Національний банк України

Великі дані для макроекономічного аналізу

Тетяна Юхименко

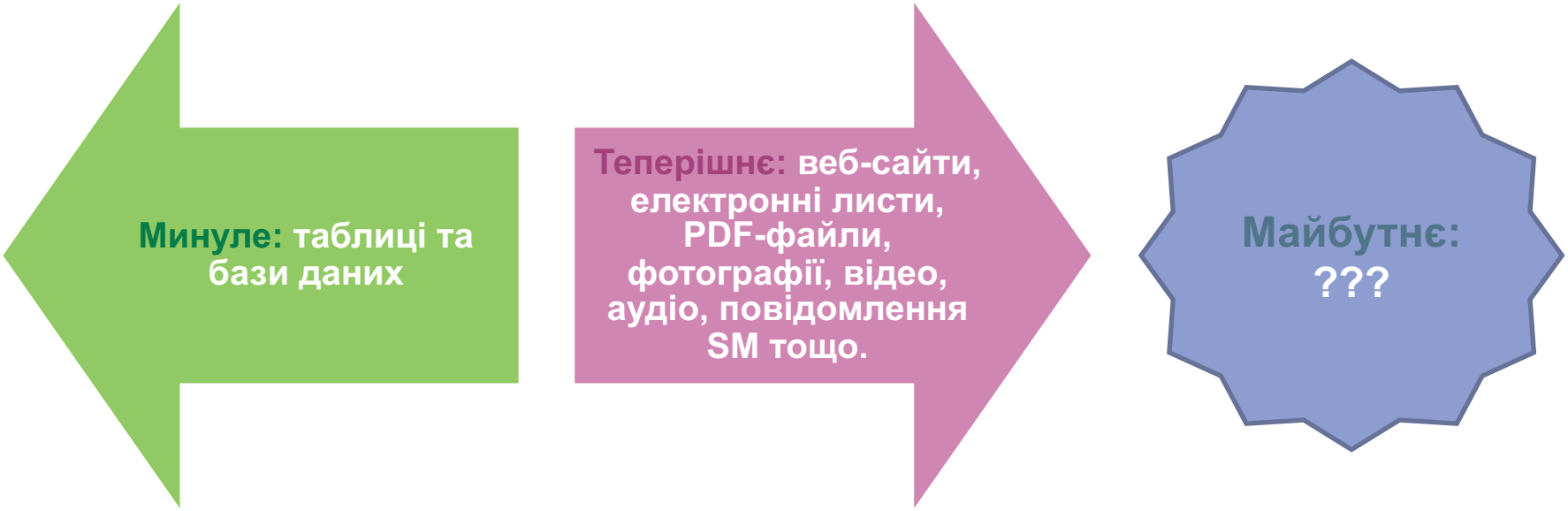
Департамент монетарної політики та економічного аналізу

27 травня 2021 року

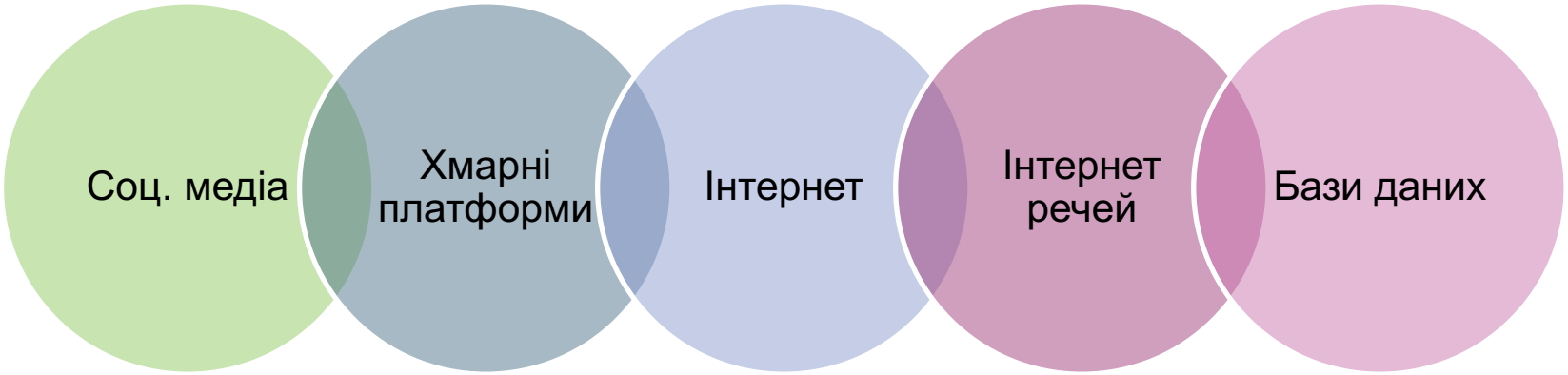
Що таке великі дані?

Великі дані – це об’ємні, швидкі чи складні дані, які важко або неможливо обробити традиційними методами

Звідки беруться великі дані?



Джерела великих даних



- Коментарі, лайки
- Загальнодоступні профілі
- Картинки, відео, аудіо тощо

- Дані, що зберігаються у хмарних сервісах

- Публічна мережа
- Відкритий API
- Dark web

- Інформація, згенерована машиною
- Дані, що генеруються з датчиків

- Традиційні та сучасні бази даних

Види даних

Структуровані

- **Попередньо визначені та відформатовані до заданої структури перед розміщенням у сховищі даних**
- **Легко використовується алгоритмами машинного навчання**
- **Легко використовується користувачами**
- **Розширений доступ до більшої кількості інструментів**

Обмежене використання

Неструктуровані

- **Зберігається у власному форматі та не обробляється, поки не використовується**
- **Свобода рідного формату**
- **Швидші темпи накопичення**

Потрібна експертиза в data science

Переваги великих даних

Нові продукти

Розумніші рішення

Дані в режимі реального часу

Ефективність витрат

Скорочення часу

Типові проблеми у роботі з великими даними

Людські ресурси

- Недостатнє розуміння
- Нестача кваліфікованих робітників
- Забагато варіантів вибору

Управління даними

- Складні системи управління даними
- Низька якість та неточні дані
- Обробка великих наборів даних
- Постійні зміни даних
- Змішана частоти та час
- Прогалини у спостереженнях

