индекса также является нетривиальной проблемой. Традиционный сбор ценовых изменений, осуществляемый Бюро Национальной Статистики Агентства стратегического планирования и реформ Казахстана (далее – БНС), основан на целевой выборке, которая предполагает наличие заранее выбранного по сегментам потребления списка репрезентативных товаров и услуг. В данный список попадают товары и услуги, потребляемые домохозяйствами, веса которых определяются по доле их затрат на приобретение данных товаров и услуг. Далее обследование цен осуществляется инспекторами локально в различных торговых точках. Методология, используемая БНС⁹ и системой Galymzhan, основана на восходящем подходе, при котором сначала централизованно определяются товары, а далее происходит сам процесс сбора цен.

Необходимо отметить что целью системы Galymzhan является построение прокси-индикатора инфляции, а не сбор всей имеющейся информации по динамике цен. Использование аналогичного с БНС подхода позволяет сохранить интерпретируемость полученных индексов и использовать их в качестве прокси-оценки потребительской инфляции в Казахстане.

3.1 Galymzhan 1.0

Galymzhan 1.0., запущенный в начале 2018 года, представляет собой специальный алгоритм, написанный на языке программирования R. Интеграция данного кода в систему Windows позволила добиться того что, ежедневно,

⁹Методология построения индекса потребительских цен.
<https://stat.gov.kz/api/getFile/?docId=ESTAT109961>

включая праздничные и выходные дни, система запускает алгоритм сбора и обработки данных с последующим получением ежедневного прокси-показателя инфляции.

Система проводит ежедневный сбор цен на 395 видов продовольственных и непродовольственных товаров, по одному наименованию на каждую категорию товара потребительской корзины в 13 магазинах по заранее указанным HTML-ссылкам. Фиксирование ценовых изменений ведется в двух крупных городах: Нур-Султане и Алматы. Узкая географическая представленность первой версии системы была связана со ограниченным количеством интернет-площадок в крупных городах Казахстана и их отсутствием в регионах. В этой связи, при построении сводного индекса делается допущение, что изменение цены на определенный товар в одном из двух городов будет восприниматься системой как общереспубликанское изменение цены на данную категорию товара, то является репрезентативным для всей генеральной совокупности товара.

Полученные в результате «полевого» сбора данные по ценам на товары преобразуются в индексы и группируются в сводный аппроксимированный ИПЦ в соответствии с весами БНС каждого товара. Итоговым результатом работы Galymzhan 1.0. является накопление ежедневного прокси-показателя ИПЦ.

3.2 Galymzhan 2.0

В настоящее время Galymzhan 2.0 функционирует в тестовом режиме и сбор ценовых изменений производится по группе социально-значимых продовольственных товаров (далее – СЗПТ), состоящей из 19 продовольственных категорий продуктов¹⁰. Galymzhan 2.0. обрабатывает 16 официальных интернет-ресурсов различных торговых марок, расположенных в шести городах Казахстана¹¹, и собирает ежедневную информацию по ценам на около 5000 продовольственных товаров-представителей. В среднем на одну категорию товара потребительской корзины приходится около 263 товаров-представителей, тогда как в первой версии системы на одну категорию товара потребительской корзины приходился один товар-представитель. Использование большего количества товаров-представителей позволило повысить репрезентативность выборки и одновременно снизить волатильность индекса.

Полное соблюдение методологии БНС предполагает сбор цен на заранее отобранные товары в течение года. Однако в отличие от физического обследования сбор цен в интернете имеет ряд особенностей. К одной из таких особенностей следует отнести частую смену ссылок на товары. В этой связи,

¹⁰ Согласно постановлению Правительства Республики Казахстан от 27 марта 2017 года № 137: мука, хлеб, рожки, крупа гречневая, рис, картофель, морковь, лук репчатый, капуста, сахар, масло подсолнечное, говядина, мясо кур, молоко, кефир, масло сливочное, яйца, соль, творог. <http://adilet.zan.kz/rus/docs/P1700000137#2>

¹¹ Нур-Султан, Алматы, Шымкент, Караганда, Павлодар, Усть-Каменогорск

было принято решение увеличить частоту фиксирования отслеживаемых товаров и ссылок на них. Так, каждые две недели система производит полное «сканирование» магазинов со сбором всех продуктов и их метаданных, которые имеются в наличии. На данном этапе записываются все ссылки на товары, их принадлежность к категориям, название магазина и его месторасположение. В общей сложности собирается информация по 140 000 товаров. Затем система проводит классификацию продукции, используя категоризацию БНС, и отбирает необходимые товары путем использования простейших алгоритмов обработки естественного языка. Характеристики отобранных товаров записываются в отдельную базу.

