ЩО ЗУМОВЛЮЄ РОЗБІЖНОСТІ МІЖ ОФІЦІЙНИМИ ТА ОНЛАЙН – ІНДЕКСАМИ ЦІН?

ОЛЕКСАНДР ФАРИНА[a], ОЛЕКСАНДР ТАЛАВЕРА[b], ТЕТЯНА ЮХИМЕНКО[c] *

[a]Національний банк України, Національний університет "Києво-Могилянська академія".
Email: oleksandr.faryna@bank.gov.ua
Автор, відповідальний за листування;
[b]Університет Суонсі, Велика Британія;
Email: o.talavera@swansea.ac.uk
[c]Національний банк України;
Email: tetiana.yukhymenko@bank.gov.ua

Анотація У статті досліджується відповідність онлайн — індексів споживчих цін офіційній статистиці. Спершу ми розраховуємо компоненти індексу споживчих цін (ІСЦ) на основі онлайн-даних, які згодом агрегуємо до загального онлайн-ІСЦ. Цей підхід застосовується до нашого унікального масиву даних, що містить близько трьох мільйонів спостережень за роздрібними цінами на споживчі товари в Інтернеті у п'яти найбільших містах України. Дані охоплюють період із січня 2016-го до грудня 2017 року включно і покривають близько 46% споживчого кошика інфляції в Україні. Результати аналізу свідчать, що онлайн-інфляція загалом узгоджується з офіційними показникам, однак розбіжності можуть виникати на рівні окремих компонент ІСЦ. Незважаючи на те, що відмінності частково зумовлені недостатнім рівнем покриття масиву даних, онлайн-ціни можуть дійсно відображати нову інформацію, не охоплену офіційною статистикою.

Класифікація JEL: C55, E31, E37

Ключові слова: онлайн-ціни, веб-скрепінг, індекс споживчих цін, мікроціни, великі дані

1. ВСТУП

Однією з основних цілей центральних банків є дотримання низького рівня інфляції. Відповідно інфляційне таргетування в останні десятиріччя стало основним підходом до здійснення економічної політики (напр., Hammond, 2011; Jahan, 2017; Roger, 2010). Однак для досягнення інфляційної цілі центральним банкам потрібен чіткий і спостережуваний показник інфляції, який може слугувати номінальним якорем для суспільства. Вимірювання інфляції не завжди є простим завданням і зазвичай виходить за рамки сфери компетенції центральних банків. Натомість поширеною практикою є використання наявних індикаторів, що офіційно публікуються державними органами статистики, як-от індекс споживчих цін (ІСЦ).

Центральні банки найчастіше обирають ІСЦ, оскільки він вимірює вартість життя в економіці та є загальнодоступним для суспільства й інституцій, відповідальних за економічну політику. Однак незважаючи на простоту і суспільне визнання, ІСЦ може бути не найкращим показником, адже включає обмежену кількість товарів та послуг в економіці й може не охоплювати загальні інфляційні тенденції, які спостерігаються суспільством. Це може вплинути на ефективність рішень центрального банку та поставити під сумнів успіх монетарної політики в цілому. Таким чином, для здійснення монетарної політики необхідно бути озброєним усіма можливими інструментами й використовувати всі доступні джерела інформації. Це дасть змогу поліпшити здатність розпізнавати й усвідомлювати всі фактори, які загрожують цінові стабільності.

Швидка інтеграція електронної торгівлі до роздрібного сектору уможливила відстежування цін на різноманітні товари та послуги в Інтернеті. Веб-скрепінг (від англ. web scraping – збір даних з онлайн-джерел через використання спеціально розробленого програмного забезпечення) став корисним інструментом збору даних щодо онлайн-цін з Інтернету з метою доповнення офіційної статистики. Багато національних статистичних організацій та інших державних інституцій уже розпочали про-

* Цей проект було реалізовано в той час, коли Олександр Талавера брав участь у Програмі запрошених дослідників Національного банку України за підтримки Проекту Канада-МВФ "Розбудова інституційної спроможності НБУ".
Висловлюємо подяку рецензентам за надання змістовних коментарів та пропозицій.
Погляди та думки, викладені в цій статті, відображають позицію авторів і можуть не збігатися з офіційною позицією їхніх афілійованих установ.

https://doi.org/10.26531/vnbu2018.243.021

екти щодо впровадження веб-скрепінгу для поліпшення процесу збору даних. Серед них — Бюро статистики праці США (Horrigan, 2013), Служба статистики Великої Британії (Breton et al., 2015), Служба статистики Нідерландів (Griffioen, de Haan, Willenborg, 2014), Служба статистики Нової Зеландії (Krsinich, 2015) та Служба статистики Норвегії (Nygaard, 2015). Порівняно з альтернативними методами збору даних веб-скрепінг має низку переваг. Крім низьких витрат на збір даних, онлайн-дані доступні в режимі реального часу та з високою частотою, що може допомогти відповідальним за економічну політику інституціям проводити постійний моніторинг інфляційних процесів на мікрорівні.

Дедалі частіше онлайн-дані використовуються в дослідницьких цілях. Проект Мільярд цін (Billion Price Project[1]), заснований Альберто Кавалло (Alberto Cavallo) та Роберто Рігобоном (Roberto Rigobon) у Массачусетському технологічному інституті в 2008 році, має на меті збирання цін із сотень роздрібних онлайн-магазинів у всьому світі. Cavallo and Rigobon (2016) у своєму дослідженні аргументують, що онлайн-ціни можуть успішно використовуватися як альтернативне джерело інформації для побудови індексів споживчих цін. Водночас у деяких дослідженнях використовуються онлайн-дані для перевірки точності офіційної статистики та відсутності маніпуляцій. Зокрема, Cavallo (2013) на основі онлайн-цін вивчає, яким чином онлайн-індекси збігаються з офіційною статистикою в п'яти країнах Латинської Америки. Автор доходить висновку, що в той час як індекси онлайн-цін для Бразилії, Чилі, Колумбії і Венесуели точно апроксимують рівень та основну динаміку офіційних показників інфляції, онлайн-інфляція в Аргентині була майже втричі вищою від даних офіційної статистики. Coupe and Petrusha (2014), своєю чергою, доходять висновку, що офіційна та онлайн-інфляція в Україні можуть значно відрізнятися в короткостроковій перспективі, однак відхилення можуть бути як позитивними, так і негативними.

