УДК 630*5 http://dx.doi.org/10.31548/forest2019.02.051

ПРОГНОЗУВАННЯ ТАКСАЦІЙНИХ ПОКАЗНИКІВ ДЕРЕВОСТАНІВ НА ОСНОВІ *k*-NN МЕТОДУ

В. В. Миронюк, *кандидат сільськогосподарських наук, доцент, докторант кафедри таксації лісу та лісового менеджменту,*
А. М. Білоус, *доктор сільськогосподарських наук, старший науковий співробітник, завідувач кафедри таксації лісу та лісового менеджменту,*
П. П. Дячук, *аспірант кафедри таксації лісу та лісового менеджменту,**
Національний університет біоресурсів і природокористування України
E-mail: victor.myroniuk@nubip.edu.ua

Дані дистанційного зондування Землі (ДЗЗ) є важливим джерелом додаткової інформації про параметри лісових насаджень, яку використовують під час інвентаризації лісів. Серед відомих технік, які дають змогу поєднати інформацію, одержану наземними та дистанційними методами, дуже поширеним став непараметричний метод відновлення пропущених значень – метод k найближчих сусідів (k-NN). Із метою дослідження ефективності цього методу за різної комбінації вхідних параметрів залучено матеріали експериментальної інвентаризації лісів на 141 вибіркових одиницях, де реалізовано принципи відбору дерев з імовірністю, пропорційною їхній кількості та розмірам. Дослідна територія розташована в центральній частині Київської області та має площу близько 56 км^2. Як джерело додаткової інформації використано три типи даних супутникової зйомки: 1) сезонні мозаїки супутникових знімків Landsat 8 OLI; 2) окрему сцену знімка SPOT 7; 3) часову серію спостережень PlanetScope. Ми встановили, що точність прогнозування сум площ поперечних перерізів дерев у насадженні для всіх типів даних ДЗЗ є вищою для реласкопічного методу наземної таксації деревостанів. Серед залучених спектральних даних менші помилки одержано для щільних часових серій супутникових спостережень Landsat та PlanetScope. При цьому виявлено, що значення темпорального розрізнення супутникових знімків для прогнозування сум площ перерізів дерев на 1 га є більшим, аніж їхнє просторове розрізнення. У роботі встановлено, що імплементація k-NN методу за k = 1, забезпечує збереження коваріації між змінними, тоді як збільшення величини k звужує діапазон мінливості таксаційних показників, особливо це стосується малих вибірок. Серед методів пошуку найближчих сусідів більшої точності досягнуто при застосуванні алгоритму Random Forest.

Ключові слова: *вибіркова інвентаризація лісу, кругові пробні площі, реласкопічна таксація, класифікація.*

* Науковий керівник – доктор сільськогосподарських наук, старший науковий співробітник А. М. Білоус.

Актуальність. Упровадження дистанційного збору даних у практику лісової інвентаризації є однією з основних сучасних тенденцій розвитку лісооблікових методів. Дистанційні технології дали змогу зменшити витрати на проведення польових робіт і додали статистичним оцінкам просторову складову у вигляді карт, що відображають видовий склад лісів і розподіл окремих біофізичних параметрів деревостанів (сума площ перерізів, запас і біомаса). На цій основі базуються сучасні підходи стосовно оцінювання показників лісового фонду з використанням так званих модельних підходів, які вже достатньо апробовані в багатьох країнах на практиці. Тому актуальним є дослідження різноманітних технік прогнозування таксаційних показників на основі даних супутникової зйомки, серед яких добре відомим став метод найближчих сусідів – *k*-Nearest Neighbors (*k*-NN).

Аналіз останніх досліджень та публікацій. Непараметричний *k*-NN метод є інтуїтивно зрозумілою технікою прогнозування значень однієї або відразу кількох змінних, який виконується на основі подібності елементів у коваріаційному просторі ознак. Завдяки своїй простоті *k*-NN метод набув великого поширення в різних завданнях лісової інвентаризації: від відновлення пропущених значень у базах даних до картографування показників лісових насаджень і статистичного оцінювання параметрів лісового фонду (McRoberts, 2012). Поєднуючи результати таксації пробних площ, дані дистанційного зондування Землі (ДЗЗ), а також іншу допоміжну інформацію, цей метод дає змогу доповнити недостатню кількість польових даних та істотно підвищити точність оцінки показників лісового фонду не лише на рівні генеральної сукупності, а й на детальнішому рівні.