Далее в течение двух последующих недель выборка продуктов остается неизменной и, исходя из данной выборки, рассчитывается прокси-показатель инфляции. По истечении двух недель процесс повторяется. То есть используемая нами процедура автоматизирует этап определения необходимых товаров-представителей и их характеристик, уменьшая временные затраты. Такой подход также позволяет снизить нагрузку на вычислительную мощность используемого компьютера и проводить сбор ежедневно.

Сам процесс веб-скрепинга начинается с загрузки всей html-страницы в среду программирования¹². Однако для этого необходимо определить тип сайта, с которого будет собираться информация. Поэтому на этом этапе определяется тип сайта. Существует два основных типа сайта – это динамический и статический. В первом, динамическом типе, данные сайта доступны только по запросу и динамически перемещаются в зависимости от действий пользователя. Запросом в таком сайте может быть первоначальная загрузка сайта, «прокручивание» веб-страницы, нажатие курсором на определенные элементы веб-страницы и т.д. Так как для сканирования динамических вебсайтов необходимо симулировать действия пользователя, простые библиотеки, возвращающие по запросу html-файл, не подходят. Для работы с подобными страницами необходимы сторонние подсистемы вроде Selenium (RSelenium), которые используют браузер и способны «воспроизвести» действия пользователя. Во втором же случае, если тип сайта статический, элементы находятся в заранее определенных местах вне зависимости от действий пользователя. Очевидно, что страницы такого типа легче поддаются веб-скрепингу, так как не требуют симуляции действий пользователя. С такими сайтами можно использовать встроенные библиотеки R для работы с веб-страницами, как *rvest*, преимуществом которого является скорость работы и простота в настройке и использовании. Каждый веб-сайт является уникальным. Поэтому практически под каждый отдельный сайт в системе Galymzhan 2.0. запрограммированы различные сценарии для извлечения данных.

Для извлечения данных с вебсайта используются уникальные идентификаторы групп элементов страницы, называемые XPath-путями. С помощью таких «путей» можно выделить все элементы, относящиеся к ценам, наименованиям и т.п.

¹² В данном случае это среда программирования и статистических вычислений R

Процесс обработки полученных данных начинается с определения дневного изменения цен на все товары-представители. Далее изменение цен на товары-представители, относящиеся к одному товару потребительской корзины в каждом регионе, группируются путем расчета общего среднегеометрического показателя. Затем полученные ценовые изменения агрегируются в инфляцию СЗПТ по регионам и республике, а также общереспубликанское изменение цен на все 19 товаров.

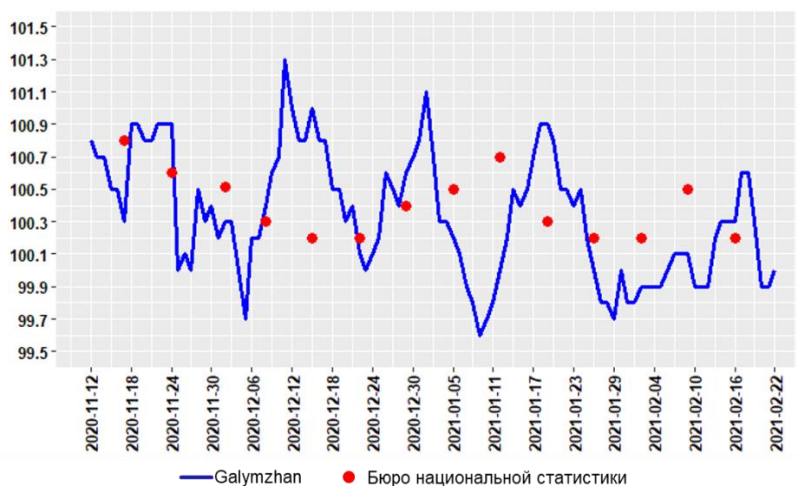
Помимо ежедневной прокси-инфляции система рассчитывает два дополнительных показателя – еженедельную инфляцию и накопленную месячную инфляцию. Для сравнения с еженедельно публикуемой инфляцией на СЗПТ ежедневный прокси-показатель инфляции трансформируется в показатель недельной частотности путем накопления предыдущих дневных изменений. Накопленный месячный показатель используется для дополнительного анализа и сравнения с официальными данными по итогам месяца. Анализ полученных индексов позволяет корректировать текущие прогнозы инфляции, а также глубже изучать и анализировать инфляционные процессы в Казахстане.