Під час порівняння офіційних та онлайн-індексів цін важливо розуміти причину виникнення можливих розбіжностей. З одного боку, онлайн-ціни можуть насправді відображати нову інформацію про довгострокову динаміку інфляції, яка не фіксується офіційною статистикою. Водночас з огляду на те, що онлайн-ринки, як правило, є гнучкішими[2], ціни в Інтернеті можуть швидше пристосовуватися до нових економічних умов, а тому можуть призвести до короткострокових відхилень, тоді як довгострокова інфляція в Інтернеті повинна відповідати офіційним оцінкам. З іншого боку, розбіжності можуть виникати через технічні проблеми, принципово різні підходи до збору даних, а також методи побудови онлайн-індексів. На відміну від офіційних, онлайн-дані, зібрані за допомогою веб-скрепінгу, зазвичай включають велику кількість товарних одиниць, тоді як охоплення роздрібних точок продажу та регіонів є обмеженим. Крім того, високочастотні онлайн-дані можуть характеризуватися великою кількістю відсутніх спостережень через технічні помилки під час веб-скрепінгу або через відсутність окремих товарів у продажу протягом певного часу. Як наслідок товарна структура онлайн-індексів цін може змінюватися в часі, що зазвичай не відповідає стандартним підходам, які використовуються статистичними організаціями. Таким чином, перш ніж робити будь-які висновки про те, чи відображають онлайн-ціни нову інформацію щодо динаміки інфляції, важливо дослідити фактори, котрі зумовлюють такі розбіжності.

У цій статті ми розробляємо онлайн — індекс споживчих цін для України, використовуючи великий масив даних онлайн-цін і порівнюємо його з офіційною статистикою, наданою Державною службою статистики України. Ми розраховуємо компоненти індексу споживчих цін на основі онлайн-даних, які згодом агрегуємо до загального онлайн-ІСЦ. Наш унікальний масив даних містить близько трьох мільйонів спостережень за роздрібними цінами на споживчі товари в Інтернеті у п'ятьох найбільших містах України. Дані охоплюють період із січня 2016-го до грудня 2017 року включно і покривають близько 46% споживчого кошика інфляції в Україні. Результати аналізу свідчать, що онлайн-інфляція загалом узгоджується з офіційними оцінками, але розбіжності можуть виникати на рівні окремих компонент ІСЦ. Ми також досліджуємо властивості нашого масиву даних, які можуть пояснити такі розбіжності. Для цього використовуємо альтернативні методи фільтрації та агрегації, які поліпшують або зменшують здатність побудованих онлайн-індексів відображати динаміку офіційної статистики. Отримані результати свідчать, що онлайн-індекси цін можуть відхилятися від офіційних значень через технічні проблеми під час збору даних та недостатнє покриття даних. Однак наш аналіз вказує на те, що онлайн-ціни можуть випереджати дані офіційної статистики й містити нову інформацію, яку не було охоплено в офіційному індексі споживчих цін.

Стаття побудована таким чином. У другому розділі описано масив онлайн-даних, що використовується для нашого аналізу. У третьому розділі подано онлайн-індекси цін для окремих компонент ІСЦ та на агрегованому рівні, а також досліджується їхня відповідність офіційній статистиці. Четвертий розділ містить висновки.

2. ОНЛАЙН-ЦІНИ НА СПОЖИВЧІ ТОВАРИ В УКРАЇНІ

У нашому аналізі використано онлайн-ціни на споживчі товари в Україні, отримані за допомогою веб-скрепінгу в Національному банку України (НБУ). У 2015 році НБУ розпочав проект веб-скрепінгу з метою поліпшення збору даних щодо споживчих ціни та доповнення офіційної статистики стосовно ІСЦ.

Індекс споживчих цін, наданий Державною службою статистики України, є основним показником для відстеження тенденцій інфляційного розвитку, що використовуються Національним банком України для проведення монетарної політики. Споживчий кошик для розрахунку ІСЦ в Україні складається з 328 компонент, 40% яких становлять продукти харчування, напої та алкоголь. У таблиці 1 наведено описову статистику щодо загального індексу споживчих цін та основних агрегованих категорій індексів споживчих цін.

Масив онлайн-даних НБУ охоплює декілька провідних роздрібних інтернет-магазинів, які, крім онлайн-платформ, мають широку мережу офлайн-супермаркетів у п'яти найбільших містах (Київ, Харків, Дніпро, Одеса та

[1] Див. для прикладу: http://www.thebillionpricesproject.com
[2] Див. для прикладу Gorodnichenko & Talavera (2017).

Львів). Ці супермаркети і їхні онлайн-платформи пропонують широкий асортимент продуктів харчування, напоїв, алкогольних та тютюнових виробів. Масив даних охоплює до 46% споживчого кошика України та включає більш як 130 компонент індексу споживчих цін. Із початку проекту масив даних НБУ налічує 75 000 найменувань товарів і майже 3 мільйони тижневих спостережень[3] (із січня 2016 року до грудня 2017 року). Більшість онлайн-цін у масиві даних походить з онлайн-магазинів Києва, який можна вважати найбільшим споживачем у сфері електронної комерції. Харків, Дніпро та Одеса мають приблизно однакові частки, водночас Львів наразі майже не представлений у масиві даних. У таблиці 2 надано описову статистику масиву онлайн-даних.

3. ПОБУДОВА ОНЛАЙН-ІНДЕКСІВ

Масив онлайн-даних НБУ надає розширену інформацію щодо цін на товари на мікрорівні в різних регіонах України. Для того, щоб з'ясувати, чи узгоджуються онлайн-ціни з офіційною статистикою, розрахуємо онлайн-індекси та порівняємо їх з офіційними даними.

3.1. Онлайн – індекси компонент ІСЦ

Наслідуючи поширену практику[4], ми спершу розраховуємо онлайн-індекси як прості середні значення щотижневих змін онлайн-цін у межах вузько визначеної групи, а саме на рівні компонент ІСЦ:

$$\Delta p_{i,t} = \sum_{j=1}^{K} \left[\frac{\left(P_{ij,t} - P_{ij,t-1}\right)}{P_{ij,t-1}} \right] \div K,$$

де Δp, $i=1,2,3...N$, означає середню щотижневу зміну цін у відсотках у межах компоненти i; $P_{ij,t}$ $j=1,2,3...K$, – ціна певного товару j в межах компоненти i.