Техніка прогнозування таксаційних показників на основі пошуку найближчих сусідів широко апробована в лісовій інвентаризації. Найчастіше її використовують для прогнозування неперервних показників насаджень: запасу деревостану в цілому та за складовими елементами лісу (McRoberts, 2009a; Reese, Nilsson, Sandström, & Olsson, 2002), запасу ділової деревини (Bernier et al., 2010), компонентів наземної біомаси (Bilous et al., 2017; Latifi et al., 2015), сум площ перерізів дерев на 1 га та густоти деревостанів (Franco-Lopez, Ek, & Bauer, 2001; Hudak, Crookston, Evans, Hall, & Falkowski, 2008), середніх діаметра, висоти, віку та зімкнутості насаджень (McRoberts, Nelson, & Wendt, 2002; Mozgeris, 2008), верхньої висоти деревостанів (Packalén, Temesgen, & Maltamo, 2012), лісистості території (McRoberts, Liknes, & Domke, 2014). Досліджень, які стосуються дискретних показників, значно менше; це переважно методи картографування типів земного покриву або специфічних для деяких країн характеристик лісових ділянок (Franco-Lopez et al., 2001; Haapanen, Ek, Bauer, & Finley, 2004), груп лісових насаджень залежно від їхнього видового складу (Beaudoin et al., 2014).

Останнім часом у літературі досить часто розглядають питання впливу конфігурації пробних ділянок на точність інвентаризації лісів за *k*-NN методу. Зазвичай для цього застосовують вибіркові одиниці фіксованої площі. Дослідження фінських учених вказують, що точність дешифрування окремих параметрів

лісових насаджень на основі кругових пробних площ постійного радіуса та комбінованих ділянок для даних лазерного сканування є приблизно однаковою (Maltamo, Korhonen, Packalen, Mehtatalo, & Suvanto, 2007). Згідно з останніми публікаціями (Tomppo, Kuusinen, Mäkisara, Katila, & McRoberts, 2017), перспективними для оцінки таксаційних показників дистанційними методами можуть бути кілька конфігурацій ділянок: кругові пробні площі радіусом 9 м; концентричні проби з поєднанням зовнішнього та внутрішнього радіусів відповідно 9,0 і 5,64 м; комбіновані ділянки з максимальним радіусом проби 9 м та реласкопічним коефіцієнтом BAF = 1. Технологія реласкопічної таксації, адаптована до відбору відповідних сигналів (пікселів) лазерного сканування, стала предметом нещодавніх досліджень німецьких учених (Kirchhoefer, Schumacher, Adler, & Kändler, 2017). У результаті опрацьовано методику, яка дає змогу узгодити сигнали лазерного сканування з відповідним деревами, які були обліковані на пробі.

Отже, ефективне використання даних наземних і дистанційних вимірювань зумовлює необхідність розгляду додаткових питань стосовно технології статистичної інвентаризації лісів. Оптимальні параметри вибіркових одиниць під час оцінки таксаційних показників лісових насаджень за цими методами можуть відрізнятися. Зокрема, зазначене стосується розміру та конфігурації пробних ділянок, адже наземна вибіркова одиниця має адекватно характеризувати відповідну ділянку супутникового знімка.

Мета дослідження полягає у діагностиці *k*-NN методу прогнозування таксаційних показників лісових насаджень залежно від конфігурації пробних ділянок і різних типів супутникової інформації.

Матеріали і методи дослідження. Дослідження базується на дослідному матеріалі, зібраному впродовж 2017–2018 рр. на території ВП НУБіП України «Боярська лісова дослідна станція». Експериментальний полігон загальною площею близько 56 км^2 розташований у Київській області між значеннями довготи 30° 00’ Сх та 30° 12’ Сх і широти 50° 12’ Пн та 50° 20’ Пн. Відповідно до даних базового лісовпорядкування підприємства 2017 р. територія на 92 % представлена лісовими насадженнями з переважанням сосни звичайної (*Pinus sylvestris* L.) – 92 %, дуба звичайного (*Querqus robur* L.) – 6 %. Решту вкритих лісовою рослинністю ділянок займають деревостани м’яколистяних деревних видів, передусім вільхи клейкої (*Alnus glutinosa* (L.) Gaerth) та берези повислої (*Betula pendula* Roth.). Як другорядні до складу лісових насаджень входять такі деревні види, як осика (*Populus tremula* L.), дуб червоний (*Querqus rubra* L.), робінія псевдоакація (*Robinia pseudoacacia* L.) та липа серцелиста (*Tilia cordata* Mill.). У віковому відношенні на території переважають лісові культури VI–IX класів віку, хоча частка насаджень віком понад 100 років також достатньо значна і становить близько 20 %. Максимальний вік природних деревостанів досягає 200 років. Завдяки сприятливим умовам продуктивність сосняків у межах експериментального полігону характеризується переважно 1^a і 1^b класами бонітету.