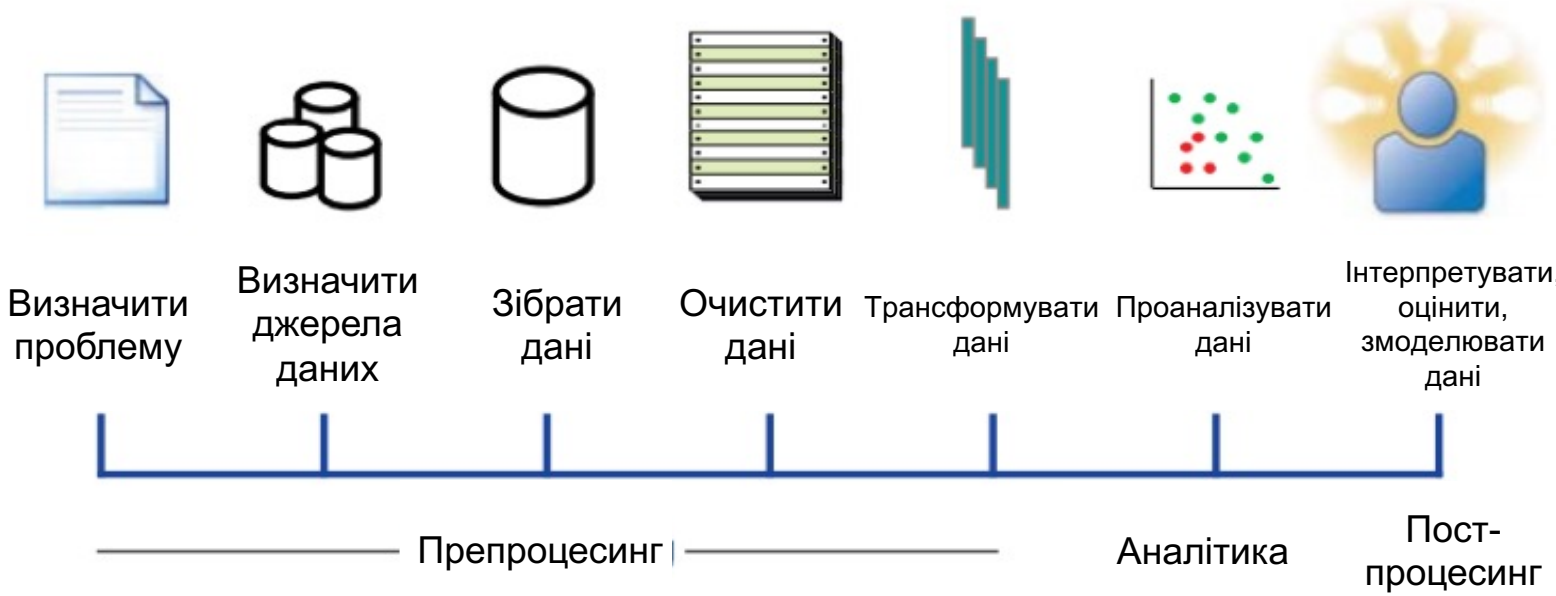
Інше

- Висока вартість рішень, у тому числі через масштабування даних
- Проблеми безпеки
- Доступність
- Темп розвитку технологій

Цінність великих даних

Великі дані **Аналітика** **Рішення**

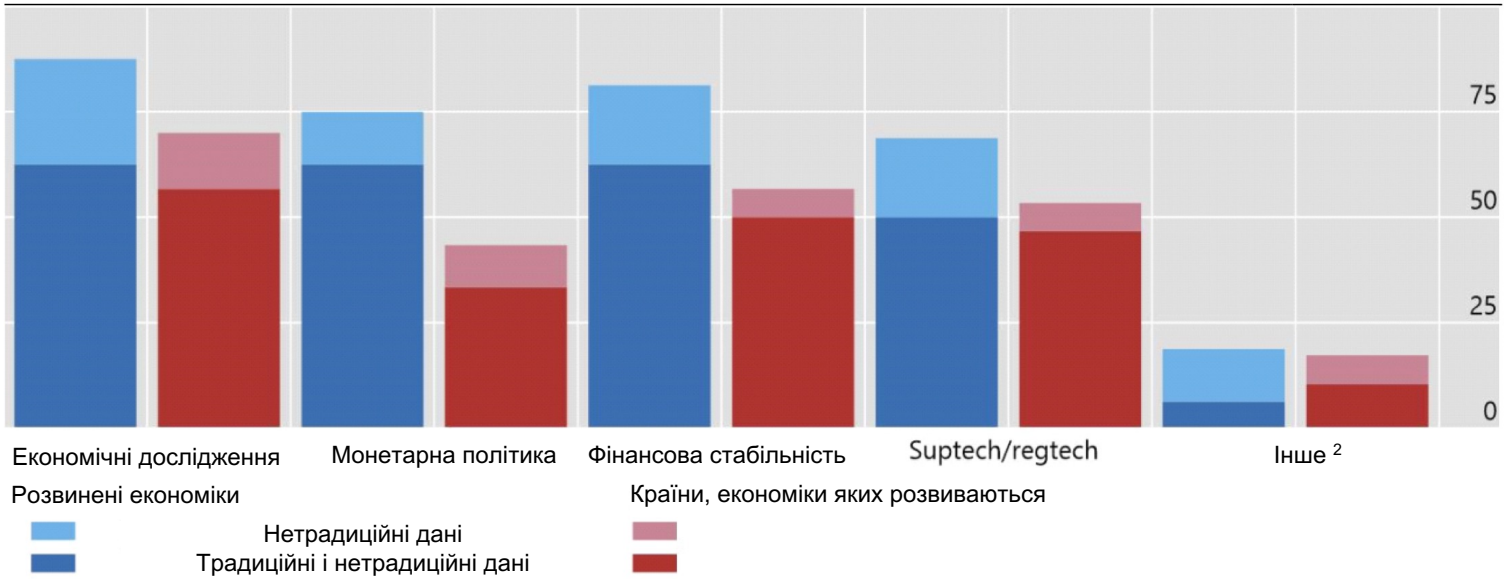
Обробка великих даних



Методи аналізу великих даних

Великі дані для центральних банків

Цілі, для яких центральні банки використовують великі дані[1]



[1] Графік відображає частку респондентів, які обрали відповідь на запитання "Для яких загальних цілей ваша установа використовує великі дані?". Респонденти могли вибрати кілька варіантів.
[2] Включає "моніторинг крипто-активів", "кібербезпеку" та "аналіз мережі".
Джерело: IFC (2020).

Застосування великих даних у макроекономіці

Великі дані створили можливості для вивчення економічних систем та прийняття рішень з високим ступенем точності

Переваги

- Доступні складні дані та існує можливість пов’язати багато наборів даних разом.
- Нові потужні аналітичні підходи на основі машинного навчання, дедалі частіше доповнюють основні економетричні інструменти.
- Великі дані дозволяють краще прогнозувати економічні явища та покращують виявлення причинно-наслідкових зв’язків.
- Дозволяють дослідникам створювати прості моделі, що описують дуже великі, складні набори даних.
- Методи машинного навчання та великі дані також дозволяють моделювати складні взаємозв'язки.

Недоліки

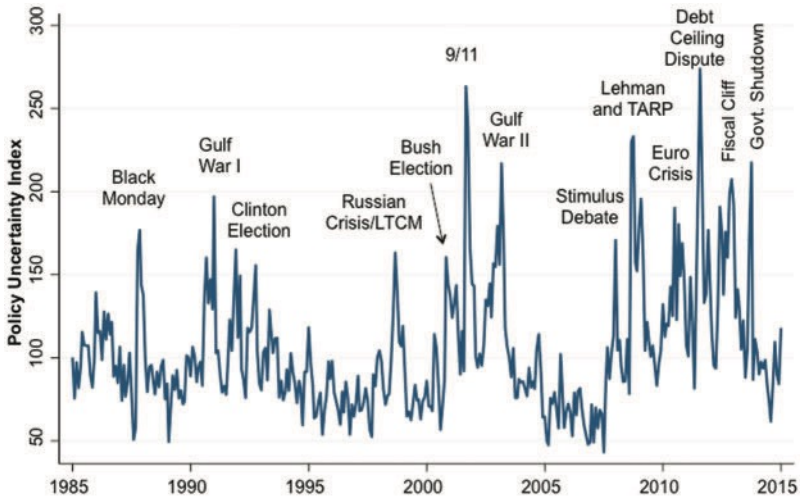
- Прогнози, засновані на великих даних, можуть мати проблеми з конфіденційністю.
- Робота з великими даними вимагає значних обчислювальних потужностей, можуть вимагати високого ступеня точного налаштування для оптимальної роботи.
- Аналіз великих даних вимагає вкладень у технології та людські навички.
- Великі дані можуть страждати від упередженості вибору залежно від того, як і ким генеруються дані.
- Доступ довеликих даних може передбачати партнерство з фірмами, які обмежують свободу дослідників.