Після цього щотижневі часові ряди в межах окремих компонент ІСЦ трансформуються в ряди зі щомісячною частотою, що полегшує порівняння з офіційною статистикою. Оскільки масив даних оновлюється щотижня, а кількість тижнів у місяці може відрізнятися, ми спершу трансформуємо онлайн-дані таким чином, щоб отримати рівно чотири спостереження протягом місяця та уникнути проблем із перетворенням частот. Це виконується шляхом поділу місяця на чотири частини та зіставлення онлайн-даних (наприклад, перші сім днів, другі сім днів, треті сім днів і решта днів). За наявності більш ніж одного спостереження протягом певного періоду в місяці береться їх просте середнє. Після цього ми генеруємо індекси в місячному вираженні й конвертуємо тижневий ряд даних у місячний ряд даних:

$$\Delta_4 p_{i,m}^{w} = \prod_{j=1}^{4} \left(\Delta p_{i,m+w-j} + 1\right),$$

де $\Delta_4 p_{i,m}^{w}$ означає місячну зміну онлайн-цін у тиждень $w=[1:4]$. У результаті отримуємо чотири щотижневі часові ряди, які відображають зміну цін за один місяць.

На графіку 1 відображено декілька онлайн-індексів окремих компонент ІСЦ разом із їхніми офіційними відповідниками. Ми наводимо онлайн-індекси, побудовані на третьому тижні кожного місяця, оскільки Державна служба статистики України зазначає, що збирає дані про ціни приблизно протягом цього часу. Відповідність онлайн-цін офіційним даним відрізняється за компонентами. Наприклад, онлайн-індекси цін на яйця, яблука, виноград та кефір досить точно передають загальні тенденції та короткострокові коливання офіційної статистики, водночас помилки не перевищують двох стандартних відхилень. Деякі онлайн-індекси, наприклад, для батонів, мороженої риби та соняшникової олії, відображають довгострокову динаміку місячної інфляції, однак можуть суттєво відрізнятися в короткостроковій перспективі. Відхилення онлайн-індексів для м'ясної вирізки та шоколаду від офіційних даних, своєю чергою, може бути тривалішим протягом деяких періодів.

Крім простої візуалізації, ми також перевіряємо узгодженість онлайн-індексів з офіційною статистикою шляхом розрахунку середньої квадратичної похибки (від англ. Root Mean Square Error – RMSE) у межах окремих компонент ІСЦ (див. таблицю 3). Для цілей порівняння ми також надаємо RMSE, скориговані на стандартні відхилення офіційної інфляції відповідної компоненти, оскільки їх волатильність може істотно відрізнятися. Ми наводимо розрахунки для чотирьох тижневих індексів місячної зміни цін. Згідно з результатами навіть скориговані RMSE можуть суттєво відрізнятися для окремих компонент ІСЦ. Середні RMSE для всіх компонент удвічі перевищують стандартне відхилення офіційної інфляції. Тоді як мінімальне значення скоригованої RMSE становить близько 0.5%, максимальне значення перевищує 11%. Водночас для майже 70% онлайн-індексів помилки, що переоцінюють офіційну інфляцію, переважають над тими, що її недооцінюють.

Для визначення факторів, які зумовлюють розбіжності між онлайн-індексами цін та офіційними індексами цін, ми застосовуємо різні методи фільтрації та вивчаємо властивості даних, які поліпшують або погіршують здатність онлайн-індексів апроксимувати офіційні дані. Для цього будуємо альтернативні онлайн-індекси шляхом вибіркового виключення товарів із масиву даних. Зокрема, проводимо 99 ітерацій, у яких кожен товар може залишитися в загальній вибірці з імовірністю 1%, 2% і до 99%. Для кожного рівня ймовірності повторюємо описану процедуру 100 разів і, зрештою, отримуємо 9 900 альтернативних масивів даних із різним складом товарів. Для кожного альтернативного масиву даних будуємо чотири тижневих онлайн-індекси місячної зміни цін, як було описано вище. Потім порівнюємо побудовані індекси з офіційною статистикою шляхом розрахунку RMSE. Враховуючи, що кожен альтернативний масив даних складається з різної кількості товарів із різною кількістю відсутніх спостережень і унікальним середнім стандартним відхиленням у межах вузько визначеної групи, ми тепер можемо дослідити, які особливості масиву даних впливають на здатність онлайн-індексів апроксимувати офіційну статистику. Для цього оцінюємо панельну регресію такої форми:

[3] Збір даних проводився на щоденній основі, однак ми використовуємо спостереження, отримані шляхом розрахунку середньої ціни за тиждень. Це дає змогу уникати проблем із надмірною кількістю втрачених спостережень і тимчасових помилок у програмному коді веб-скрепінгу.
[4] Наш підхід подібний до підходу Cavallo (2013), однак ми використовуємо просте середнє змін онлайн-цін замість геометричного середнього.

$$RMSE_{i,k} = \beta_o + \beta_1 Obs_{i,k} + \beta_2 Members_{i,k} + \\ + \beta_3 SD_{i,k} + u_i + \varepsilon_{ik},$$

де $RMSE_{i,k}$ позначає RMSE усереднену для чотирьох тижневих індексів у межах компоненти i для ітерації k; $Obs_{i,k}$, $Members_{i,k}$, *та* $SD_{i,k}$ позначають відповідно середню кількість наявних спостережень у групі, кількість товарів у групі, а також середнє стандартне відхилення групи; u_i означає крос-компонентний незмінний у часі фіксований ефект, що дає змогу відобразити особливості окремої компоненти; насамкінець ε_{ik}, позначає залишки.

Оцінені коефіцієнти свідчать (див. таблицю 4), що за вищої волатильності онлайн-цін у межах окремої компоненти розбіжність побудованих індексів з офіційною статистикою зростає. Водночас зменшення кількості товарів у масиві даних впливає на зростання RMSE, і, як наслідок, рівень відповідності офіційним даним знижується. Таким чином, розбіжності між онлайн-інфляцією та офіційною інфляцією можуть бути спричинені недостатнім покриттям масиву даних. Неочікувані результати отримано для середньої кількості наявних спостережень, адже кореляція з помилками прогнозу є позитивною. Відповідно збільшення кількості наявних спостережень у вибірці впливає на збільшення RMSE.

Далі ми відфільтровуємо товари, що погіршують здатність онлайн-індексів відображати офіційну статистику. Спершу виключаємо з вибірки ті товари, ціни на які характеризуються високими стандартними відхиленнями порівняно із середніми значеннями в межах окремих компонент. Ми будуємо альтернативні онлайн-індекси шляхом виключення верхнього та нижнього процентиля стандартних відхилень (а саме 121 ітерація від 0 до 50-го верхнього та нижнього процентиля) і розраховуємо частку індексів, RMSE яких нижча від середнього значення серед усіх ітерацій. На графіку 2 наведено результати експерименту.