Полученные результаты визуализируются в виде дашбордов в Tableau – системе интерактивной аналитики, которая позволяет анализировать динамику каждой из 19 категорий товаров во всех регионах максимально быстро и удобно.

4. Обсуждение результатов

Конечным результатом деятельности системы Galymzhan является построение агрегированного прокси-индикатора инфляции, в данном случае инфляции СЗПТ. Для иллюстрации текущих результатов системы Galymzhan на рисунке 1 представлены официальная недельная динамика цен на 19 социально-значимых товаров, публикуемая БНС, и еженедельная прокси-инфляция, рассчитанная на основе данных, полученных технологией веб-скрепинга.

Рисунок 1. Динамика индекса цен на СЗПТ и прокси-инфляции, в % к соотв. дню пред. недели

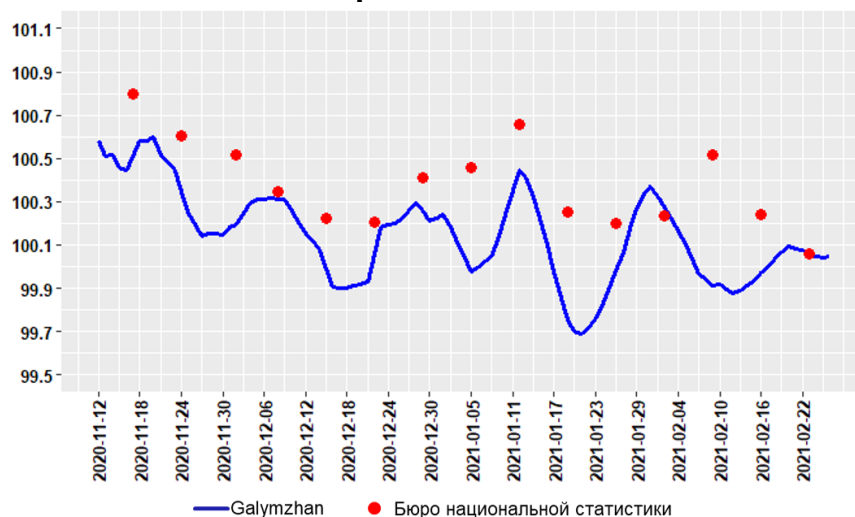


Источник: БНС, расчеты авторов

В целом как видно из рисунка, несмотря на высокий размах расчетных значений прокси-инфляции, она в большинстве случаев конвергируется с фактической недельной инфляцией социально-значимых товаров.

В ходе анализа было определено, что данные с онлайн-магазинов в сравнении с ценами, фиксируемыми классическим методом обследования, имеют тенденцию к более быстрому изменению, что возможно связано с высокой оборачиваемостью товаров и отсутствием явных издержек на обновление ценника. Подобное явление в макроэкономике называется «эффектом меню». При этом, темпы роста онлайн-цен в большинстве случаев ниже темпов роста офлайн-цен. Вероятно, данный факт связан с тем, что наличие собственного онлайн-магазина в основном является прерогативой крупных продовольственных сетей, где стоимость конечных продуктов может быть ниже в сравнении с единичными торговыми точками за счет эффекта масштаба. Наличие временного лага между онлайн-ценами и ценами, фиксируемыми классическим методом обследования, предоставляет основания использовать показатель, рассчитанный системой Galymzhan 2.0 в качестве опережающего индикатора официальных данных об изменении цен на социально-значимые продукты. На рисунке 2 представлена еженедельная прокси-инфляция, взятая с 10 дневным лагом, и индекс цен на социально-значимые продовольственные товары. Дополнительно для снижения визуального шума онлайн-индекс был сглажен семидневной скользящей средней.

