УДК 004.738.5:005
DOI: https://doi.org/10.37320/2415-3583/10.29

Струтинська І.В.
кандидат економічних наук,
Тернопільський національний технічний університет імені Івана Пулюя
ORCID: https://orcid.org/0000-0001-5667-6569

Дмитроца Л.П.
кандидат технічних наук,
Тернопільський національний технічний університет імені Івана Пулюя

Козбур Г.В.
старший викладач,
Тернопільський національний технічний університет імені Івана Пулюя

МЕТОДОЛОГІЯ ВИЗНАЧЕННЯ РІВНЯ ЦИФРОВОЇ ЗРІЛОСТІ БІЗНЕС-СТРУКТУР МЕТОДОМ КЛАСТЕРИЗАЦІЇ

Адаптація та трансформація бізнесу через цифрові технології є серйозною проблемою у вирішенні викликів світового ринку. Інформаційні технології дають змогу будь-якій компанії змінювати власну бізнес-модель, щоб диференціюватися від усього світового ринку. Враховуючи прогалини у статистичному забезпеченні моніторингу розвитку цифрової економіки та побудови інформаційного суспільства, доцільно активізувати роботу основних стейхолдерів щодо виконання «Плану заходів із реалізації Концепції розвитку цифрової економіки та суспільства України на 2018–2020 роки». Враховуючи актуальність даного питання, у статті проаналізовано методику збору важливих даних шляхом опитування та способу їх аналітики методом кластеризації респондентів. Розроблено інноваційну методику статистичного дослідження щодо цифрової трансформації бізнес-структур малого та середнього розмірів, а саме: запропоновано індикатори, здійснено опитування респондентів, підготовлено дані (запропоновано техніку їх очищення та подальшої обробки включно з кодуванням), здійснено кластеризацію суб'єктів дослідження та проведено аналіз відповідних результатів.

Ключові слова: *цифрова економіка, цифрові технології, кластеризація даних, опитування респондентів, статистичні методики, бізнес-структури МСП.*

Постановка проблеми. Сьогодні відбувається «цифровий перехід» від свого роду «аналогових» систем і процесів промислової економіки й інформаційного суспільства до «цифрової» економіки та «цифрового» суспільства. Вітчизняні бізнес-структури мають величезний потенціал у напрямі цифрової трансформації, компанії відкриті для всього нового, підприємці шукають нові можливості для бізнесу. Під цифровою трансформацією розуміють перетворення бізнесу (бізнес-стратегії або цифрової стратегії, моделей, операцій, продуктів, підходу до маркетингу, цілей тощо) через використання цифрових технологій.

Особливий інтерес представляють малі та середні підприємства. Проте ці бізнес-структури часто не володіють необхідною інформацією щодо застосування тих чи інших інноваційних цифрових технологій, що підвищують ефективність моделювання та ведення бізнесу [1; 2]. Адаптація до викликів ринку та побудова конкурентоспроможних бізнес-моделей підприємств такого типу потребують розроблення дорожніх карт цифрової трансформації. Є очікування певного стимулювання та напрацювання орієнтирів розвитку цифрового ринку з боку держави, у тому числі зазначення позиції малого та середнього підприємництва у процесі цифрового розвитку країни у часовій перспективі. Саме тому необхідно вдосконалити методику збору, статистичного дослідження, аналізу та візуалізації даних щодо використання ІКТ підприємствами в розрізі цифрового зростання економіки держави у цілому.

Аналіз останніх досліджень і публікацій. Проблему кластеризації респондентів за результатами відповідей на запитання опитувальників було розглянуто в [3]. Цю роботу присвячено наданню рекомендацій респондентам дослідження щодо результатів їхніх політичних уподобань та порівнянню методу кластеризації з іншими загальноприйнятими методиками надання таких рекомендацій. Також у роботі було розглянуто й інші популярні техніки обробки результатів опитувань подібного типу, такі як Party-Coding Similarity та Average Voter, та проведено порівняння їхніх результатів до запропонованої кластеризації. У роботі [4] досліджено проблематику кластеризації категорійних даних із використанням алгоритму GACUC із метрикою якості Category utility, заснованою на ймовірнісному підході до появи різних значень того чи іншого атрибуту. У роботах [5-9] висвітлено основні положення щодо кластеризації даних змішаного типу та використання вибраних у роботі алгоритмів та метрик. Деякі з методів та підходів до обробки результатів опитувань було розглянуто в [10], а саме: One-Way Tables, Cross-Tabulation, Higher-Way Tables,

Tabulation & the Assessment of Accuracy, а також вибір груп респондентів, валідацію отриманих результатів, формування профілів тощо.