Результати аналізу свідчать, що виключення товарів аж до верхнього 20-го процентиля у вузько визначеній групі з високими стандартними відхиленнями підвищує рівень відповідності онлайн-індексів офіційним даним, що також підтверджує результати оцінки панельної регресії. Натомість виключення нижнього процентиля не зменшує RMSE. З одного боку, це може вказувати на те, що масив онлайн-даних містить певні викиди, які з технічних причин з'являються в процесі веб-скрепінгу. Зокрема, ціна може суттєво змінитися, якщо змінюється одиниця виміру кількості товару. Якщо роздрібний продавець змінює одиницю виміру кількості товару, але використовує ту саму сторінку в Інтернеті, програмний код веб-скрепінгу не може розпізнати такі зміни без зовнішнього втручання. Водночас частка таких викидів, спричинених технічними проблемами, мала б бути незначною. У нашому випадку, однак, дана частка може перевищувати 20 відсотків, а це вказує на те, що вкрай волатильні ціни можуть насправді відображати нову інформацію у короткостроковій перспективі, неохоплену офіційною статистикою.

З метою перевірки того, чи впливає кількість наявних спостережень на здатність онлайн-цін апроксимувати офіційну статистику, ми повторюємо описану вище процедуру шляхом виключення товарів із великою кількістю відсутніх спостережень. Однак RSME зростає зі збільшенням вимог до фільтрування, що підтверджують результати оцінки панельної регресії. Враховуючи те, що кількість товарів у групі має негативну кореляцію з RMSE, додаткове виключення товарів погіршує рівень відповідності онлайн-індексів офіційним даним. Згідно з отриманими результатами наявність у масиві онлайн-даних товарів із великою кількістю спостережень у вибірці не обов'язково гарантує кращі результати, тому включення рідко відстежуваних товарів із великою кількістю пропущених спостережень не зменшує рівень збігів.

Зрештою, ми тестуємо рівень відповідності онлайн-індексів, згенерованих для різних тижнів місяця. Мета цього завдання — дослідити здатність онлайн-індексів апроксимувати офіційну статистику з появою оновлених щотижня даних. Крім тижневих індексів, що відображають зміну цін за останні чотири тижні, розраховуємо середню онлайн-інфляцію з плином часу. Наприклад, наприкінці першого тижня місяця ми маємо інформацію про те, як змінилися ціни порівняно з першим тижнем попереднього місяця. Наприкінці наступного тижня, на доповнення до першотижневої інфляції, отримуємо дані про інфляцію на другому тижні. З метою кращого врахування динаміки цін можемо також розрахувати середнє значення інфляції за перший та другий тижні. Те саме стосується наступних тижнів. Ми також порівнюємо офіційні показники інфляції за певний місяць з онлайн-інфляцією за останній тиждень попереднього місяця та перший тиждень наступного місяця. На графіку 3 відображено значення RMSE, усереднене для всіх компонент ІСЦ для різних тижневих індексів місячної зміни цін.

Згідно з результатами частка онлайн-індексів із найнижчим рівнем RMSE є найвищою для онлайн-інфляції за другий тиждень. Варто зауважити, що Державна служба статистики України збирає дані про ціни на початку другої половини місяця, що узгоджується з нашими результатами. Водночас середнє значення RMSE для всіх онлайн-індексів, які відображають усереднене значення індексів за попередні тижні, зменшується з плином часу та появою оновлених даних. Це є додатковим підтвердженням того, що онлайн-інфляція може випереджати офіційні показники, а тому є гарним орієнтиром для прогнозів.

Отже, згідно з результатами нашого аналізу онлайн-інфляція загалом відповідає офіційним даним, однак розбіжності можуть виникати на рівні окремих компонент ІСЦ. Ці відмінності можна пояснити як властивостями масиву даних, як-от покриттям товарів, так і тим фактом, що онлайн-ціни насправді відображають нову інформацію, яка не фіксується офіційною статистикою. Зокрема, онлайн-ціни можуть бути значно волатильнішими і швидше реагувати на нові економічні умови.

3.2. Онлайн — індекси агрегованих категорій ІСЦ

У попередньому розділі ми побудували онлайн-індекси споживчих цін у межах окремих компонент ІСЦ. Надалі переходимо до побудови загального онлайн-індексу споживчих цін та окремих агрегованих категорій, що дасть змогу дослідити відповідність онлайн-цін загальним інфляційним тенденціям у країні.

Ми використовуємо кілька альтернативних підходів до агрегування онлайн-індексів. По-перше, це просте

середнє значення всіх онлайн-індексів окремих категорій ІСЦ. Зокрема, для агрегування загальної інфляції використовуємо просте середнє значення всіх онлайн-індексів, водночас для онлайн-цін на продукти харчування включаємо лише ті індекси, які належать до категорії продуктів харчування. Крім цього, також використовуємо ваги структури споживчого кошика, що надається Державною службою статистики України, та агрегуємо онлайн-індекси як зважене середнє. Оскільки масив онлайн-даних містить до 46% споживчого кошика в Україні (134 із 328 компонентів), розраховуємо відносні ваги з використанням лише компонентів, представлених у масиві даних. Насамкінець для порівняння результатів за різних типів агрегації розраховуємо індекс, що містить середню динаміку цін на всі товари в масиві даних без побудови індексів окремих компонент. У таблиці 5 відображено RMSE агрегованих онлайн-індексів для загального ІСЦ, розрахункового ІСЦ (містить лише компоненти, представлені в масиві онлайн-даних), ІСЦ на продукти харчування та його окремі категорії, ІСЦ на напої, а також на алкогольні та тютюнові вироби. На графіку 4 візуалізовано отримані результати і відображено офіційну та онлайн-інфляцію.

Для більшості агрегованих онлайн-індексів зважене середнє онлайн-індексів окремих компонент поліпшує рівень відповідності онлайн-даних офіційній статистиці. Зокрема, для розрахункового ІСЦ, що включає лише компоненти, подані в онлайн-масиві даних, RMSE зменшується з 1.06% до 0.93%. Те саме стосується агрегованих індексів на продукти харчування, оскільки наш масив онлайн-даних охоплює здебільшого саме продукти харчування. Незважаючи на те, що результати агрегованих онлайн-індексів для загального ІСЦ є змішаними, RMSE не перевищує 1%. Це вказує на те, що в той час як частка компонент, яка не представлена в масиві даних, відіграє важливу роль, наш масив онлайн-даних спроможний охопити загальну динаміку цін у країні, оскільки середня квадратична похибка для більшості індексів не перевищує одного стандартного відхилення офіційної статистики.