Територіальну основу вибірки утворює регулярна мережа вибіркових одиниць, розташованих серед вкритих

лісовою рослинністю ділянок з інтервалом 1' за широтою та довготою. Під час збору дослідних даних практикувалося два принципи таксації лісу – відбір, пропорційний кількості дерев (кругові проби фіксованої площі) та розмірам дерев (реласкопічні перелічувальні ділянки). Кругові ділянки радіусом 12,62 м закладалися з використанням програмного комплексу Field-Map за методикою, що передбачає картографування та обмір облікових дерев. Кожне з них розглядалося як точка, спроектована в систему координат проекцій EPSG:32636. Одночасно з цим виконували облік дерев на реласкопічних площадках, використовуючи спільний центр і реласкопічний коефіцієнт BAF = 1. Для всіх облікових дерев вимірювали діаметри дерев на висоті 1,3 м. У модельних дерев (приблизно ¼ від загальної кількості) додатково вимірювали висоту стовбура. В камеральних умовах на основі координат облікових дерев були обчислені відстані до кожного з них від центра ділянки. Це дало змогу розглядати дві додаткові схеми відбору (табл. 1) – концентричні пробні площі та комбіновані ділянки (поєднання кругової та реласкопічної вибірки). Результати таксації деревостанів одержано на 141 пробах кожної конфігурації.

Як джерело дистанційних даних використано супутникові знімки відразу трьох сенсорів: Landsat 8 OLI, SPOT 7 і PlanetScope. За своїм характером це достатньо різнопланові дані. Знімки Landsat являють собою часову серію супутникових спостережень із просторовим розрізненням 30 × 30 м, одержану протягом 2014–2016 рр. Їх поєднано за принципом відбору «найбільш зеленого» пікселя у три композитні мозаїки для таких сезонів: квітень–жовтень, літо та осінь (Myroniuk, 2018). У дослідженні розглянуто 36 змінних, доцільність використання яких було обґрунтовано на етапі опрацювання класифікаційної моделі Random Forest (Myroniuk, 2017). Класичний варіант дистанційного зондування представлений супутниковим знімком SPOT 7, який одержано 9 серпня 2015 р. В аналізі використано три наявні спектральні канали видимого та один ближнього інфрачервоного діапазонів із розміром пікселя 6 × 6 м, які приведено до значень відбиття на сенсорі. Геометричну корекцію знімка виконано з використанням значень RPC коефіцієнтів, що постачалися зі знімком. Крім цього, на території експериментального полігону створено часову серію спостережень з ортотрансформованих знімків PlanetScope із просторовим розрізненням 3,125 м рівня обробки Analytic (Planet Team, 2017). Два з них відображають стан території у весняний період – 21 квітня та 5 травня 2018 р., а ще два – літній (21 червня та 26 серпня 2018 р.). Таким чином досліджували часову серію спостережень у блакитному, зеленому, червоному та ближньому інфрачервоному діапазонах (усього 16 спектральних каналів).

Відповідно до термінології *k*-NN методу, багатовимірний вектор залежних змінних позначають як *Y*. Його утворюють *n* спостережень вибірки з кінцевої популяції обсягом *N*. Вектор допоміжних змінних, представлений спостереженнями для всіх елементів популяції, позначають як *X*. Ці змінні утворюють *n*-вимірний простір – простір ознак. Вибіркові одиниці, для яких доступні як значення відгуку (відгуків), так і незалежних змінних, формують опорний набір даних. Відповідно, цільовий

набір утворюють елементи популяції, для яких необхідно виконати прогноз відгуку (відгуків). Для неперервних показників значення залежної змінної для *i*-го цільового елемента популяції розраховують за формулою (1):

$$\tilde{y}_i = \sum_{j=1}^{k} w_{ij} y_j^i, \qquad (1)$$

де y_j^i – набір значень досліджуваних змінних для елементів опорного набору даних, які є найближчими або найбільш схожими до *i*-го елемента цільового набору в просторі ознак із урахуванням відповідної міри відстані d_{ij} $(j = 1, 2, 3 \ldots, k)$; w_{ij} – ваговий коефіцієнт, який призначають *j*-му найближчому сусіду з урахуванням умови, що .

Із наведеної формули випливає, що застосування *k*-NN методу потребує розгляду трьох питань: 1) вибору міри відстані; 2) визначення схеми зважування відстаней; 3) встановлення кількості сусідів . Утім, першочергове значення для цього методу має технологія збору даних, яку характеризують розмір та форма первинних одиниць вибірки, просторове розрізнення супутникових знімків. Зазначені параметри формують передумови прогнозування значень таксаційних показників за даними ДЗЗ.

Найпростішою та інтуїтивно зрозумілою мірою відстані є Евклідова (EUC) відстань. Оптимізацію розрахунків при цьому виконують тільки вибором доречних змінних, який може бути покроковим (Chirici, McRoberts, Fattorini, Mura, & Marchetti, 2016) або здійснюватися за допомогою генетичного алгоритму (Tomppo & Halme, 2004). Під час обробки даних за *k*-NN методом можливі й інші міри відстаней між найближчими «сусідами», які обчислюють за результатами канонічного кореляційного аналізу (MSN), канонічного кореспонденс-аналізу (GNN), компонентного аналізу (ICA). Усі вони реалізовані у спеціальному статистичному пакеті `yaImpute` для системи R (Crookston & Finley, 2008). Серед можливих варіантів пошуку найближчих сусідів розробники пакета додатково реалізували один з алгоритмів машинного навчання – Random Forest (RF), якому приділяють велику увагу останнім часом (Breiman, 2001).