Рисунок 2. Динамика индекса цен на СЗПТ и прокси-инфляции*, в % к соотв. дню пред. недели



*Данные системы Galymzhan сглажены 7 дневным МА и взяты с 10 дневным лагом без учета цен на овощи

Источник: БНС, расчеты авторов

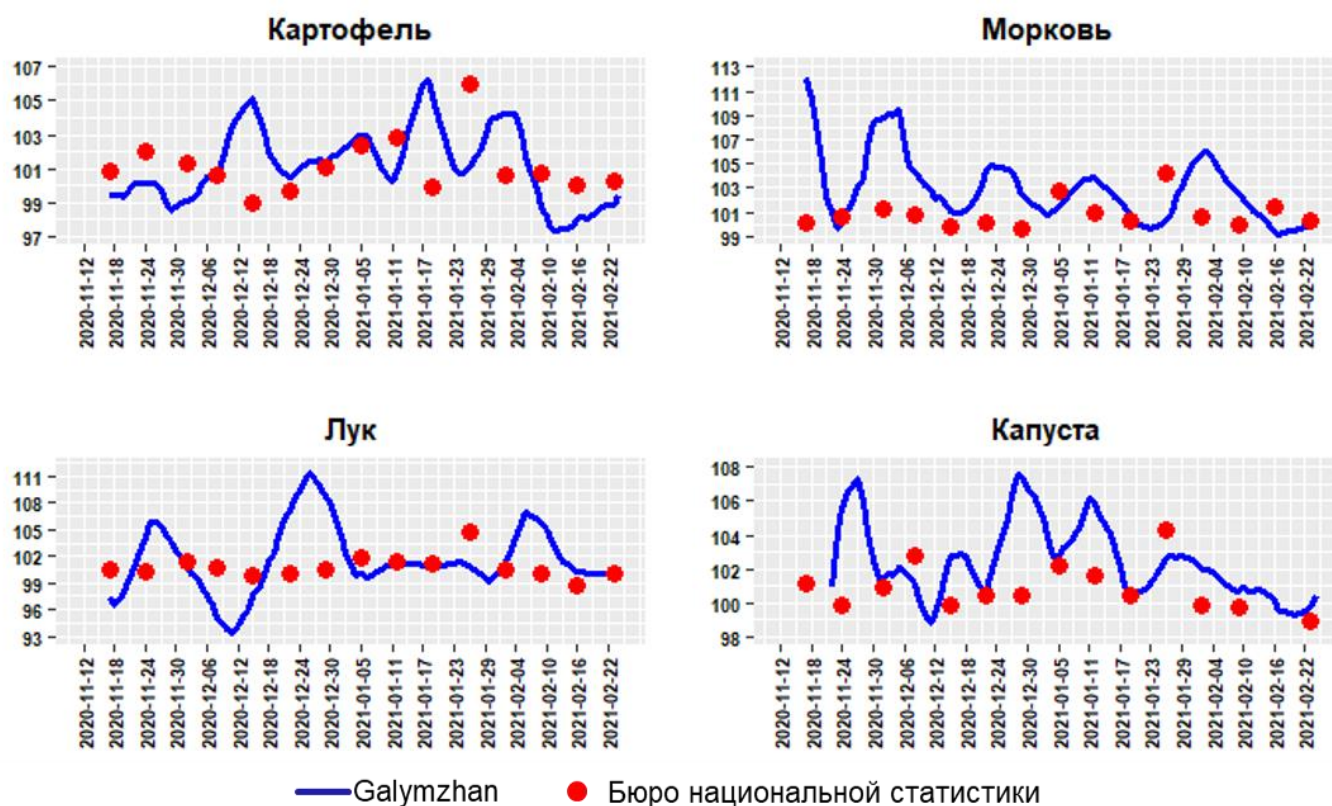
Вместе с тем, по ряду товаров, в частности по овощной продукции, наблюдается некоторое расхождение с официальной статистикой. Данный факт связан с малой представленностью уникальных товаров-представителей в онлайн-магазинах (рисунок 3).

В дальнейшем по мере расширения системы (добавления новых магазинов и увеличения выборки) ожидается, что дифференциал между данными БНС и Galymzhan 2.0 будет сокращаться.

В качестве примеров региональной разбивки были выбраны города Нур-Султан и Алматы, как города с наибольшим количеством онлайн-магазинов (4 и 8 онлайн-площадок, соответственно). При этом, в городе Нур-Султан наблюдается сильная зависимость между данными системы Galymzhan и официальными показателями БНС (рисунок 4).