4. ВИСНОВКИ

Швидкий розвиток електронної торгівлі впродовж останніх десятиліть розширює можливості державних органів, відповідальних за здійснення економічної політики, спостерігати за поточними тенденціями в економіці країни за допомогою аналізу великих масивів даних. У цій статті ми розраховуємо онлайн-індекс споживчих цін, використовуючи багатий масив даних онлайн-цін, отриманий у процесі веб-скрепінгу в Національному банку України, і порівнюємо його узгодженість із даними офіційної статистики. Спершу ми генеруємо онлайн-індекси в межах окремих компонент ІСЦ, а потім агрегуємо їх до загального ІСЦ, а також окремих категорій ІСЦ. Наш унікальний масив даних містить близько трьох мільйонів спостережень за роздрібними онлайн-цінами на споживчі товари в п'ятьох найбільших містах України за період із січня 2016 року до грудня 2017 року. Онлайн-дані охоплюють близько 46% споживчого кошика України.

Ми досліджуємо властивості масиву даних, які поліпшують чи погіршують його здатність відображати офіційну статистику. Згідно з нашими результатами онлайн-індекси цін загалом узгоджуються з офіційною статистикою, однак рівень відповідності онлайн-даних відрізняється для окремих компонент ІСЦ. Розбіжності лише частково можна пояснити технічними особливостями масиву даних. Зокрема, важливу роль відіграє кількість товарів у масиві даних, а з цього випливає, що здатність онлайн-індексів збігатися з даними офіційної статистики зростає, якщо онлайн-масив даних охоплює широкий спектр товарів у вузько визначеній групі. Натомість товари з великою кількістю спостережень у вибірці не обов'язково гарантують кращий рівень збігів, а це означає, що включення рідко відстежуваних товарів із великою кількістю відсутніх спостережень не впливає на рівень відповідності онлайн-індексів офіційним даним. Насамкінець використання офіційних ваг структури споживчого кошика під час побудови агрегованих індексів зменшує відхилення індексів онлайн-цін від даних офіційної статистики.

Водночас онлайн-ціни можуть насправді відображати нову інформацію, яка не фіксується офіційною статистикою. Онлайн-ціни на деякі товари можуть бути набагато волатильніші, а тому виключення таких товарів збільшує рівень відповідності онлайн-індексів. Здатність високочастотних онлайн-даних наближатися до офіційної місячної інфляції зростає за умови врахування ширшого періоду змін онлайн-цін. Це вказує на те, що онлайн-ціни можуть швидше реагувати на нові економічні умови, а отже, можуть слугувати орієнтиром для прогнозування офіційної статистики.

Наш аналіз підтверджує значну кількість фактів у літературі (наприклад, Cavallo and Rigobon, 2016; Breton et. al., 2015) про те, що онлайн-ціни можуть використовуватися як додаткове джерело інформації для спостереження поточної динаміки інфляції. Вони можуть також використовуватися для так званого наукастингу або короткострокового прогнозування, оскільки онлайн-дані доступні в режимі реального часу і з високою частотою. Таким чином, мета подальших наших досліджень — розробка уніфікованої методології для наукастингу для інфляції на основі онлайн-даних разом із більш традиційними підходами. Зокрема, онлайн-ціни здатні потенційно поліпшувати результати динамічних факторних моделей, які зазвичай використовуються для наукастингу макроекономічних показників.

ЛІТЕРАТУРА

• Breton R., Clews G., Metcalfe L., Milliken N., Payne C., Winton J., Woods A. (2015). Research Indexes Using Web Scraped Data. Office for National Statistics, UK.

• Cavallo A., Rigobon R. (2016). The Billion Prices Project: Using Online Prices for Measurement and Research. Journal of Economic Perspectives, Vol. 30, No. 2, pp. 151-178.

• Cavallo A. (2013). Online and Official Price Indexes: Measuring Argentina's Inflation. Journal of Monetary Economics, Vol. 60, Issue 2, pp. 152-165.

• Coupe T., Petrusha E. (2014). Can We Trust Official Inflation Measures? A Check Based on Inflation at Ukrainian Online Supermarkets. Focus Ukraine.

• Griffioen R., Haan J., Willenborg L. (2014). Collecting Clothing Data from the Internet. Proceedings of Meeting of the Group of Experts on Consumer Price Indexes, May 26–28.

• Gorodnichenko Y., Talavera O. (2017). Price Setting in Online Markets: Basic Facts, International Comparisons, and Cross-Border Integration. American Economic Review, Vol. 107, No. 1, pp. 249-282.

• Hammond G. (2011). State of the Art Inflation Targeting. Center for Central Banking Studies Handbook, No. 29, Bank of England, London.

• Horrigan M. W. (2013). Big Data: A Perspective from the BLS. Amstat News, January 1. Available at http://magazine.amstat.org/ blog/ 2013/ 01/ 01/ sci-policy-jan2013/

• Jahan S. (2017). Inflation Targeting: Holding the Line. Finance & Development (IMF). Available at http: // www.imf.org / external / pubs / ft / fandd / basics / target.htm

• Krsinich F. (2015). Price Indexes from Online Data Using the Fixed-Effects WindowSplice (FEWS) Index. Paper presented at the Ottawa Group, Tokyo, Japan, May 20-22, 2015.

• Nygaard R. (2015). The Use of Online Prices in the Norwegian Consumer Price Index. Paper prepared for the meeting of the Ottowa Group, Tokyo, Japan, May 20-22, 2015.

• Roger S. (2010). Inflation Targeting Turns 20. Finance & Development, Vol. 47, No. 1, IMF, pp. 46-49.