Джерело: https://wol.iza.org/articles/big-data-in-economics/long

Приклади застосування великих даних для макроекономічного аналізу

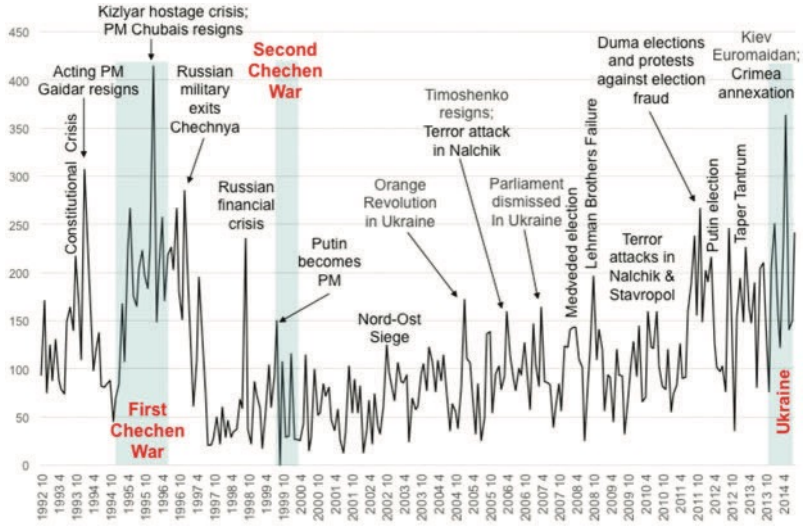
Індекс невизначеності економічної політики (Baker, 2016)

Індекс EPU для США



Index reflects scaled monthly counts of articles containing 'uncertain' or 'uncertainty', 'economic' or 'economy', and one or more policy relevant terms: 'regulation', 'federal reserve', 'deficit', 'congress', 'legislation', or 'white house'. The series is normalized to mean 100 from 1985-2009 and based on queries run on 2 February, 2015 for the USA Today, Miami Herald, Chicago Tribune, Washington Post, LA Times, Boston Globe, SF Chronicle, Dallas Morning News, NY Times, and the Wall Street Journal.

Індекс EPU для Росії



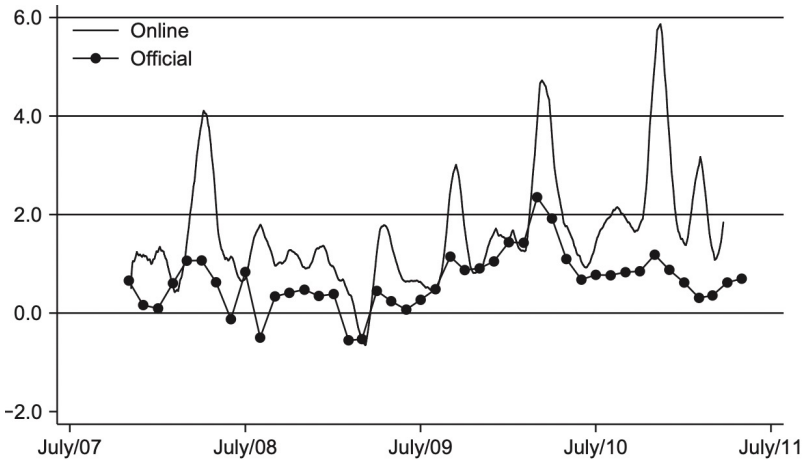
Index reflects scaled monthly counts of articles in Kommersant with Russian terms for 'uncertain' or 'uncertainty', 'economic' or 'economy', and one or more selected policy terms. The series is normalized to 100 and runs from October 1992 to August 2014.

Джерело: Baker (2016)

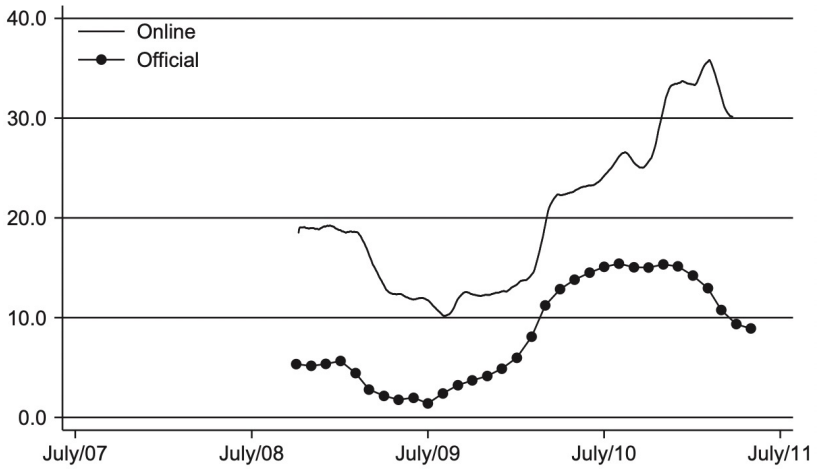
- У дослідженні розроблено новий індекс невизначеності економічної політики (EPU) на основі частоти висвітлення окремих тем у газетах.
- Результати свідчать про те, що підвищена невизначеність політики в США та Європі протягом останніх років могла погіршити макроекономічні показники.

Онлайнові та офіційні Індекси цін: вимірювання інфляції в Аргентині (Cavallo, 2012)

Офіційна та онлайн-інфляція в Аргентині,% м/м



Офіційна та онлайн-інфляція в Аргентині,% р/р

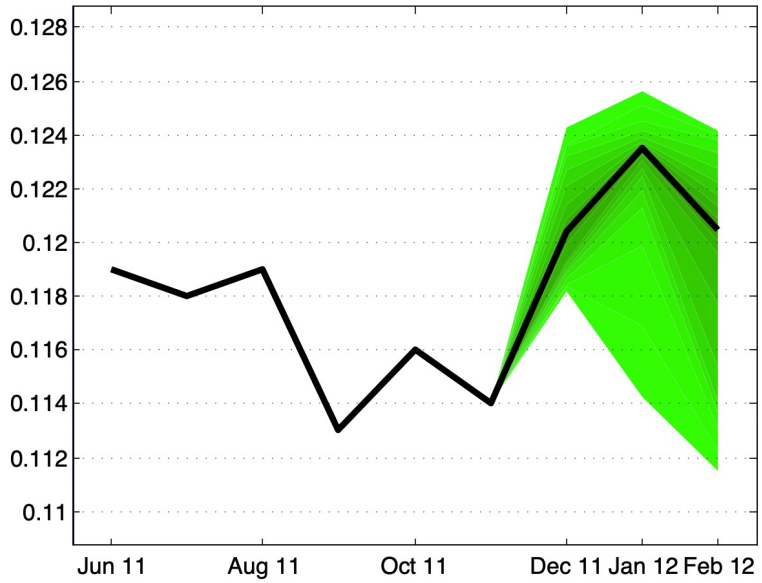


Джерело: Cavallo (2012)

- Дослідження вивчало відповідність офіційних оцінок інфляції в п'яти країнах Латинської Америки, приділяючи основну увагу Аргентині, де офіційна статистика зазнавала жорсткої критики.
- Дані були зібрані з жовтня 2007 року по березень 2011 року з найбільшого онлайн-супермаркету в кожній країні.
- У Бразилії, Чилі, Колумбії та Венесуелі індекси цін в Інтернеті приблизно відповідають рівню та основній динаміці офіційної інфляції. Натомість, рівень інфляції в Інтернеті в Інтернеті майже втричі перевищує офіційну оцінку.

Наукаст безробіття в Туреччині: запитаємо Google (Chadwick, 2012)

Наукаст безробіття в Туреччині (без с/г)

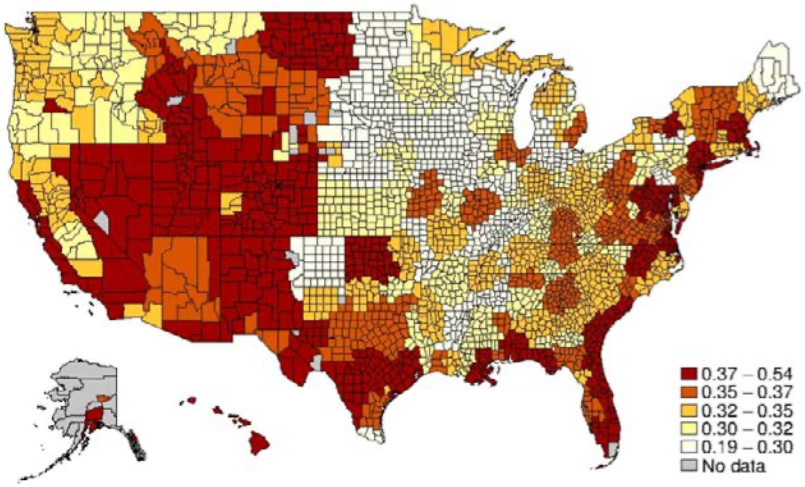


Джерело: Chadwick (2012).