Проте питання підготовки даних різних типів до кластеризації з метою їх глибинного опрацювання, що є надзвичайно важливим у специфіці опрацювання відповідних статистичних досліджень (анкет-опитувальників, фокус-груп та ін.), сьогодні не є вирішеним.

Мета статті полягає у розробленні інноваційної методології кластеризації бізнес-структур МСП щодо рівня їхньої цифрової зрілості за результатами опрацювання вибіркового опитування представників бізнес-структур малого та середнього розмірів, зареєстрованих у Тернопільській області, для подальшого обчислення індексу цифрової зрілості підприємства та вироблення рекомендацій стосовно його поліпшення.

Виклад основного матеріалу. У дослідженні кластеризації бізнес-структур за рівнем цифрової зрілості використовується набір даних, зібраних під час вибіркового анкетування підприємців Тернопільської області за допомогою опитувальника, створеного у сервісі Google Forms. Учасниками опитування стали підприємці – представники різних сфер діяльності, які зареєстровані в Тернопільській області. На момент проведення кластеризації набір даних містив відповіді 34-х підприємців.

Математичний опис даних. Дані являють собою набір N респондентів $U=\left\{\overrightarrow{u_1},\overrightarrow{u_2},\ldots,\overrightarrow{u_N}\right\}$ та M запитань $Q=\left\{q_1,q_2,\ldots,q_M\right\}$. Уважатимемо, що кожен учасник $\overrightarrow{u_i}\in U$ відповів на кожне із запитань $q_k\in Q$. Таким чином, було сформовано матрицю відповідей, де кожен опитаний учасник представлений у вигляді такого кортежу: $\overrightarrow{u_i}=\left\{u_{i1},u_{i2},\ldots,u_{ik},\ldots,u_{iM}\right\}$, де u_{ik} є відповіддю і-го учасника опитування на k-те запитання. Надалі такий кортеж називатимемо точкою.

Специфіка введених даних. Учасникам було запропоновано відповісти на 35 запитань, що стосуються двох аспектів функціонування бізнес-структур:

- даних про бізнес загалом (форма організації, кількість працівників, сфера діяльності, ведення імпорту чи експорту тощо);
- інформатизації бізнес-діяльності (використання соціальних мереж, SEO-оптимізації сайтів, аналітичних та рекламних продуктів, систем планування тощо).

Використано такі категорії запитань:

- множина варіантів (відповіді надавалися у формі вибору одного варіанту відповіді з декількох);
- множинна відповідь (із можливим вибором кількох відповідей із запропонованого переліку);
- відкриті запитання (у формі власноручного введення інформації).

Однак під час подальшої обробки відповіді, що є категорійними даними великого розміру, вважалися не важливими, адже вони містили особисте ставлення респондента з того чи іншого питання, а не сам факт використання чи невикористання тієї чи іншої технології у своєму бізнесі. Проте вони виявилися надзвичайно корисними під час глибшого аналізу та формування списку рекомендацій чи зауважень для конкретної ситуації.

Оскільки запитання опитувальника передбачали різні формати варіантів відповідей, елементами матриці відповідей стали як числові, так і категорійні дані, що потребували попередньої обробки перед запуском роботи алгоритму кластеризації. Під час первинного аналізу цих даних було виправлено відмінності у написанні різних варіантів відповідей, однакових за суттю, внесених за допомогою варіанту «Інше» в опитувальнику. Розмірність матриці було зменшено вручну за рахунок видалення запитань, відповіді на які однозначно не впливатимуть на результати кластеризації; також виправлено механічні помилки у введенні числових даних. Також, зважаючи на опцію обов'язковості для відповідей на усі запитання форми, не було виявлено пропущених значень. Власноруч уведені пропуски відповідей вважалися за дану відповідь «Ні» чи її еквівалент у тому чи іншому запитанні. Зважаючи на невелику кількість тестових даних, виконання такої очистки вручну було можливим.