ДОДАТОК. ТАБЛИЦІ Й ГРАФІКИ

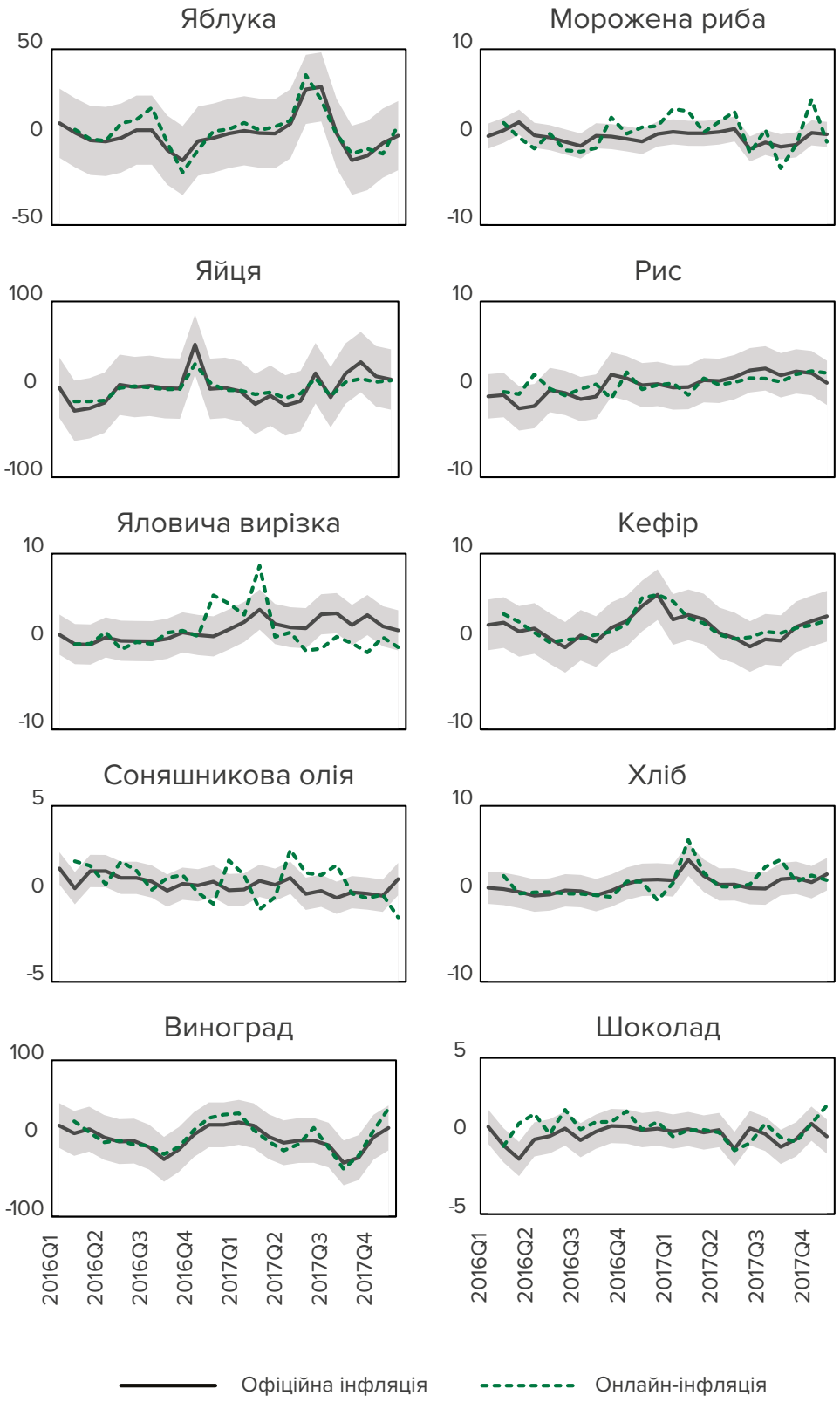
Таблиця 1. Описова статистика офіційної інфляції

Індекс	% споживчого кошика	Кількість компонент	Ст. відх.	Мін.	Сер.	Макс.
ІСЦ	100	328	1.00	-0.36	1.04	3.52
Продукти харчування	39.6	113	1.26	-1.06	0.81	3.42
- Хліб	7.3	21	0.69	-1.15	0.66	1.96
- М'ясо	10.1	23	1.61	-1.32	1.33	5.15
- Риба	2.2	9	0.59	-0.90	0.15	1.26
- Молоко	6.2	14	3.83	-5.15	1.29	11.43
- Жири	4.6	6	1.13	-0.36	1.16	3.60
- Фрукти	2.3	10	5.46	-5.70	1.05	13.45
- Овочі	2.4	16	10.04	-21.50	-0.31	16.69
- Цукор	3.4	7	0.92	-1.90	0.36	2.21
Напої	1.4	7	0.28	-0.04	0.42	1.03
Алкоголь	9.2	12	1.28	-1.91	1.69	3.33

Таблиця 2. Описова статистика масиву онлайн-даних

Індекс	% споживчого кошика	Відн. частка	Кількість компонент	К-сть товарів, 1000	К-сть спостережень, 1м	Сер. ст. відх.
ІСЦ	45.7	45.7	134	75.1	2.48	4.96
Продукти харчування	34.1	86.2	93	34.3	1.11	5.57
- Хліб	6.8	93.0	19	8.19	0.29	3.39
- М'ясо	7.1	70.7	16	3.51	0.11	3.42
- Риба	2.2	100	9	2.57	0.09	4.01
- Молоко	5.4	86.0	11	4.96	0.15	3.80
- Жири	4.4	96.8	4	0.72	0.03	3.99
- Фрукти	1.4	60.2	5	0.74	0.02	9.66
- Овочі	2.3	99.0	15	1.90	0.06	13.18
- Цукор	3.4	100	7	6.93	0.20	3.41
Напої	1.4	98.9	6	10.5	0.40	4.71
Алкоголь	6.2	67.6	7	9.92	0.37	3.59

Графік 1. Вибрані онлайн-індекси компонент ІСЦ разом із даними офіційної статистики в місячному вираженні (%)



Таблиця 3. Рівень відповідності онлайн-індексів офіційній статистиці

		Тиждень I	Тиждень II	Тиждень III	Тиждень IV
RMSE	*Сер.*	4.97	4.97	5.09	5.30
	Мін.	0.73	0.78	0.65	0.78
	Макс.	84.75	110.30	75.82	86.17
Скоригована RMSE*	*Сер.*	2.68	2.54	2.61	2.73
	Мін.	0.49	0.48	0.39	0.50
	Макс.	11.13	11.70	11.76	28.08
Сер. помилка переоцінки		3.99	3.81	3.95	4.11
Сер. помилка недооцінки		2.89	2.67	2.83	3.14
Частка індексів із переважно помилками переоцінки		70%	72%	72%	72%