Більшість публікацій свідчить, що основним критерієм діагностики *k*-NN методу має бути запас деревостанів або сума площ перерізів дерев на 1 га (McRoberts, 2009b; Tomppo et al., 2017). З практичного погляду в основу обґрунтування технології дистанційної оцінки таксаційних показників лісових насаджень доречніше покласти суму площ перерізів. Це пояснюється кількома факторами: тісним взаємозв'язком зазначених таксаційних показників; сума площ поперечних перерізів обчислюється безпосередньо, тому оцінка помилок є простішою і не залежить від методу визначення запасу. Отже, в дослідженні розглядали вектор Y_j, утворений значеннями сум площ перерізів дерев на 1 га, та вектор X_j із відповідними спостереженнями для і-го пікселя у просторі спектральних ознак *j*-го сенсора. Для того щоб досягти узгодження між розміром пікселя Landsat 8 OLI та інших сенсорів, для кожного центрального пікселя показники відбиття усереднювалися у вікні розміром 9×9 пікселів – для сенсора PlanetScope та 5×5 пікселів – SPOT 7.

Результати дослідження та їх обговорення. Методика польових досліджень дала змогу розглядати

чотири типи пробних ділянок: фіксованої площі (F 1262–8, N 564–12), реласкопічних (AC-1) та комбінованих (TA 1262–1) проб (табл. 1).

Дослідження виконували у статистичному пакеті `yaImpute` для системи R (Crookston & Finley, 2008)». Вплив конфігурації пробних площ на величину середнього квадратичного відхилення між фактичними та прогнозними значеннями таксаційних показників (RMSD) аналізували за базової імплементації *k*-NN методу, яка передбачає використання Евклідової відстані та $k = 1$ (табл. 2). Розрахунок здійснювали за формулою (2):

$$RMSD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2} \quad (2)$$

де $(y_i - \tilde{y}_i)^2$ – відхилення між змодельованими та встановленими на пробі значеннями відгуку; n – обсяг вибірки.

Точність прогнозування сум площ поперечних перерізів на основі всіх типів даних ДЗЗ виявилася найвищою для реласкопічного методу таксації. На нашу думку, цей факт можна пояснити методикою відбору дерев повнотоміром: до вибірки з більшою ймовірністю потрапляють дерева більших розмірів, вони визначають параметри зображення деревостану «зверху», а отже, дані на реласкопічних пробах тісніше корелюють із спектральними показниками супутникових знімків. Таку особливість, а саме, залежність мінливості оцінок показників лісових насаджень від того, наскільки тісно вони пов'язані з параметрами верхнього ярусу насаджень, виявлено в дослідженнях Ohmann, Gregory & Roberts (2014). Під час визначення сум площ за даними різних сенсорів спостерігається істотна перевага часових серій супутникових спостережень. Найбільші помилки за всіх конфігурацій пробних ділянок одержані за окремим знімком SPOT 7, тоді як найбільшу точність забезпечила обробка часової серії знімків PlanetScope.

Вибір величини k є суб'єктивним рішенням, яке приймають на основі багатьох факторів. Коректніше у цьому відношенні спрацьовують малі значення k, які дають змогу зберегти коваріацію між різними показниками вибірки та перенести її на цільовий набір даних (Hudak et al., 2008; McRoberts et al., 2002). Через те, що *k*-NN метод з однаковим успіхом можна використовувати як для прогнозування значень окремого параметра, так і набору показників, з'являються два аспекти цього питання:

1. Параметри конфігурацій пробних ділянок

Шифр конфігурації вибіркових ділянок*	Параметри вибіркових одиниць				
	основна проба			підпроба	
	радіус проби, м	діаметр облікових дерев, см	BAF	радіус підпроби, м	діаметр облікових дерев, см
F 1262–8	12,62	≥ 8,0	–	–	–
N 564–12	12,62	≥ 12,0	–	5,64	6,0–11,9
TA 1262–1	12,62	≥ 8,0	1	–	–
AC-1	–	≥ 8,0	1	–	–

* F – кругові пробні площі постійного радіуса; N – концентричні проби; TA – комбіновані ділянки; AC – реласкопічні пробні площі.