Задача кластеризації даних із використанням мови Python. Задачу кластеризації відносять до широкого класу задач Data Mining, а саме до класу задач навчання без учителя. Мета кластеризації полягає у групуванні множини об'єктів на підмножини (кластери) так, щоб об'єкти одного кластеру були максимально схожі один на одного, тоді як об'єкти з різних кластерів – максимально несхожі. Вимірюють схожість елементів за певним критерієм.

Основним інструментом для технічного виконання задачі кластеризації бізнес-структур за рівнем інформаційної зрілості було вибрано Python – високорівневу мову програмування, яка набула найширшого використання для вирішення завдань, пов'язаних із машинним навчанням, штучним інтелектом та статистичними задачами [11]. Зручність застосування Python забезпечується доступністю низки безкоштовних бібліотек, які містять функції для розв'язання задач машинного навчання. Бібліотеки sklearn та scipy являють собою широкий набір інструментів для кластеризації та замірів показників якості проведеної кластеризації. У даному дослідженні було вибрано бібліотеку sklearn, що дає змогу вносити зміни в алгоритм обробки.

Підготовка даних. Для підготовки даних до кластеризації було використано кодування варіантів відповідей, які є категорійними даними

Деякі запитання опитувальника містили варіанти відповідей, непридатні для ранжування через неможливість порівняння їх із позиції «краще – гірше» або «більше – менше». Використання таких типів запитань потребує особливих підходів під час подальшої обробки. У зв'язку із цим, а також із малою кількістю тестових даних кодування непорядкових категорійних даних не проводилося, оскільки під час використання такого типу кодування нейтралізується будь-яка математична цінність даних, що унеможливлює коректну роботу вибраних алгоритмів. При цьому кортеж відповідей кожного окремого респондента складався б із числових даних, математичні операції над якими позбавлені сенсу.

На противагу відповідям порядкового категорійного типу було застосоване кодування. Після попередньої обробки отриманих даних кожному елементу вказаного типу з кожного кортежу було надано числове значення в інтервалі від нуля до одиниці: нуль присвоювався категорійній відповіді «Ні», інші значення ранжовано відповідно до збільшення рівня ствердності відповіді щодо успішності використання цифрових технологій у бізнес-структурі.

Агломеративна ієрархічна кластеризація. Агломерація є одним із підвидів ієрархічного підходу до кластеризації. Згідно з агломеративним підходом, кожна з окремих точок спочатку вважається окремим кластером. На кожному наступному кроці два найближчі кластери об'єднуються, в кінцевому підсумку утворюючи визначену кількість груп або зводячись до одного. Такий підхід дістав назву висхідного, або «знизу – вгору» [12].

Засоби бібліотеки sklearn. У бібліотеці sklearn алгомеративна кластеризація представлена функцією sklearn.cluster. Agglomerative Clustering приймає на вхід параметри:

- n_clusters – кількість кластерів, до якої потрібно звести алгоритм;
- affinity – метрика відстані між точками;
- linkage – метрика відстані між кластерами.

У бібліотеці sklearn [13] використовуються метрики Ward, Complete та Average для вимірювання відстаней між кластерами. Для обчислення відстаней між точками доступні метрики Euclidian, Manhattan, cosine, precomputed.

Усі, крім останньої із цих метрик, використовують математичні операції над числами – елементами векторів, що порівнюються. Використання precomputed метрики вимагає власноручного утворення матриці відстаней між елементами за будь-яким алгоритмом. Оскільки отриманий під час дослідження датасет містить у собі дані змішаних типів, то нами було вибрано використання precomputed-метрики за алгоритмом, наведеним нижче.

Метод Говера для обчислення відстані між елементами кластера. Одним із методів обчислення відстані між точками-кортежами у наборах даних змішаного типу є метрика відстані Говера (1), запропонована ще в 1971 р. [8; 9; 14]:

$$d\left(\overrightarrow{u_i},\overrightarrow{u_j}\right)=\frac{\sum_{k=1}^{M} w_{ijk} * d_{ijk}}{\sum_{k=1}^{M} w_{ijk}}, \qquad (1)$$

де w_{ijk} – вага окремо взятого компоненту – відповіді на запитання, d_{ijk} – відстань між двома векторами в окремо взятому k-му запитанні, M – кількість атрибутів (відповідей на запитання) у кортежі.