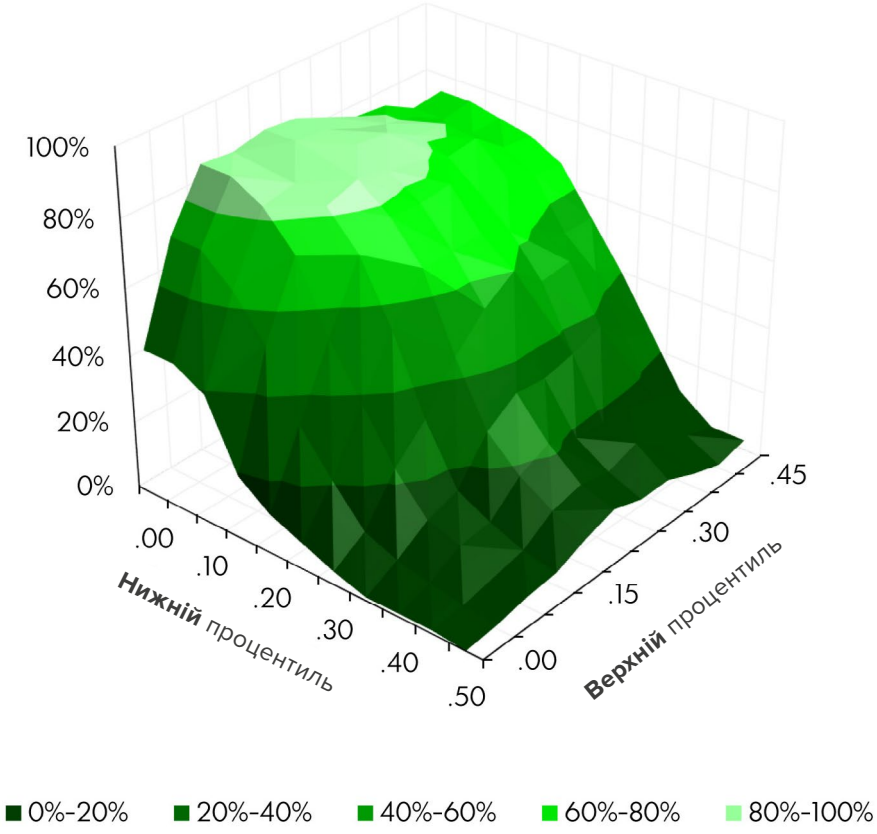
* Середні квадратичні похибки, скориговані відповідно до стандартного відхилення відповідної компоненти офіційної інфляції для цілей порівняння.

Таблиця 4. Детермінанти рівня відповідності онлайн-індексів офіційній статистиці (панельна регресія)

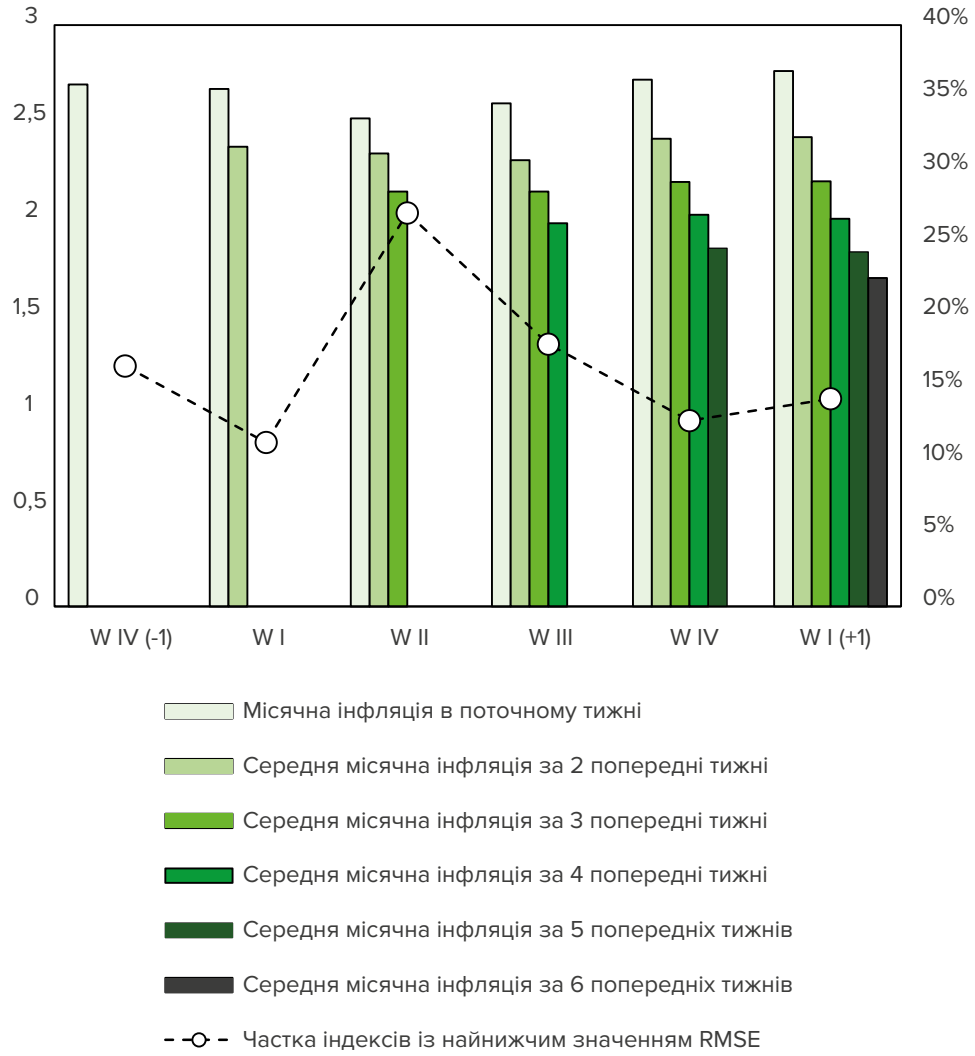
RMSE	**1)**	**2)**	**3)**	**4)**	**5)**	**6)**	**7)**
Середня кількість спостережень	8.661* (0.137)			8.470* (0.137)	3.957* (0.073)		3.784* (0.073)
Кількість найменувань товарів		-0.001* (0.000)		-0.001* (0.000)		-0.001* (0.000)	-0.001* (0.000)
Сер. ст. відх. групи			1.870* (0.001)		1.868* (0.001)	1.870* (0.001)	1.867* (0.001)
Фіксований ефект	*V*	*V*	*V*	*V*	*V*	*V*	*V*
R^2	0.844	0.838	0.955	0.844	0.955	0.955	0.956

Примітка: Символ « * » означає значущість на рівні 1%.