2. Величина оцінки сум площ перерізів дерев на 1 га за різних конфігурацій пробних ділянок і даних ДЗЗ, м²·га⁻¹

Сенсор	Шифр конфігурації пробних ділянок			
	F 1262–8	N 564–12	TA 1262–1	AC-1
PlanetScope	11,6	11,6	11,7	9,2
SPOT 7	11,2	11,2	11,4	10,4
Landsat 8 OLI	10,1	10,0	10,2	9,7

збереження просторової семіваріації однієї змінної або коваріації між різними показниками. У першому випадку найінформативнішим є побудова емпіричної семіваріограми зі встановленим просторовим лагом, яку необхідно порівняти з аналогічною, побудованою на підставі прогнозних значень (McRoberts, 2009b). Якщо відразу прогнозують кілька показників, доречнішим стане аналіз коваріації (3) між парами різних змінних:

$$\mathrm{Cov}(y_p, y_q) = \frac{1}{n}\sum_{i=1}^{n}(y_{pi} - \bar{y}_p)(y_{qi} - \bar{y}_q), \quad (3)$$

де p, q – позначення векторів двох змінних; $\bar{y}_p$, $\bar{y}_q$ – відповідно середні арифметичні значення спостережень для p-ї та q-ї змінної; – обсяг вибірки.

Якість прогнозу оцінюють на основі відношення коваріації між фактичними та модельними значеннями показників. Спочатку k-NN метод реалізується за k = 1 із залученням різних сусідів і-го порядку, наприклад, i=1,5,10...n. Якщо на якомусь з етапів відношення коваріації істотно змінилося щодо k = 1, слід очікувати значного порушення коваріаційної структури між змодельованими показниками.

Для дослідження впливу різної кількості найближчих сусідів на якість прогнозу до значень сум площ перерізів дерев, встановлених на реласкопічних пробних (AC-1) ділянках, додано відповідні оцінки їхньої густоти. На основі цих показників обчислено середній діаметр деревостанів (4), який за даними натурної таксації 141 пробних ділянок становить 33,2 см:

$$D = 200 \cdot \sqrt{\frac{G}{\pi \cdot N}}, \quad (4)$$

де G – сума площ перерізів дерев на 1 га, м²; – кількість дерев на 1 га, шт.

Отже, поряд з аналізом відношення коваріацій змінних за різної кількості найближчих сусідів, порівнювалося середнє значення діаметра деревостанів (табл. 3).

У нашому випадку використано стандартне зважування відстаней, яке базується на рівнянні (5) при t=1:

$$w_{ij} = \frac{d_{ij}^{-t}}{W_i} = \frac{d_{ij}^{-t}}{\sum_{j=1}^{k} d_{ij}^{-t}}, \quad (5)$$

де k – кількість найближчих сусідів; t – дійсне число більше ніж 0 (t=2).

Порівняння відношень коваріації між сумами площ перерізів і кількістю дерев на 1 га свідчить, що коваріація зберігається для k-NN методу тільки за k =1. Це має важливе значення за одночасного прогнозування кількох змінних, адже такий варіант методу унеможливлює появу нетипових співвідношень між значеннями таксаційних показників. Зазначене стосується показників, які обчислю-

3. Відношення коваріацій між фактичними і змодельованими значеннями таксаційних показників та оцінки середнього діаметра деревостанів за різної кількості найближчих сусідів (дані Landsat 8 OLI, метод EUC)

k	*k =1*		*k >1*	
	відношення коваріації	середній діаметр, см	відношення коваріації	середній діаметр, см
1	0,92	33,5	–	–
2	1,39	33,5	2,76	31,6
3	1,05	33,5	4,36	31,2
4	1,15	33,5	5,11	31,0
5	1,19	33,5	6,36	30,7
6	1,01	33,5	7,55	30,5
7	0,99	33,5	9,51	30,5
8	1,15	33,5	12,1	30,5
9	1,00	33,5	14,1	30,4
10	1,27	33,5	15,6	30,4

ються на підставі інших, як-от середнього діаметра насаджень. Якщо $k > 1$, середній діаметр точніше оцінюється за незначної кількості найближчих сусідів. У разі збільшення їхньої кількості від 2 до 10, істотно порушується співвідношення між значеннями сум площ перерізів і кількістю дерев на 1 га, про що свідчить збільшення відношень коваріації до 15,6.

Все ж, окрім прагнення перенести вихідну коваріаційну структуру між змінними у вибірці на цільовий набір даних, необхідно також орієнтуватися на точність *k*-NN методу, яку визначає середнє квадратичне значення відхилень між фактичними і змодельованими значеннями змінної. Зазвичай, при збільшенні кількості найближчих сусідів величина зменшується (рис. 1). Звідси випливає додаткове завдання: як знайти компроміс між різними параметрами, що визначають якість прогнозу. Вибір кількості сусідів здійснюють експериментальним шляхом, намагаючись зменшити величину середньоквадратичного відхилення .

Відповідно до представлених даних, величина *k* у діапазоні від 6 до 10 забезпечує найменші середні квадратичні помилки визначення суми площ перерізів дерев на 1 га для всіх знімальних систем. Візуально простежується, що більшим значенням *k* відповідають менші помилки RMSD, однак доцільність їх використання є сумнівною. У цьому можна переконатися, порівнявши співвідношення між фактичними та змодельованими значеннями сум площ перерізів за $k = 1$ та $k = 9$ (рис. 2).