Нехай ваги всіх запитань позначено рівними одиниці: $w_{ijk}=1$. У такому разі формула набуде вигляду:

$$d\left(\overrightarrow{u_i},\overrightarrow{u_j}\right)=\frac{1}{M} * \sum_{k=1}^{M} d_{ijk}. \qquad (2)$$

Очевидно, що $d\left(\overrightarrow{u_i},\overrightarrow{u_j}\right)\in[0;1]$. Також варто зазначити, що вагомою перевагою є те, що кожен доданок відстані d_{ijk} обчислюється окремо та прямо залежить від типу даних у запитанні k. Це дає змогу стверджувати, що до кожного елементу кортежу відповідей буде застосований метод, який підходить саме для його типу, а відстань між двома окремими точками-векторами обчислюється як середня величина попарних відстаней для всіх ознак. Для числових даних відстань d_{ijk} виражається формулою:

$$d_{ijk}=\frac{\left|x_{ik}-x_{jk}\right|}{\max\left(x_k\right)-\min\left(x_k\right)}. \qquad (3)$$

Тобто відстань (3) дорівнює модулю різниці значень x_{ik} та x_{jk}, поділену на різницю між максимальним та мінімальним значеннями k-го елементу кортежу в усьому наборі даних.

Для категорійних даних із неможливістю впорядкування відстань обчислюється так:

$$d_{ijk}=\begin{cases}0, & x_{ik}=x_{jk},\\ 1, & x_{ik}\neq x_{jk}\end{cases}. \qquad (4)$$

Згідно з формулою (4), якщо категорійні значення рівні між собою, то відстань між ними дорівнює нулю, в іншому разі – одиниці.

Завдяки такому підходу алгоритм ураховує особливості різних типів змінних та дає змогу якомога ефективніше використати значення непорядкових категорійних змінних в обчисленні відстані між елементами кластера.

Обґрунтування оптимальної кількості кластерів. Оптимальна кількість кластерів в агломеративному методі знаходилася за допомогою покрокового обчислення показника якості кластеризації залежно від кількості кластерів. На рис. 1 представлено графік такої залежності, обчисленої для кількості кластерів від двох до десяти. Найвищу якість демонструє поділ лише на два кластери, проте це не дає змоги досягнути поставленої мети – оптимально

розбити вхідну множину на підмножини, придатні для подальшого аналізу. Із цієї причини оптимальною кількістю кластерів було вибрано п'ять кластерів як точку на графіку, після якої якість поділу даних на кластери стрімко погіршується.

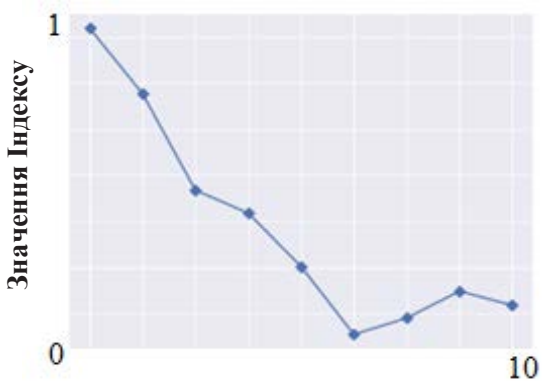

Рис. 1. Залежність показника якості кластеризації від кількості кластерів

Обчислення метрики якості кластеризації. Як метрику якості кластеризації даних у цьому дослідженні використано індекс Силуетта (5):

$$sil_i = \frac{b_i - a_i}{\max(a_i, b_i)}, \quad (5)$$

де a_i – середня відстань від *i*-го елемента кластера до всіх інших елементів того ж кластера; b_i – середня відстань від *i*-го елемента кластера до всіх інших елементів найближчого сусіднього кластера.

Очевидно, що $-1 \le sil \le 1$, при цьому значення, близькі до –1, свідчать про погану якість кластеризації, а близькі до 1 – про високу якість розбиття на кластери. Також варто зауважити, що індекс Силуетта має зміст тоді, коли кількість кластерів більша за 2, але менша за кількість точок у наборі даних.