- У роботі досліджено, чи можуть дані пошукових запитів Google покращити поточну ефективність щомісячного рівня безробіття в Туреччині (за виключенням с/г) за період з січня 2005 року по січень 2012 року.
- Дані пошукових запитів Google суттєво покращили наукаст щомісячного рівня безробіття. У порівнянні з базовою моделлю (проста авторегресійна модель), RMSE найкращої моделі з даними пошукових запитів Google виявилася на 47,8% кращим.

Виявлення регіонів ризику за допомогою Google Trends: вплив Covid-19 на ринки праці в США (Doerr, 2020)

Вплив Covid-19 на зайнятість



Кожна панель показує частку зайнятості у галузях, які, ймовірно, найбільше постраждають від Covid-19. Більш темні кольори означають, що більша частка зайнятих у зоні ризику.
Джерело: Doerr (2020).

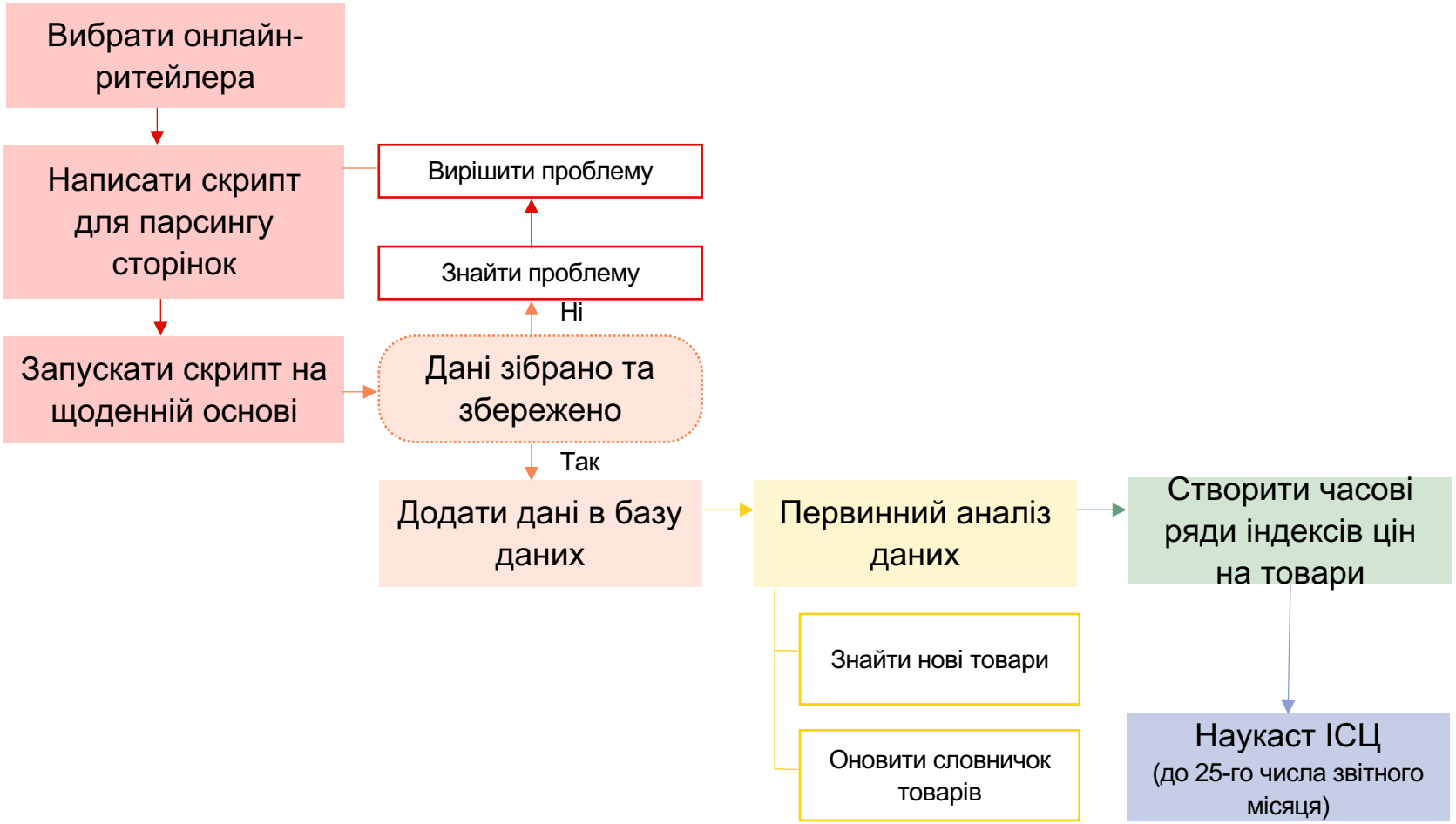
- У дослідженні перевірено гіпотезу, що в районах з вищим ступенем поширення Covid-19 більше пошукових запитів Google, пов'язаних з пандемією та виплатами по безробіттю.
- Поєднання офіційної статистики та даних у режимі реального часу з нетрадиційних джерел може додатково покращити розуміння політиками неоднорідного впливу шоку Covid-19 та їх здатності виробляти адекватні відповіді.

Застосування великих даних для макроекономічного аналізу в НБУ

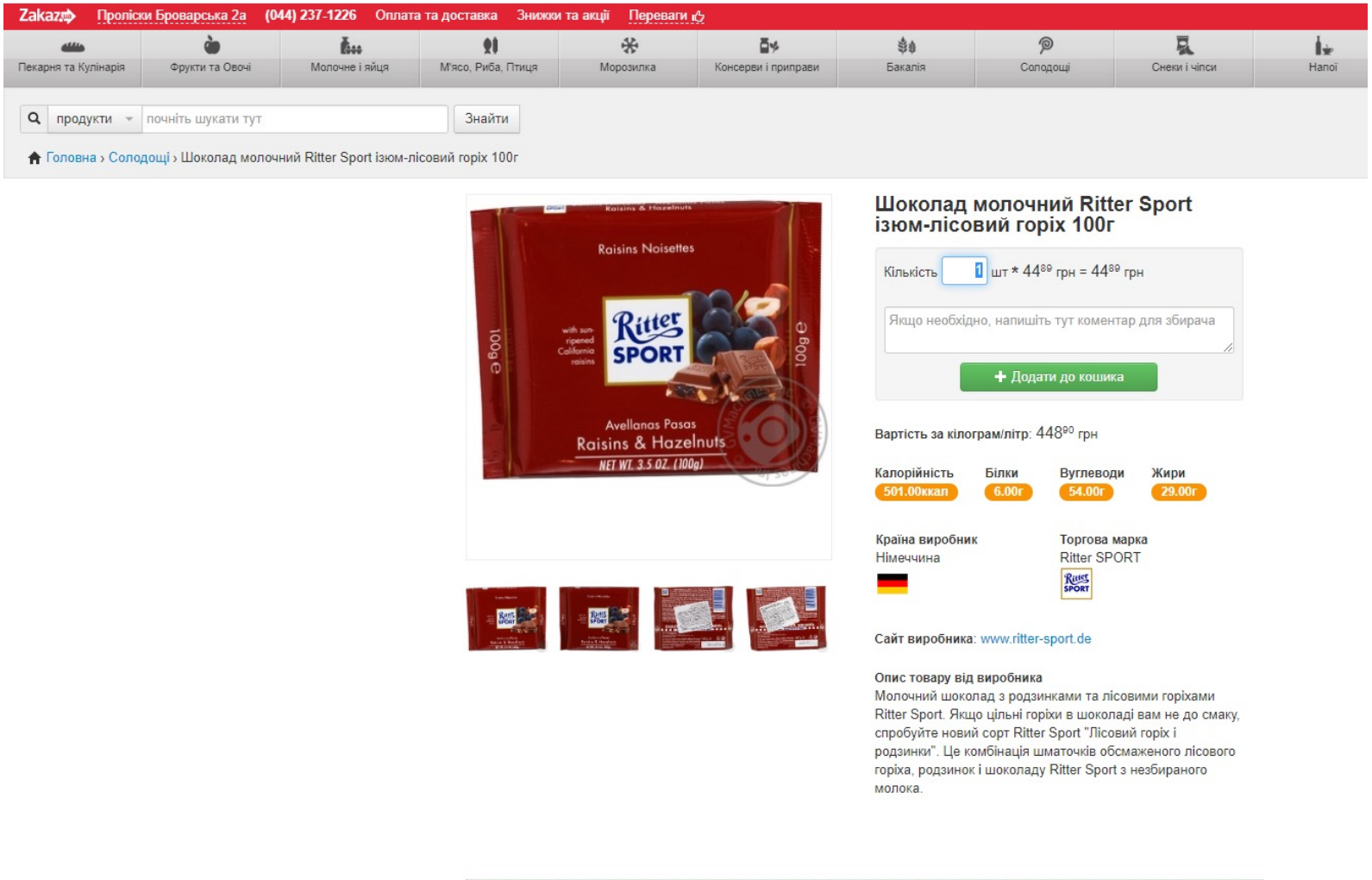
Традиційні статистичні дані часто публікуються з величезною затримкою