На основі аналізу одержаних результатів за знімками Landsat 8 OLI, SPOT 7 та PlanetScope, можна зробити певні висновки про вплив просторового розрізнення супутникових знімків та інформації про сезонний стан рослинності на точність *k*-NN методу. З даних рис. 1 випливає, що значення часових серій супутникових спостережень для прогнозування суми площ перерізів істотно більше, ніж просторове розрізнення знімків. Оцінка значень цього показника на підставі окремої сцени SPOT 7 по-

ступається точністю як за серією супутникових знімків PlanetScope, так і сезонними мозаїками Landsat 8 OLI. При цьому, не простежується суттєвої переваги даних комерційних сенсорів PlanetScope, які майже в 30 разів детальніші від знімків Landsat. Звідси випливає більша економічна доцільність інвентаризації великих територій із використанням часових серій безкоштовних супутникових знімків Landsat 8 OLI (Hou et al., 2018).

Графічний аналіз дає змогу зробити висновок, що, збільшуючи кількість сусідів, на основі яких здійснюють прогнозування досліджуваного показника за *k*-NN методом, істотно обмежують діапазон мінливості його значень. У разі обробки сезонних мозаїк Landsat 8 OLI метод обчислення відстаней EUC забезпечує найменшу помилку RMSD=7,1 м^2 за $k = 9$, що на 2,6 м^2 менше, ніж за $k = 1$. Проте ця перевага може проявлятися лише під час оцінки середнього значення сум площ перерізів дерев на 1 га або за наявності великого обсягу навчальної вибірки. Цілком очевидно, що для коректного прогнозування значень, які перебувають на краях факторного простору, необхідно забезпечити достатню кількість спостережень, інакше оцінки найменших значень будуть завищеними, а найбільших – навпаки, заниженими. Для відтворення просторового розподілу значень досліджуваного показника, перевагу варто віддати варіантам із меншим значенням *k*. Зокрема, найбільше узгодження фактичних і модельних значень спостерігається у варіанті з одним найближчим сусідом. При цьому фактичні та прогнозні значення сум площ перерізів рівномірно розподіляються навколо лінії 1 : 1, а це вказує, що немає систематичної помилки (рис. 2).

Ми виявили, що найменші значення помилки для всіх конфігурацій пробних ділянок забезпечує алгоритм RF. Порівняння точності прогнозування сум площ перерізів на 1 га за сезонними мозаїками Landsat 8 OLI на основі різних методів пошуку найближчих сусідів наведено в табл. 4.

Прогноз значень на основі методу RF є незміщеним для всіх конфігурацій пробних ділянок, зокрема

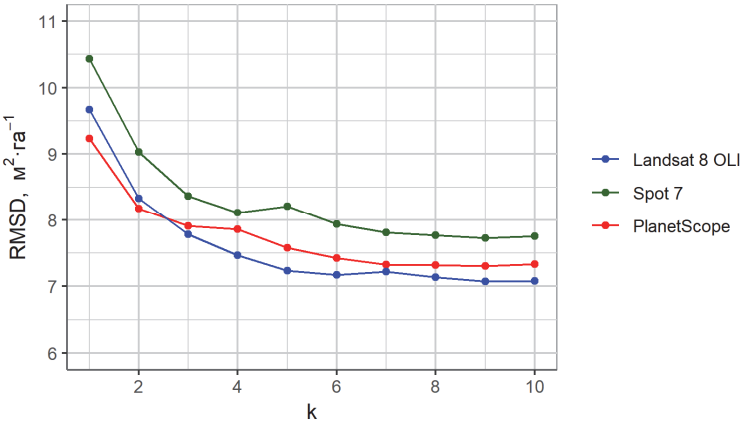

Рис. 1. Залежність величини помилки для різних сенсорів від різної кількості сусідів *k* (метод EUC)

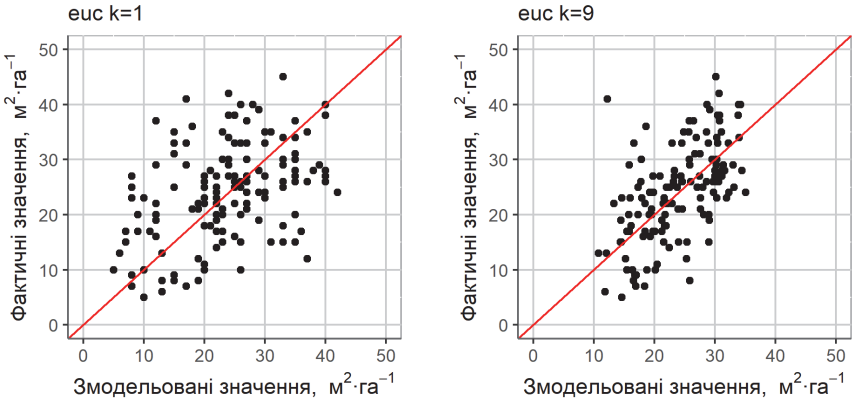

Рис. 2. Співвідношення між фактичними (АС-1) і змодельованими значеннями сум площ перерізів дерев (дані Landsat 8 OLI, метод EUC)

кругових ділянок фіксованої площі. Середні квадратичні відхилення між фактичними та прогнозними значеннями сум площ перерізів дерев на 1 га за методу RF є систематично меншими, аніж інші способи пошуку найближчих сусідів за k=1. Аналогічної точності вони досягають лише при збільшенні величини k, однак при цьому оцінки екстремальних значень стають зміщеними (рис. 2). Отже, для всіх конфігурацій пробних ділянок (крім АС-1) метод RF перевершує інші підходи до вибору найближчих сусідів.