Алгоритм було запущено на заданому наборі даних для його розбиття на кластери в інтервалі від двох до шести з метою практичного знаходження оптимальної кількості кластерів, за якої дані в отриманих підмножинах матимуть достатньо спільних рис та достатньо відрізнятимуться від інших кластерів.

Під час програмного виклику алгоритму агломеративної кластеризації вкажемо кількість кластерів *(m)*, метрику між точками та між кластерами і застосуємо метод fit_predict на матриці попередньо обчислених відстаней за метрикою Говера (*distances)*. На виході очікується отримання назв компаній, які потрапили у той чи інший кластер та список міток-кластерів для заданого переліку.

model = AgglomerativeClustering(n_clusters = m, affinity = «precomputed», linkage = «complete»)

labels = model.fit_predict(gower_distances)

У результаті роботи алгоритму отримано стійку кластеризацію із задовільною метрикою якості для кожного випадку: $sil > 0$. Візуалізацію такого розбиття зображено на рис. 2.

За первинного розбиття набору даних на два кластери отримано чіткий поділ на дві категорії: компанії, які не мають досвіду використання цифрових інструментів, та компанії, які мають такий досвід, що свідчить про коректну роботу алгоритму в цілому.

Під час утворення трьох кластерів категорія компаній, що вже мали досвід із цифровими технологіями, утворює дві нових. Відділяється кластер учасників, які застосовували платну допомогу спеціалістів у налаштуванні реклами, SEO-оптимізації сайту чи SMM-просуванні. Він залишається сталим до кінця тестування та налічує двох опитаних учасників із 34. Також виділяється категорія компаній, великий відсоток учасників якої використовував цифрові технологій самостійно і не задоволені результатами такої діяльності.

Розбиття на чотири кластери розділяє групу опитаних, які зазначили про відсутність досвіду із цифровими технологіями на компанії, що використовували лише ведення соціальних мереж, та компанії, які не мають ні сайту, ні сторінок у соціальних мережах, ні інших цифрових інструментів. Перша з груп залишається сталим кластером та налічує 10 компаній із 34.

Поділ на п'ять кластерів ініціює розбиття групи учасників, які самостійно ведуть кампанію з інформатизації бізнесу. Таким чином, явно виділяється статичний кластер-лідер – його єдиний учасник використовує аналітичні та рекламні інструменти, використовує спеціалізоване програмне забезпечення, веде соціальні мережі та має високі показники продажів через Інтернет. Інші учасники попереднього кластера формують групу компаній, що не повною мірою використовували цифрові інструменти.

Разом із тим утворення шести кластерів із початкового набору даних призвело до розбиття групи компаній, що не використовують цифрові технології, на два кластери за атрибутом приналежності до двох різних організаційних форм. Однак така деталізація не є доречною у даному дослідженні, тому з огляду на це кластеризацію було припинено.

Оптимальною кількістю кластерів було визначено п'ять, оскільки за такого розбиття з'явилася можливість виділення всіх необхідних категорій цифрової зрілості бізнес-структур: компанії, що не мали досвіду використання цифрових інструментів (16 учасників); компанії, що використовують лише соціальні мережі (10 учасників); компанії, що частково використовують SMM, SEO та аналітику (5 учасників); компанії, що використовують