Інфляція	• 8-10 днів після закінчення звітного місяця
ВВП	• 45 днів після закінчення звітного кварталу
Безробіття	• ≈ 85 днів після закінчення звітного кварталу

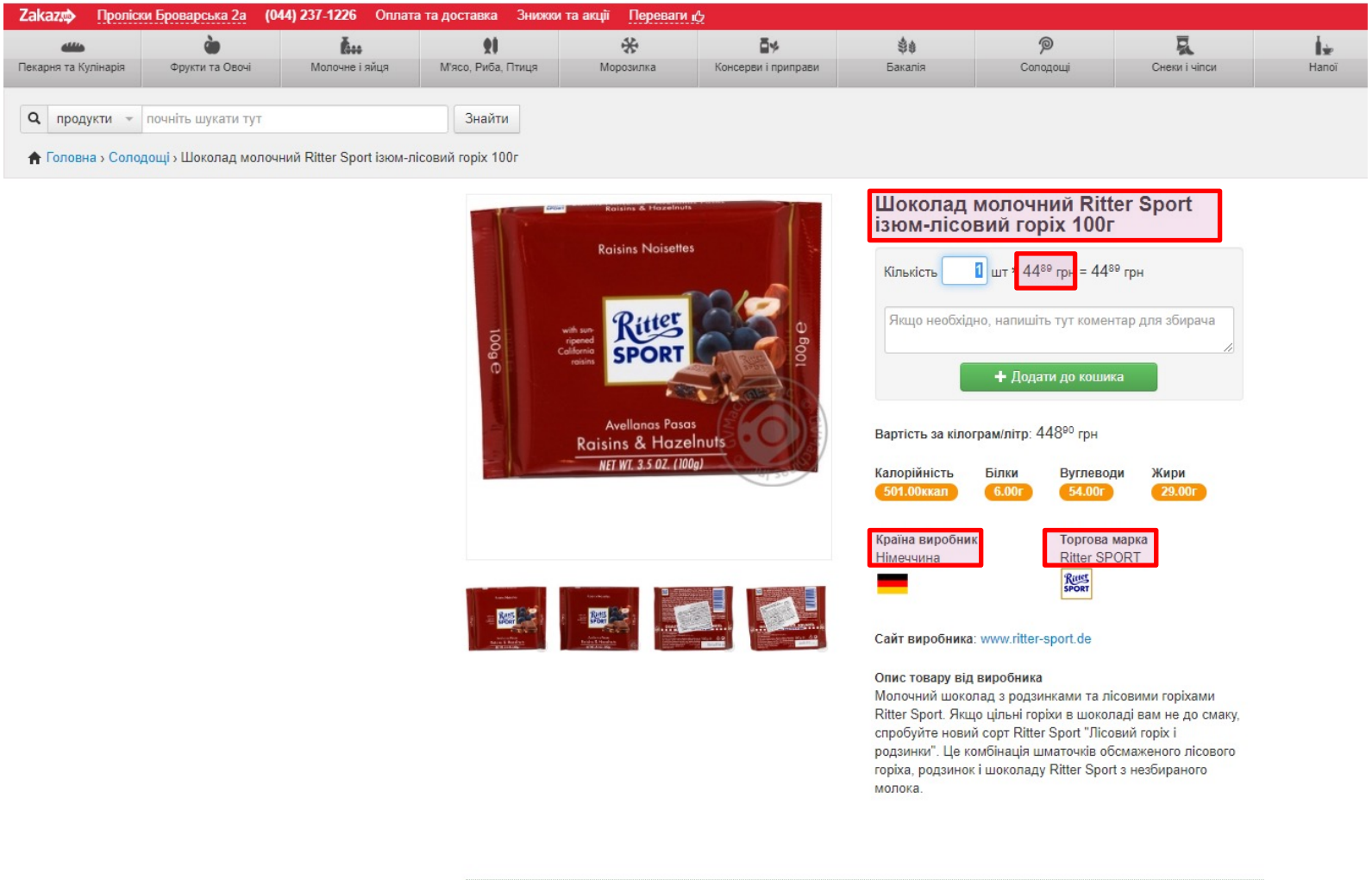
Процес наукасту інфляції за допомогою веб-скрепінгу в НБУ



Приклад


Zakaz
Проліски Броварська 2а
(044) 237-1226
Оплата та доставка
Знижки та акції
Переваги
Пекарня та Кулінарія
Фрукти та Овочі
Молочне і яйця
М'ясо, Риба, Птиця
Морозилка
Консерви і приправи
Бакалія
Солодощі
Снеки і чіпси
Напої
продукти
почніть шукати тут
Знайти
Головна › Солодощі › Шоколад молочний Ritter Sport ізюм-лісовий горіх 100г
Raisins Noisettes
Ritter SPORT
Avellanas Pasas
Raisins & Hazelnuts
NET WT. 3.5 OZ. (100g)
Шоколад молочний Ritter Sport ізюм-лісовий горіх 100г
Кількість 1 шт * 44⁸⁹ грн = 44⁸⁹ грн
Якщо необхідно, напишіть тут коментар для збирача
+ Додати до кошика
Вартість за кілограм/літр: 448⁹⁰ грн
Калорійність 501.00ккал
Білки 6.00г
Вуглеводи 54.00г
Жири 29.00г
Країна виробник
Німеччина
Торгова марка
Ritter SPORT
Сайт виробника: www.ritter-sport.de
Опис товару від виробника
Молочний шоколад з родзинками та лісовими горіхами Ritter Sport. Якщо цільні горіхи в шоколаді вам не до смаку, спробуйте новий сорт Ritter Sport "Лісовий горіх і родзинки". Це комбінація шматочків обсмаженого лісового горіха, родзинок і шоколаду Ritter Sport з незбираного молока.