Висновки і перспективи. Сучасні риси статистичної інвентаризації лісів визначають такі поняття, як територіальна основа вибірки, інтенсивність вибірки конфігурація пробних ділянок. Утім, упровадження технологій ДЗЗ вимагає врахування додаткових параметрів вибіркового методу обліку лісів. Проведені дослідження дають змогу зробити низку важливих висновків, які сприяють удосконаленню науково-методичного рівня лісової інвентаризації.

Першочергове значення для обліку лісів на основі дистанційних методів досліджень мають форма та розмір первинних одиниць вибірки. Відповідно до одержаних результатів, реласкопічний метод таксації лісу забезпечив найвищу точність прогнозування таксаційних показ-

4. Порівняння величини помилки при пошуку найближчих сусідів для прогнозу сум площ перерізів дерев на основі методу RF та інших мір відстаней за k=1, м²·га⁻¹ (дані Landsat 8 OLI)

Конфігурація пробних ділянок	Метод пошуку відстані				
	EUC	MAL	MSN	GNN	RF
F 1262–8	10,1	10,8	10,9	13,0	6,4
N 564–12	10,0	10,8	10,8	13,0	6,4
TA 1262–1	10,2	11,0	10,9	13,2	6,6
AC-1	9,7	9,4	8,4	10,9	8,6

ників лісових насаджень на основі *k*-NN методу. Ми виявили, що фізичний розмір пікселя супутникових знімків менше позначається на точності *k*-NN методу, ніж темпоральне розрізнення. При цьому найбільшої точності було досягнуто при використанні щільних часових серій супутникових знімків Landsat 8 OLI, які перебувають у відкритому доступі. Метод пошуку найближчих сусідів за алгоритмом RF виявився найдосконалішим, забезпечуючи меншу величину випадкової помилок за збереження коваріаційної структури вибірки. Остання умова для інших мір відстані забезпечується тільки за $k = 1$, а збільшення величини k зумовлює систематичні помилки для екстремальних спостережень.

Список літератури

Beaudoin, A., Bernier, P. Y., Guindon, L., Villemaire, P., Guo, X. J., Stinson, G., ... Hall, R. J. (2014). Mapping attributes of Canada's forests at moderate resolution through k NN and MODIS imagery. Canadian Journal of Forest Research, 44 (5), 521–532. https://doi.org/10.1139/cjfr-2013-0401

Bernier, P. Y., Daigle, G., Rivest, L.-P., Ung, C.-H., Labbé, F., Bergeron, C., & Patry, A. (2010). From plots to landscape: A k-NN-based method for estimating stand-level merchantable volume in the Province of Québec, Canada. The Forestry Chronicle, 86 (4), 461–468. https://doi.org/10.5558/tfc86461-4

Bilous, A., Myroniuk, V., Holiaka, D., Bilous, S., See, L., & Schepaschenko, D. (2017). Mapping growing stock volume and forest live biomass: a case study of the Polissya region of Ukraine. Environmental Research Letters, 12 (10), 13. https://doi.org/10.1088/1748-9326/aa8352

Breiman, L. (2001). Random forests. Machine Learning, 45 (1), 5–32. https://doi.org/10.1023/a:1010933404324

Chirici, G., McRoberts, R. E., Fattorini, L., Mura, M., & Marchetti, M. (2016). Comparing echo-based and canopy height model-based metrics for enhancing estimation of forest aboveground biomass in a model-assisted framework. Remote Sensing of Environment, 174, 1–9. https://doi.org/10.1016/j.rse.2015.11.010

Crookston, N. L., & Finley, A. O. (2008). yaImpute : An R Package for k NN Imputation. Journal of Statistical Software, 23 (10). https://doi.org/10.18637/jss.v023.i10

Franco-Lopez, H., Ek, A. R., & Bauer, M. E. (2001). Estimation and mapping of forest stand density, volume, and cover type using the k-nearest neighbors method. Remote Sensing of Environment, 77 (3), 251–274. https://doi.org/10.1016/S0034-4257(01)00209-7

Haapanen, R., Ek, A. R., Bauer, M. E., & Finley, A. O. (2004). Delineation of forest/nonforest land use classes using nearest neighbor methods. Remote Sensing of Environment, 89 (3), 265–271. https://doi.org/10.1016/j.rse.2003.10.002