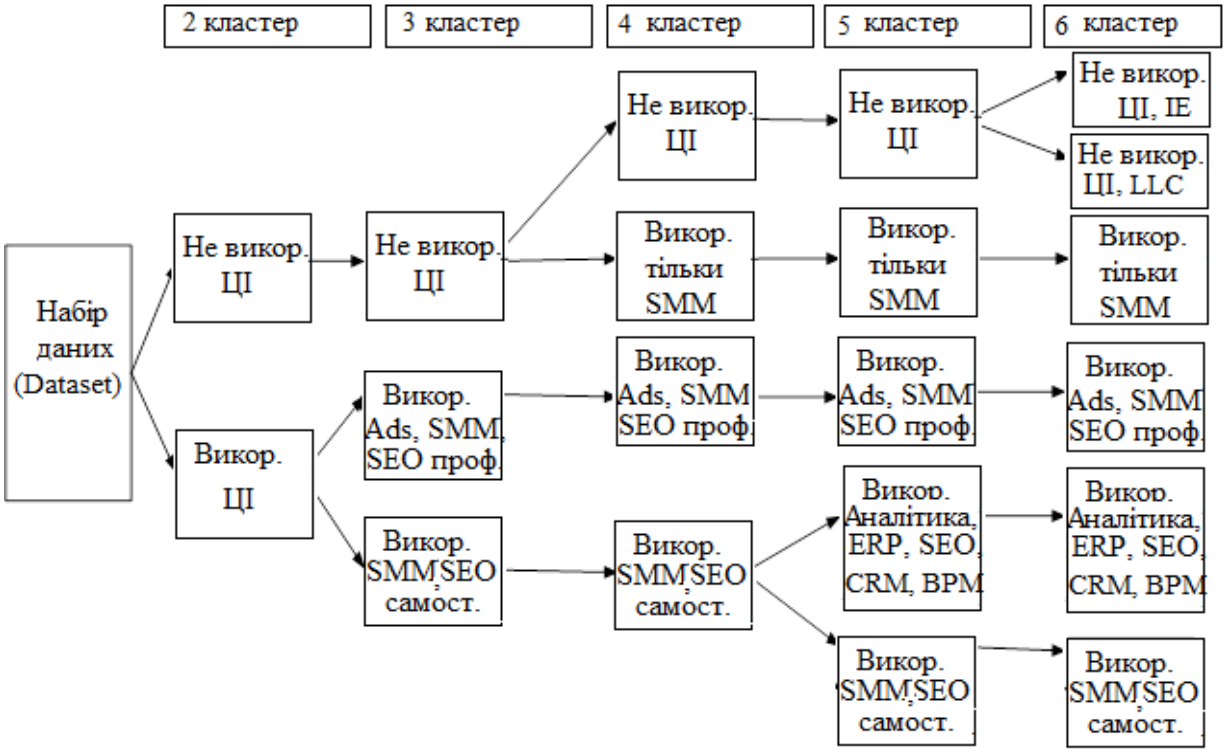

Рис. 2. Схематичне зображення етапів розбиття набору даних на кластери

послуги професіоналів (2 учасники), та компанії-лідери, які активно працюють з аналітикою та спеціалізованими додатками (1 учасник).

Таким чином, оптимальна кількість кластерів, вибрана теоретично, підтверджується аналізом практичних результатів. Метрика якості такої кластеризації sil = 0.19 є задовільною й явно залежить від кількості даних. Зауважимо, що за основними характеристиками (використання аналітичних інструментів, наявність сайту, його оптимізація, ведення сторінок у соцмережах) кожен з отриманих кластерів має достатню ступінь відмінності від інших кластерів та достатню подібність елементів усередині нього. Результати кластеризації задовольняють поставленим вимогам та дають змогу достатньою мірою окреслити окремі кластери та проблеми їхніх учасників.

Окрім розглянутої методики, також було опрацьовано й інші алгоритми кластеризації таких результатів (KMeans, EM-algorithm, GACUC, Agglomerative with Euclidian distance, Agglomerative with Gower distance and average linkage etc.), а також використано вищенаведені методи з непорядковими категорійними даними. Проте результати таких алгоритмів виявилися не завжди стійкими та менш придатними до подальшого аналізу. Порівняльний аналіз різних підходів та отриманих результатів вартує окремої публікації.

Висновки. У даному дослідженні нами було здійснено спробу виділити кластери для бізнес-структур Тернопільської області за ступенем цифрової зрілості. Інформацію щодо використання цифрових засобів та інструментів у моделюванні та веденні бізнесу було зібрано за допомогою вибіркового опитування представників підприємств, зареєстрованих у регіоні.

Результати цього наукового дослідження допоможуть глибше зрозуміти проблеми та реальний стан цифрової зрілості бізнес-структур на прикладі МСП Тернопільської області, розробити методику знаходження індексу цифрової зрілості окремого підприємства чи галузі економіки у цілому, а також розробити конкретні рекомендації (дорожні карти) для впровадження, які сприятимуть зростанню цифрової зрілості та подальшим трансформаціям на мікро- та макрорівні.