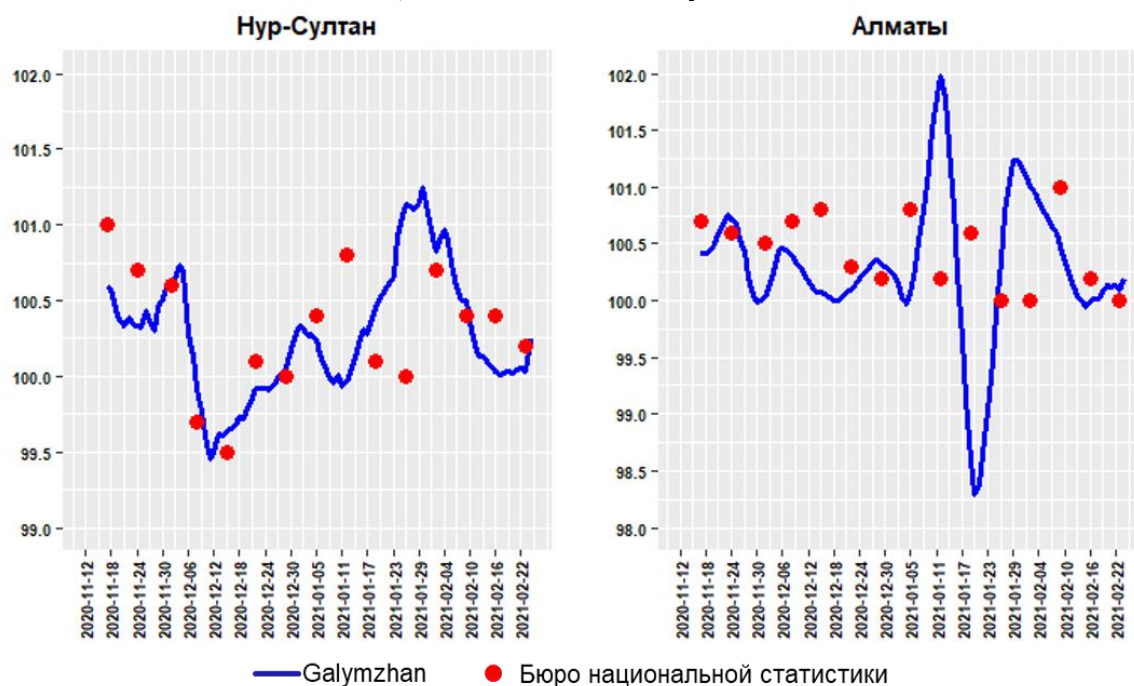
В отличие от г. Нур-Султан, в г. Алматы наблюдается повышенная волатильность. Так, в предновогодний период (26-27 декабря 2020 года) было зафиксировано резкое повышение цен на большинство продуктов питания. Возможным объяснением данного повышения цен является возросший спрос вследствие новогодних праздников. При этом, рост цен отмечается в крупных торговых сетях, в то время как цены на локальных рынках практически не изменились. Таким образом, преимуществом данной системы также является возможность проведения глубокого анализа ценовой динамики в периоды аномального поведения рынков. Так, в настоящее время Galymzhan позволяет отслеживать динамику цен определенных товаров-представителей на полках различных торговых точек.

Рисунок 3. Динамика индекса цен на СЗПТ и прокси-инфляции на овощи*, в % к соотв. дню пред. недели



*Данные системы Galymzhan сглажены 7-дневным МА и взяты с 10 дневным лагом
Источник: БНС, расчеты авторов

Рисунок 4. Динамика индекса цен на СЗПТ и прокси-инфляции* в Нур-Султане и Алматы, в % к соотв. дню пред. недели



*Данные системы Galymzhan сглажены 7-дневным МА и взяты с 10 дневным лагом

Источник: БНС, расчеты авторов

В дальнейшем по мере улучшения системы, а именно расширения географической представленности, добавления новых товаров-представителей и увеличения общего количества категорий товаров потребительской корзины ожидается, что конвергенция между индексом, рассчитываемым системой Galymzhan, и официальными данными будет сокращаться, а расширенный индекс будет максимально приближен по товарной и региональной представленности к индексу потребительских цен и будет использоваться при построении краткосрочных прогнозов, для целей анализа и принятия решений по денежно-кредитной политике.

5. Выводы и рекомендации для дальнейших исследований

В рамках текущего исследования был изучен мировой опыт использования технологии веб-скрепинга при построении инфляции как в официальных статистических органах, так и в частных организациях. Кроме того, приведено описание методологии, используемой системой Galymzhan 2.0 для сбора и дальнейшего построения прокси-показателя инфляции на социально-значимые продовольственные товары в Казахстане.

Полученные результаты свидетельствуют о наличии корреляции между официальными данными и построенным индексом. При этом, было отмечено, что цены в онлайн-магазинах меняются быстрее, чем в офлайн торговых точках, и темпы роста цен в реальных магазинах несколько выше.