Приклад


Zakaz
Проліски Броварська 2а
(044) 237-1226
Оплата та доставка
Знижки та акції
Переваги
Пекарня та Кулінарія
Фрукти та Овочі
Молочне і яйця
М'ясо, Риба, Птиця
Морозилка
Консерви і приправи
Бакалія
Солодощі
Снеки і чіпси
Напої
продукти
почніть шукати тут
Знайти
Головна › Солодощі › Шоколад молочний Ritter Sport ізюм-лісовий горіх 100г
Raisins Noisettes
Ritter SPORT
Avellanas Pasas
Raisins & Hazelnuts
NET WT. 3.5 OZ. (100g)
Шоколад молочний Ritter Sport ізюм-лісовий горіх 100г
Кількість 1 шт * 44⁸⁹ грн = 44⁸⁹ грн
Якщо необхідно, напишіть тут коментар для збирача
+ Додати до кошика
Вартість за кілограм/літр: 448⁹⁰ грн
Калорійність 501.00ккал
Білки 6.00г
Вуглеводи 54.00г
Жири 29.00г
Країна виробник Німеччина
Торгова марка Ritter SPORT
Сайт виробника: www.ritter-sport.de
Опис товару від виробника
Молочний шоколад з родзинками та лісовими горіхами Ritter Sport. Якщо цільні горіхи в шоколаді вам не до смаку, спробуйте новий сорт Ritter Sport "Лісовий горіх і родзинки". Це комбінація шматочків обсмаженого лісового горіха, родзинок і шоколаду Ritter Sport з незбираного молока.

Приклад

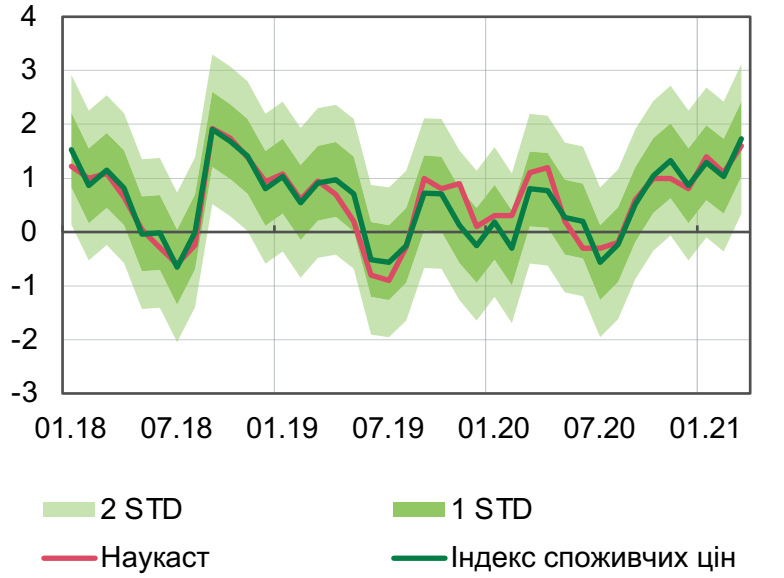
```
<form
    class="add-to-cart-js"
    data-id="04000417022004"
    data-name="Шоколад молочний Ritter Sport ізюм-лісовий горіх 100г"
    data-price="4489"
    data-quantity="1"
    method="POST"
><div style='display:none;'><input type='hidden' id='csrfmiddlewaretoken' name='csrfmiddlewaretoken' value='f972fbca557ba0fe88c8a7d7278848f7' /></div>
<div class="l-product-details-info-chunk well  product-details-info-ations">
    Кількість
    <span class="product-details-quantity">
        <div class="product-quantity-input">

            <script type="text/json-data" id="tier-prices">
            []
            </script>
            <input id="product-quantity"
                data="item" class="span1"
                type="text" name="amount"
                value="1"></span>


        <td>
            <div class="l-product-details-info-chunk product-details-info-tm">
                <h5>Країна виробник</h5>
                <div>Німеччина</div>

                    <div><img src="https://img2.zakaz.ua/upload.version_1.0.ebc7060c1f826bd704328763899cbd3c.32x32.png"></div>

            </div>
        </td>



        <td>
            <div class="l-product-details-info-chunk product-details-info-tm">
                <h5>Торгова марка</h5>
                <div><span itemprop="brand">Ritter SPORT</span></div>

                    <div><img src="https://img2.zakaz.ua/upload.version_1.0.473c328161636572192ab446f5283336.32x32.png"></div>

            </div>
        </td>
```

Data Structure

filename	good_name	zakaz_id	price	country	brand	avail	category	top_categ
P_bread_1_201	Батон Киевхлеб Киевский Нарезной (нарезка) 500г Украина	4820011192701	7.72	Украина	КИЇВХЛІБ	Yes	bread	bakery
P_bread_2_201	Батон Киевхлеб Киевский нарезной (целый) 500г Украина	4820011192718	7.38	Украина	КИЇВХЛІБ	Yes	bread	bakery
P_bread_3_201	Батон Кулиничи обычный нарезной половинка в/с 250г	4820174300289	4.35	N/A	КУЛИНИЧ	Yes	bread	bakery
P_bread_4_201	Батон Кулиничи нарезанный 500г	4820140110294	7.72	Украина	КУЛИНИЧ	Yes	bread	bakery
P_bread_5_201	Батон ТАК Отрубной низкокалорийный 240г Украина	4820041020227	8.28	Украина	Так	Yes	bread	bakery
P_bread_6_201	Батон Пшеничный ФР пекарни 370г	fozzy000000027139	4.55	N/A	N/A	Yes	bread	bakery
P_bread_7_201	Багет с отрубями ФР пекарни* 250г	fozzy000000027099	4.29	N/A	Фоззи	Yes	bread	bakery
P_bread_8_201	Багет Фоззи ФР пекарни 200г	fozzy000000027097	3.39	N/A	Фоззи	Yes	bread	bakery
P_bread_9_201	Багет гречневый 300г	fozzy000000027193	14.69	N/A	Фоззи	Yes	bread	bakery
P_bread_10_20	Багет Французский Люкс 300г	fozzy000000027487	10.45	N/A	Фоззи	Yes	bread	bakery
P_bread_11_20	Багет ХРУСteam со вкусом томатов и зелени 60г	4823063111843	6.32	N/A	Хрустим	Yes	bread	bakery
P_bread_12_20	Багет ХРУСteam со вкусом буженины и франц горчицы 60г	4823063111850	6.32	N/A	Хрустим	Yes	bread	bakery
P_bread_13_20	Багет ХРУСteam со вкусом королевского краба 60г	4823063111836	6.32	N/A	Хрустим	Yes	bread	bakery
P_bread_14_20	Багет Французская пекарня Пикантный 250г	fozzy000000027592	4.99	N/A	N/A	Yes	bread	bakery
P_bread_15_20	Багет бутербродный Флют ФР пекарни* 350г	fozzy000000027099	8.49	N/A	N/A	Yes	bread	bakery
P_bread_16_20	Хлеб Чиабат сендвич 100г	fozzy000000027019	4.49	N/A	Фоззи	Yes	bread	bakery
P_bread_17_20	Чиабатта Телятина 245г	fozzy000000027937	36	N/A	Фоззи	Yes	bread	bakery
P_bread_18_20	Чиабата с сыром ФР пекарни 150г	fozzy000000027139	8.45	N/A	Фоззи	Yes	bread	bakery
P_bread_19_20	Чиабатта ржаная 200г	fozzy000000027231	6.99	N/A	Фоззи	Yes	bread	bakery

Наукаст інфляції за допомогою веб-скрепінгу*

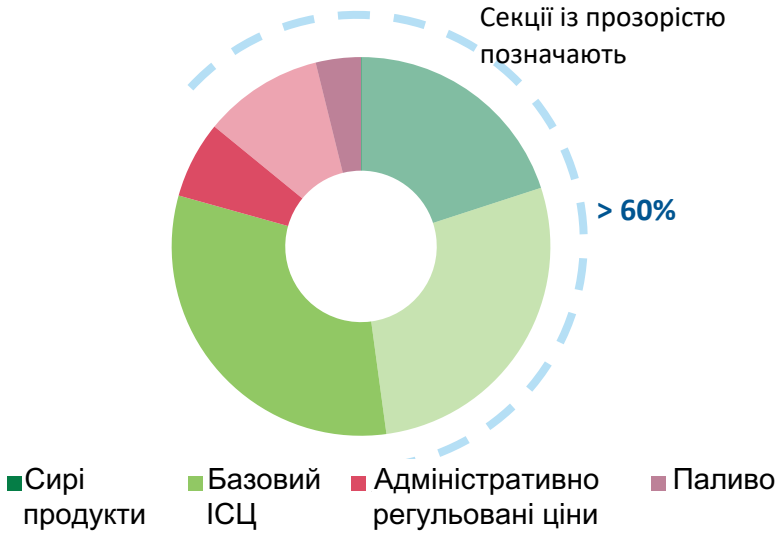
Порівняння наукасту та офіційних даних щодо ІСЦ, % м/м



RMSE	MAE
0.27	0.20

Джерело: ДССУ, НБУ.