Список використаних джерел:

1. Про схвалення Концепції розвитку цифрової економіки та суспільства України на 2018–2020 роки. URL : https://www.kmu.gov.ua/ua/npas/pro-shvalennya-koncepciyi-rozvitku-cifrovoyi-ekonomiki-ta-suspilstva-ukrayini-na-20182020-roki-ta-zatverdzhennya-planu-zahodiv-shodo-yiyi-realizaciyi (дата звернення: 17.10.2019).

2. Коляденко С.В. Цифрова економіка: передумови та етапи становлення в Україні й у світі. *Економіка. Фінанси. Менеджмент*. 2016. № 6. С. 106–107.

3. Clustering Online Poll Data: Towards a Voting Assistance System / I. Katakis et al. *2012 Seventh International Workshop on Semantic and Social Media Adaptation and Personalization*. URL : http://www.katakis.eu/wp-content/uploads/2014/11/katakis_smap12.pdf. DOI : 10.1109/SMAP.2012.19 (дата звернення: 20.10.2019).

4. McCaffrey J. Machine Learning Using C#. Syncfusion, 2014. P. 148. URL : https://pt.b-ok.org/book/3097267/7356b0 (дата звернення: 17.10.2019).

5. Cluster analysis with balancing weights on mixed-type data / S.S. Chae et al. *The Korean communications in statistics*. 2002. № 13(3). DOI : org/10.5351/CKSS.2006.13.3.719.

6. Gower J.C. A comparison of some methods of cluster analysis. *Biometrics*. 1967. № 23. P. 623–637. DOI : 10.2307/2528417.

7. Rand W.M. Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association.* 1971.Vol. 66. № 336. P. 846–850. DOI : 10.2307/2284239.

8. Hoven J. van den. Clustering with optimised weights for Gower's metric. *University of Amsterdam*. 2015. P. 14–17. URL : https://www.math.vu.nl/~sbhulai/papers/thesis-vandenhoven.pdf (дата звернення: 17.10.2019).

9. Gower J.C. A General Coefficient of Similarity and Some of Its Properties. *Biometrics*. 1971. № 27(4). P. 859. DOI : 10.2307/2528823.

10. Statistical Services Centre, Approaches to the Analysis of Survey Data. *The University of Reading Statistical Services Centre Biometrics Advisory and Support Service to DFID*. 2001. March. URL : https://www.ilri.org/biometrics/TrainingResources/Documents/University%20of%20Reading/Guides/Guides%20on%20Analysis/ApprochAnalysis.pdf (дата звернення: 20.10.2019).

11. Welcome to Python.org. URL : https://www.python.org (дата звернення: 20.10.2019).

12. Malik U. Hierarchical Clustering with Python and Scikit-Learn. *Stack Abuse*. 2018. July. URL : https://stackabuse.com/hierarchical-clustering-with-python-and-scikit-learn/ (дата звернення: 20.10.2019).

13. Scikit. Clustering documentation. URL : *Scikit learn.* https://scikit-learn.org/stable/modules/clustering.html (дата звернення: 18.10.2019).

14. Filaire T. Clustering on mixed type data. *Medium*. 2018. July. URL : https://towardsdatascience.com/clustering-on-mixed-type-data-8bbd0a2569c3 (дата звернення: 18.10.2019).

References:

1. Plan of measures on implementation of the Conception of development of digital economy and society of Ukraine for the period from 2018 to 2020. URL: https://mtu.gov.ua/en/news/29453.html (access date October 17, 2019).

2. S.V. Koliadenko (2016) *Tsyfrova ekonomika: peredumovy ta etapy stanovlennia v Ukraini i u sviti* [Digital economy: preconditions and stages of formation in Ukraine and in the world], *Ekonomika. Finansy. Menedzhment*, no. 6, pp. 106–107, 2016. (in Ukrainian)

3. I. Katakis, N. Tsapatsoulis, C. Tziouvas and F. Mendes. (2012). Clustering Online Poll Data: Towards a Voting Assistance System, *2012 Seventh International Workshop on Semantic and Social Media Adaptation and Personalization.* URL: http://www.katakis.eu/wp-content/uploads/2014/11/katakis_smap12.pdf. doi: 10.1109/SMAP.2012.19 (access date October 20, 2019).