Графік 2. Виключення верхніх та нижніх процентилів стандартних відхилень інфляції: частка компонентів ІСЦ, RMSE яких є нижчою від середнього значення



Графік 3. Рівень відповідності онлайн-індексів на різних тижнях протягом місяця: усереднене значення RMSE (ліва шкала), частка компонент із найнижчим значенням RMSE (права шкала)



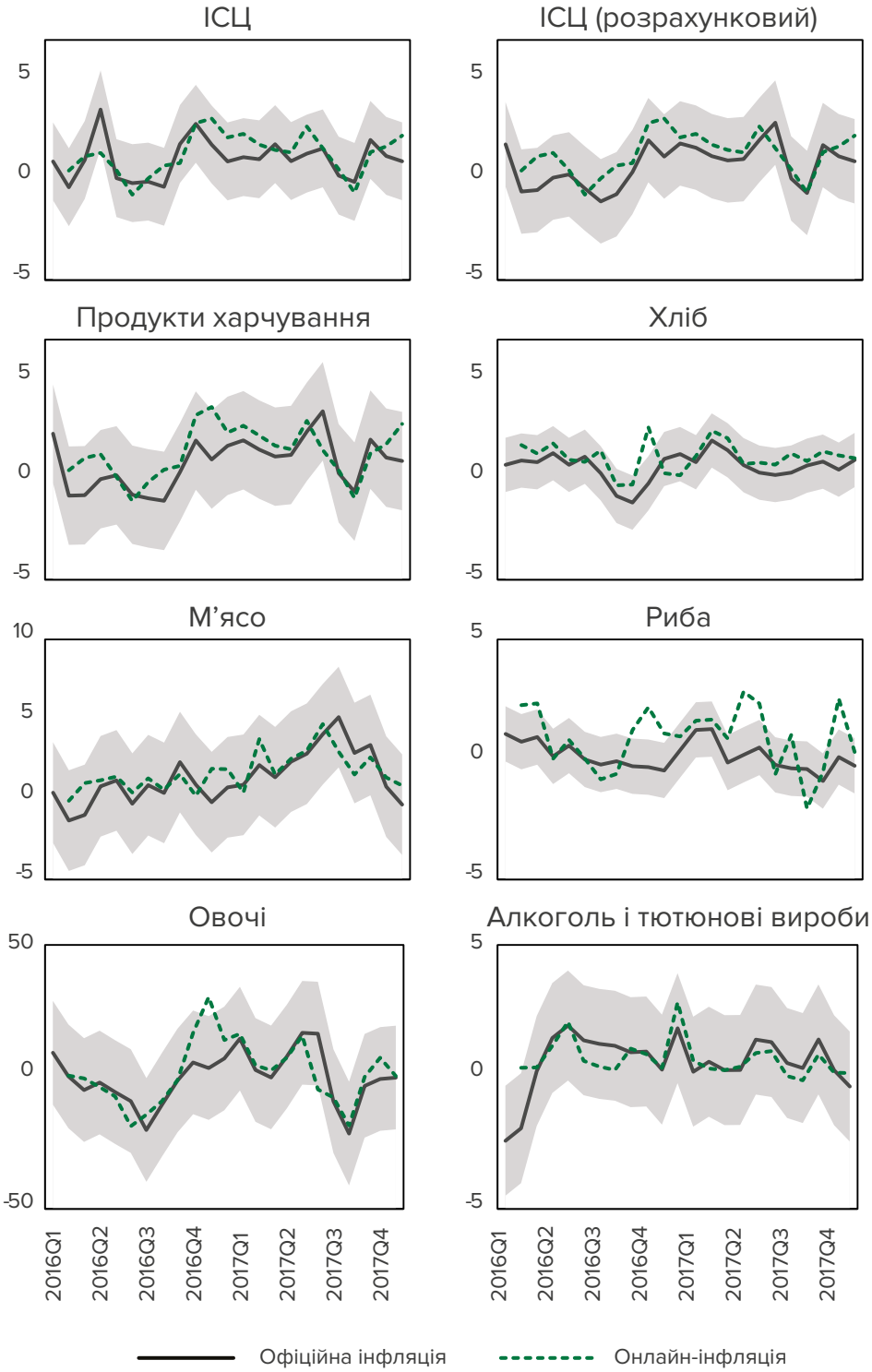
Примітка. RMSE скориговані відповідно до стандартного відхилення відповідної категорії офіційної інфляції для цілей порівняння.

Таблиця 5. RMSE агрегованих онлайн-індексів місячної зміни цін

Індекс	Просте середнє всіх товарів	Просте середнє онлайн-індексів компонент	Зважене середнє онлайн-індексів компонент
ІСЦ	**0.81 (0.82)**	1.89 (1.90)	0.90 (0.90)
ІСЦ (розрахунковий)	1.06 (0.99)	1.92 (1.78)	**0.93 (0.87)**
Продукти харчування	1.21 (0.97)	1.98 (1.58)	**1.14 (0.91)**
- Хліб	**0.79 (1.14)**	2.31 (3.33)	0.85 (1.23)
- М'ясо	1.37 (0.85)	2.96 (1.84)	**1.07 (0.67)**
- Риба	**1.16 (1.97)**	2.60 (4.43)	1.33 (2.26)
- Молоко	2.95 (0.77)	3.37 (0.88)	**1.61 (0.42)**
- Жири	4.65 (4.11)	**1.68 (1.48)**	**2.79 (2.47)**
- Фрукти	**4.06 (0.74)**	5.85 (1.07)	5.43 (1.00)
- Овочі	**8.67 (0.86)**	9.04 (0.90)	9.45 (0.94)
- Цукор	1.11 (1.21)	2.56 (2.79)	**0.78 (0.85)**
Напої	0.81 (2.93)	2.47 (8.90)	**0.80 (2.90)**
Алкоголь	1.20 (0.94)	2.51 (1.97)	**0.82 (0.64)**

Примітка. Значення в дужках відображають RMSE, скориговану на стандартне відхилення офіційної інфляції відповідної категорії для цілей порівняння.

Графік 4. Агрегована онлайн- та офіційна інфляція в місячному вираженні (%)



Примітка. Агреговані онлайн-індекси побудовані як зважене середнє індексів компонент, які також є середнім значенням чотирьох тижневих онлайн-індексів місячної зміни цін.