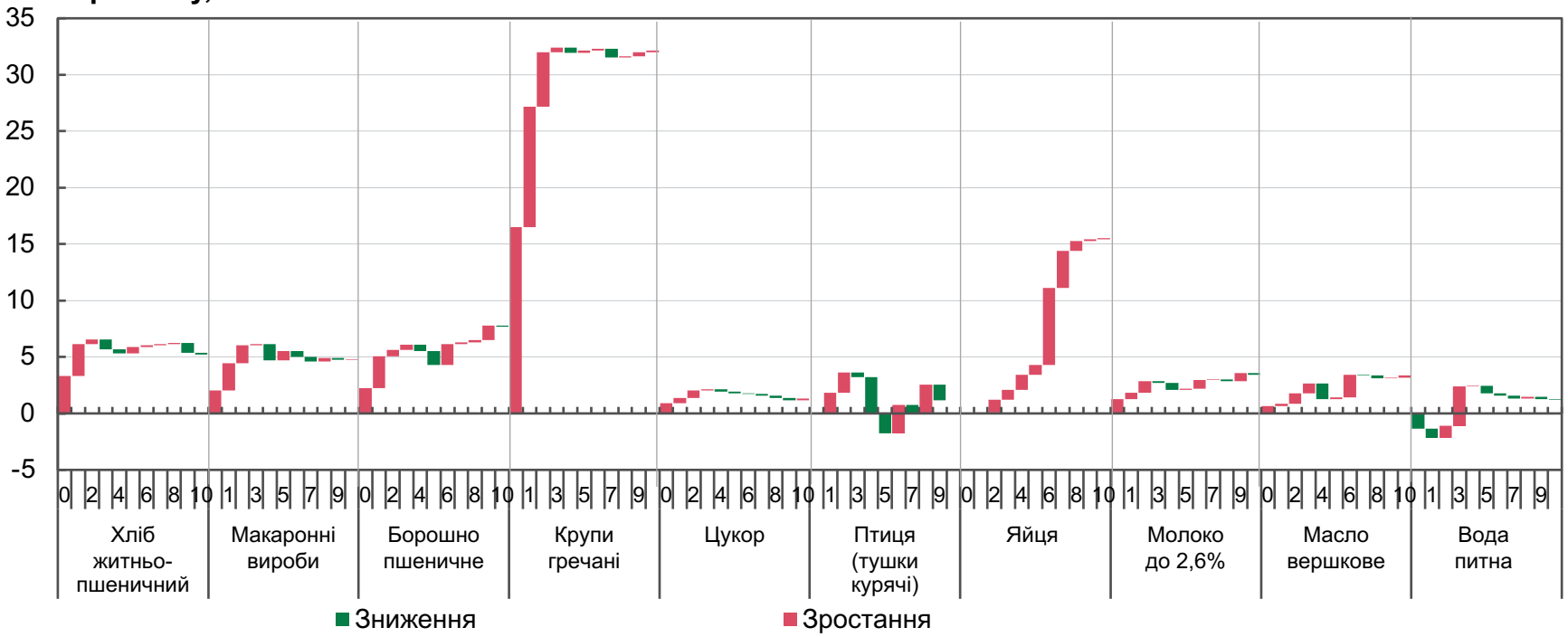
Покриття веб-скрепінгом споживчого набору для розрахунку ІСЦ



- Модель базується на фактичній оцінці зміни цін на товари та послуги за поточний місяць, а також доповнюється модельними розрахунками
- Розрахунок здійснюється до 25 числа звітного місяця, що дає змогу отримати інформацію щодо інфляційної динаміки на два тижні раніше офіційних значень
- Модель успішно експлуатується з кінця 2015 року

Запровадження карантинних обмежень призвело до значного коливання цін на товари та послуги

Потижнева динаміка (до попереднього тижня) цін на окремі товари з початку запровадження карантину, %*

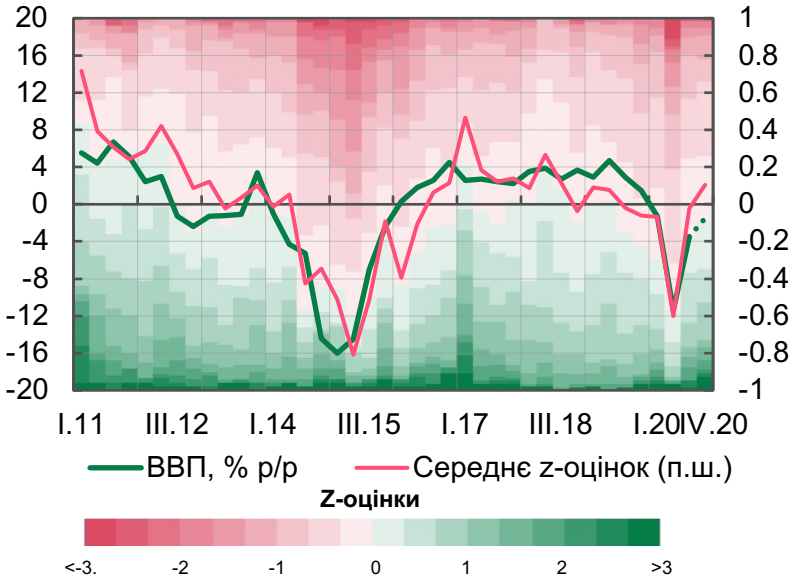


* Тиждень 0 – з 13 до 19 березня, цифри з 1 до 10 позначають наступні тижні.
Джерело: онлайн-супермаркети.

- Оперативний аналіз цінової ситуації за допомогою методу збору даних веб-скрепінг свідчить, що ціни на деякі продукти харчування та окремі товари особистого догляду дійсно доволі різко зросли на тлі ажіотажного попиту в другій половині березня
- Однак вже з квітня темпи зростання цін послабилися, а на окремі товари навіть знизилися

Теплова карта економічної активності (Z-score)

Теплова карта економічної активності України



Джерело: ДССУ, розрахунки НБУ.

Показники, використані для побудови теплової карти

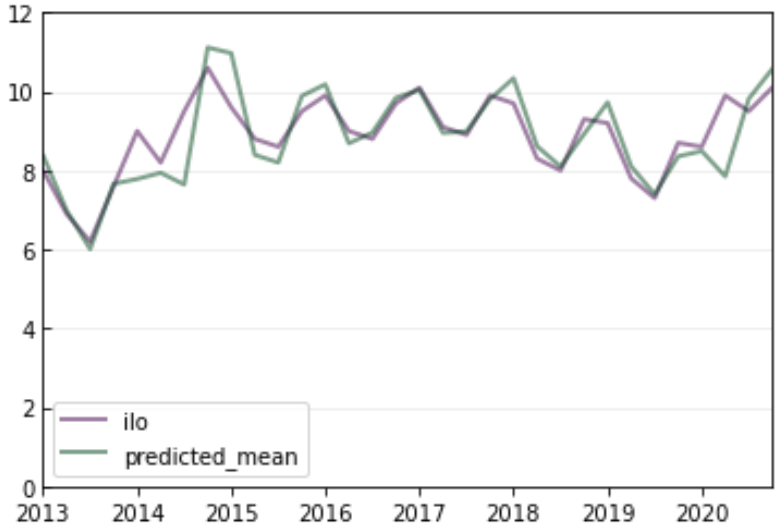
Галузь/показник	Кількість
Промисловість	33
Рослинництво, тваринництво	27
Будівництво	3
Операції з нерухомістю та землею	2
Пасажирооборот, вантажооборот	2
Комунальні послуги	1
Оборот роздрібної/оптової торгівлі	2
Перша реєстрація авто	1
Кредити ДГ та НФК	8
Платежі	1
Доходи/витрати депозитних корпорацій	4
Експорт/Імпорт	4
Надходження бюджету	10
Видатки бюджету	17
Кількість фізичних осіб-підприємців	3
Індекс споживчих настроїв населення	1
Усього	**119**

Джерело: ДССУ, розрахунки НБУ, ДПС, Укравтопром, Info Sapiens, Opendatabot, Міністерство юстиції.