Стоит отметить, что система Galymzhan 2.0 в настоящее время функционирует в тестовом режиме. В целях дальнейшего развития системы авторами планируется проведение работ по расширению географического охвата и добавления большего количества магазинов, категорий товаров потребительской корзины и товаров-представителей.

Потенциал дальнейшего развития процесса визуализации связан с включением в дашборд возможности отслеживания ценовых корректировок не только в регионах и магазинах, но и конкретных товаров-представителей на онлайн-полках определенных магазинов.

Дальнейшее расширение системы позволит максимально приблизиться по географическому и товарному охвату к индексу потребительских цен, что позволит реплицировать его в полной мере, а также использовать результаты системы Galymzhan 2.0 в качестве инструмента при построении краткосрочных прогнозов и анализа монетарной политики.

Список литературы

1. A Guide to Data Integration for Official Statistics, *UNECE High Level Group for the Modernisation of Official Statistics Data Integration Project*, January 2017;
2. Alan Bentley, Frances Krsinich, Towards a big data CPI for New Zealand, *Ottawa Group*, 2017;
3. Alberto Cavallo, Roberto Rigobon, The Billion Prices Project: Using Online Prices for Measurement and Research, *Journal of Economic Perspectives*, Spring 2016;
4. Alberto Cavallo, Scraped Data and Sticky Prices, *MIT Sloan School of Management*, April 2012;
5. Antonio G. Chessa, A new methodology for processing scanner data in the Dutch CPI, *EURONA — Eurostat Review on National Accounts and Macroeconomic Indicators*, January 2016;
6. Antonio G. Chessa, Processing scanner data in the Dutch CPI: A new methodology and first experiences, *Statistics Netherlands*;
7. Antonio G. Chessa, Robert Griffioen, Comparing Price Indices of Clothing and Footwear for Scanner Data and Web Scraped Data, *Economics and Statistics*, April 2019;
8. Australian Bureau of Statistics (ABS), Maximizing the use of Web Scraped Data in the Australian CPI;
9. Ben Powell, Guy Nason, Duncan Elliott, Matthew Mayhew, Jennifer Davies, Joe Winton, Tracking and modelling prices using web-scraped price microdata: towards automated daily consumer price index forecasting, *Journal of the Royal Statistical Society*, 2017;
10. Christian Blaudow, Florian Burg, Dynamic pricing as a challenge for consumer price statistics, *EURONA — Eurostat Review on National Accounts and Macroeconomic Indicators*, March 2018;
11. Federal Statistical Office of Germany, *Your Benefit. Our Mission.*, 2018;
12. Federico Polidoro, Web scraping techniques to collect data on consumer electronics and airfares for Italian HICP compilation, *Statistical Journal of the IAOS* 31 (2015) 165–176;
13. Heymerik van der Grient, Jan de Haan, The use of supermarket scanner data in the Dutch CPI, *Statistics Netherlands*, July 2010;
14. Janine Boshoff, Xuxin Mao, Garry Young, Outlier detection methodologies for alternative data sources, *ESCoE Technical Report 07*, July 2020;
15. Karola Brunner, Automated price collection via the internet, *Wirtschaft und Statistik*, April 2014;
16. Ken Van Loon, Dorien Roels, Integrating big data in the Belgian CPI, *Meeting of the Group of Experts on Consumer Price Indices Geneva, Switzerland*, May 2018;

17. Ken Van Loon, Dorien Roels, Web scraping and online data collection and processing for the consumer price index, *Belgian statistical office*, February 2018;
18. Lincoln T. da Silva, Ingrid L. de Oliveira, Tiago M. Dantas, Vladimir G. Miranda, *Studies of new data sources and techniques to improve CPI compilation in Brazil*, Brazilian Institute of Geography and Statistics;
19. Matthew Mayhew, Comparison of index number methodology used on UK web scraped price data, *Office for National Statistics*;
20. Robert Breton, Tanya Flower, Matthew Mayhew, Elizabeth Metcalfe, Natasha Milliken, Christopher Payne, Thomas Smith, Joe Winton, Ainslie Woods, Research indices using web scraped data, *Office for National Statistics*, May 2016;
21. Robert Griffioen, Olav ten Bosch, On the use of internet data for the Dutch CPI, *Meeting of the Group of Experts on Consumer Price Indices Geneva, Switzerland*, May 2016;
22. Официальный сайт Австралийского бюро статистики, Режим доступа: <https://www.abs.gov.au/articles/web-scraping-australian-cpi>.