4. McCaffrey J. Machine Learning Using C#. Syncfusion, 2014, p. 148. URL: https://pt.b-ok.org/book/3097267/7356b0 (access date October 17, 2019).

5. Chae, S.S., Kim, J.-M. & Yang, W.Y., (2006). Cluster analysis with balancing weights on mixed-type data. The Korean communications in statistics, 13(3) doi.org/10.5351/CKSS.2006.13.3.719.

6. Gower J.C. (1967) A comparison of some methods of cluster analysis. Biometrics 23:623–637. doi: 10.2307/2528417.

7. William M. Rand (1971). Objective Criteria for the Evaluation of Clustering Methods. Journal of the American Statistical Association. Vol. 66, No. 336 (Dec., 1971), pp. 846–850. doi: 10.2307/2284239.

8. J. van den Hoven. (2015). Clustering with optimised weights for Gower's metric. University of Amsterdam. pp. 14–17. URL: https://www.math.vu.nl/~sbhulai/papers/thesis-vandenhoven.pdf (access date October 17, 2019).

9. Gower, J.C., (1971). A General Coefficient of Similarity and Some of Its Properties. *Biometrics*, 27(4), pp. 859. doi: 10.2307/2528823.

10. Statistical Services Centre, (2001, March) Approaches to the Analysis of Survey Data. *The University of Reading Statistical Services Centre Biometrics Advisory and Support Service to DFID. URL:* https://www.ilri.org/biometrics/TrainingResources/Documents/University%20of%20Reading/Guides/Guides%20on%20Analysis/ApprochAnalysis.pdf.

11. Welcome to Python.org. URL: https://www.python.org (access date October 20, 2019).

12. U. Malik. (2018 July). Hierarchical Clustering with Python and Scikit-Learn. *Stack Abuse*. URL: https://stackabuse.com/hierarchical-clustering-with-python-and-scikit-learn/ (access date October 20, 2019).

13. Scikit. Clustering documentation. *Scikit learn*. URL: https://scikit-learn.org/stable/modules/clustering.html (access date October 18, 2019).

14. T. Filaire. (2018, July). Clustering on mixed type data. *Medium.* URL: https://towardsdatascience.com/clustering-on-mixed-type-data-8bbd0a2569c3 (access date October 18, 2019).

Strutynska Iryna, Dmytrotsa Lesia, Kozbur Halyna
Ternopil Ivan Puluj National Technical University

METHODOLOGY OF DETERMINATION OF DIGITAL MATURITY LEVEL OF BUSINESS STRUCTURES BY CLUSTERING METHOD

Adaptation and transformation of business through digital are the major problem in solving the challenges of the global market. Information technologies allow any company to change its own business model to differentiate itself from the global market. Given the gaps in the statistical support for monitoring the development of the digital economy and building the information society, it is advisable to intensify the work of key stakeholders in the implementation of the Action Plan for the implementation of the Concept of the development of the digital economy and society of Ukraine for 2018-2020. The development of a system of indicators of digital business transformation, the provision of regular assessments of digital development and the introduction of regular, systematic statistical observations are particularly noteworthy. This, in turn, involves modifying existing statistical forms on the use of the Internet by the public and information and communication technology (ICT) at enterprises, and developing new indicators, methodological and organizational support for the collection and analysis of new data. Given the relevance of this issue, the article analyzes the method of collecting important data through a survey and the method of their analysis by the respondents' clustering method. An innovative statistical survey methodology was developed on the digital transformation of small and medium-sized business structures, namely: indicators were proposed, respondents were surveyed, data were prepared (techniques for their purification and further processing, including coding were proposed); respondents (business entities registered in the Ternopil oblast) were surveyed regarding their level of digital maturity; the data of the relevant surveys (studies) were elaborated by solving the problem of data analysis, namely, the clustering of the research subjects with the use of advanced information technologies of data analysis was carried out and the corresponding results were analyzed. The result of this research will be an in-depth understanding of the problems and the real state of use of digital by domestic business entities in their own business activities.

Key words: *digital economy, digital technologies, data clustering, survey of respondents, statistic techniques, business-structures of SMEs.*

JEL Classification: M21, G14.