- Для теплової карти економічної активності обрано показники економічної активності, які відображають структуру ВВП, з максимально доступною деталізацією, а також додано опосередковані показники. Усі показники, обрані для теплової карти публікуються раніше, ніж виходять дані щодо ВВП. Завдяки цьому зберігається прогнозна здатність цього підходу.
- Як кольорові індикатори теплової карти для економічної активності використано z-оцінки річних змін реальних показників діяльності
- Кореляція між реальною зміною ВВП та середньою z-оцінок у день закінчення кварталу перевищує 70%, а за наявності повного набору даних зростає до 80%
- Модель експлуатується з кінця 2020 року, специфікації опубліковані у вставці до Інфляційного звіту (Січень 2021 року)

Короткострокове прогнозування безробіття за допомогою Google Trends

Безробіття та прогноз за допомогою Google Trends



Характеристики моделі

1. Ключові слова:
Пошук роботи, безробіття, назви популярних сайтів для пошуку роботи, робота за кордоном

2. Агрегований індекс:
На основі індексів з найбільшою кореляцією

3. Специфікації SARIMAX:
(1, 1, 1)x(1, 1, 1, 4)

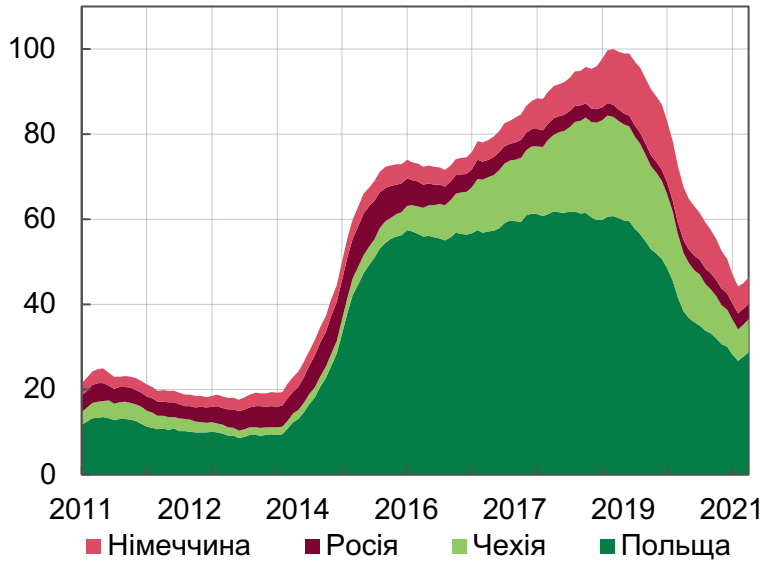
RMSE	MAE
1.14	0.44

Джерело: МОП, розрахунки НБУ.

- Базується на моделі SARIMAX (сезонна ARIMA з можливістю включення екзогенного фактору)
- Модель використовує тижневі дані запитів Google Trends, а також агрегований індекс цих запитів, конвертовані в квартальні та квартальні дані безробіття
- Прогнозування відбувається приблизно на квартал швидше, ніж публікуються офіційні дані
- Проводяться роботи з покращення результатів моделі, оскільки є під час кризових періодів збільшуються помилки

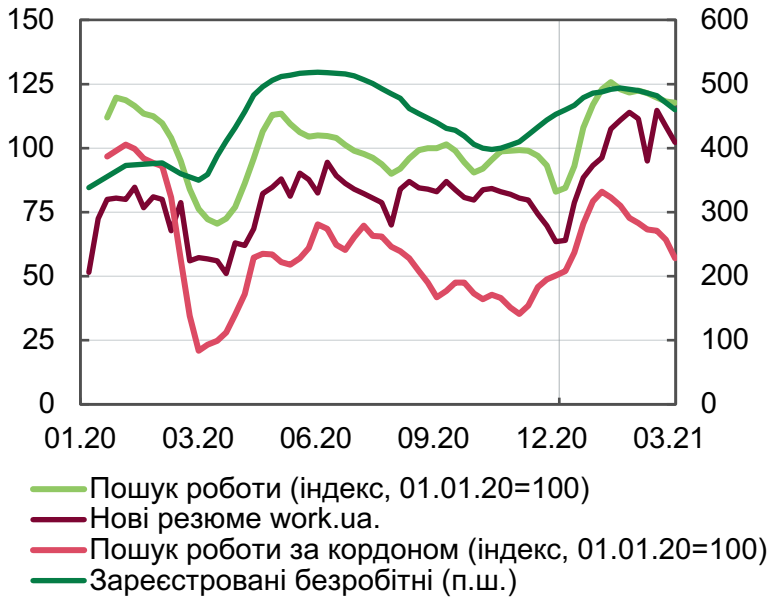
Дані ринку праці в режимі реального часу

Індекс міграційних настроїв



Джерело: Google Trends.

Пропозиція на ринку праці



Джерело: work.ua, Google Trends.

Перспективні проекти НБУ у сфері використання великих даних для макроекономічного аналізу

Цінові індекси

- Процеси ціноутворення
- Регіональні відмінності
- Ланцюг створення вартості
- Ціни на вітчизняні та імпортні товари

Показники реального сектору

- Прогнозування на основі даних платежів, а також даних податкової та митниці
- Використання новітніх методів прогнозування на основі машинного навчання

Текстовий аналіз

- Взаємозв'язок медіасередовища та макроекономічних показників
- Дослідження настроїв населення через тональність медіа простору
- Вивчення ефективності комунікації НБУ

Корисна література

- Baker, Scott & Bloom, Nicholas & Davis, Steven. (2016). Measuring Economic Policy Uncertainty. The Quarterly Journal of Economics. 131. qjw024. 10.1093/qje/qjw024.
- Cavallo, A., Online and official price indexes: Measuring Argentina's inflation. Journal of Monetary Economics (2012), http://dx.doi.org/10.1016/j.jmoneco.2012.10.002
- Krukovets, D. (2020). Data Science Opportunities at Central Banks: Overview. Visnyk of the National Bank of Ukraine, 249, 13-24. https://doi.org/10.26531/vnbu2020.249.02
- Faryna, O., Talavera, O., Yukhymenko, T. (2018). What drives the difference between online and official price indexes? Visnyk of the National Bank of Ukraine, 243, 21-32. https://doi.org/10.26531/vnbu2018.243.021
- Meltem Gulenay Chadwick & Gonul Sengul, 2012. "Nowcasting Unemployment Rate in Turkey : Let's Ask Google," Working Papers 1218, Research and Monetary Policy Department, Central Bank of the Republic of Turkey.
- Doerr, S and L Gambacorta (2020): "Identifying regions at risk with Google Trends: the impact of Covid-19 on US labour markets", *BIS Bulletin,* no 8, April.
- IFC big data survey (2020) https://www.bis.org/ifc/publ/ifc_report_13.pdf

National Bank
of Ukraine

Web ▪ Facebook ▪ Twitter ▪ Flickr ▪ Youtube ▪ Instagram