Hou, Z., McRoberts, R. E., Ståhl, G., Packalen, P., Greenberg, J. A., & Xu, Q. (2018). How much can natural resource inventory benefit from finer resolution auxiliary data? Remote Sensing of Environment, 209, 31–40. https://doi.org/10.1016/j.rse.2018.02.039

Hudak, A. T., Crookston, N. L., Evans, J. S., Hall, D. E., & Falkowski, M. J. (2008). Nearest neighbor imputation of species-level, plot-scale forest structure attributes from LiDAR data. Remote Sensing of Environment, 112 (5), 2232–2245. https://doi.org/10.1016/j.rse.2007.10.009

Kirchhoefer, M., Schumacher, J., Adler, P., & Kändler, G. (2017). Considerations towards a Novel Approach for Integrating Angle-Count Sampling Data in Remote Sensing Based Forest Inventories. Forests, 8 (7), 239. https://doi.org/10.3390/f8070239

Latifi, H., Fassnacht, F. E., Hartig, F., Berger, C., Hernández, J., Corvalán, P., & Koch, B. (2015). Stratified aboveground forest biomass estimation by remote sensing data.

International Journal of Applied Earth Observation and Geoinformation, 38, 229–241. https://doi.org/10.1016/j.jag.2015.01.016

Maltamo, M., Korhonen, K., Packalen, P., Mehtatalo, L., & Suvanto, A. (2007). Testing the usability of truncated angle count sample plots as ground truth in airborne laser scanning-based forest inventories. Forestry, 80 (1), 73–81. https://doi.org/10.1093/forestry/cpl045

McRoberts, R. E. (2009a). A two-step nearest neighbors algorithm using satellite imagery for predicting forest structure within species composition classes. Remote Sensing of Environment, 113 (3), 532–545. https://doi.org/10.1016/j.rse.2008.10.001

McRoberts, R. E. (2009b). Diagnostic tools for nearest neighbors techniques when used with satellite imagery. Remote Sensing of Environment, 113 (3), 489–499. https://doi.org/10.1016/j.rse.2008.06.015

McRoberts, R. E. (2012). Estimating forest attribute parameters for small areas using nearest neighbors techniques. Forest Ecology and Management, 272, 3–12. https://doi.org/10.1016/j.foreco.2011.06.039

McRoberts, R. E., Liknes, G. C., & Domke, G. M. (2014). Using a remote sensing-based, percent tree cover map to enhance forest inventory estimation. Forest Ecology and Management, 331, 12–18. https://doi.org/10.1016/j.foreco.2014.07.025

McRoberts, R. E., Nelson, M. D., & Wendt, D. G. (2002). Stratified estimation of forest area using satellite imagery, inventory data, and the k-Nearest Neighbors technique. Remote Sensing of Environment, 82 (2–3), 457–468. https://doi.org/10.1016/S0034-4257(02)00064-0

Mozgeris, G. (2008). Estimation and Use of Continuous Surfaces of Forest Parameters: Options for Lithuanian Forest Inventory. Baltic Forestry, 14 (2), 9.

Myroniuk, V. (2017). Variable selection in the context of forest cover mapping using seasonal Landsat mosaics. Scientific Herald of NULES of Ukraine, 278, 66–76 (in Ukrainian).

Myroniuk, V. (2018). Forest cover mapping using Landsat-based seasonal composited mosaics. Scientific Bulletin of NFWU of Ukraine, 28 (1), 28–33 (in Ukrainian). https://doi.org/10.15421/40280105

Ohmann, J. L., Gregory, M. J., & Roberts, H. M. (2014). Scale considerations for integrating forest inventory plot data and satellite image data for regional forest mapping. Remote Sensing of Environment, 151, 3–15. https://doi.org/10.1016/j.rse.2013.08.048

Packalén, P., Temesgen, H., & Maltamo, M. (2012). Variable selection strategies for nearest neighbor imputation methods used in remote sensing based forest inventory. Canadian Journal of Remote Sensing, 38 (5), 557–569. https://doi.org/10.5589/m12-046

Reese, H., Nilsson, M., Sandström, P., & Olsson, H. (2002). Applications using estimates of forest parameters derived from satellite and forest inventory data. Computers and Electronics in Agriculture, 37 (1–3), 37–55. https://doi.org/10.1016/S0168-1699(02)00118-7

Tomppo, E., & Halme, M. (2004). Using coarse scale forest variables as ancillary information and weighting of variables in k-N-N estimation: a genetic algorithm approach. Remote Sensing of Environment, 92 (1), 1–20. https://doi.org/10.1016/j.rse.2004.04.003

Tomppo, E., Kuusinen, N., Mäkisara, K., Katila, M., & McRoberts, R. E. (2017). Effects of field plot configurations on the uncertainties of ALS-assisted forest resource estimates. Scandinavian Journal of Forest Research, 32(6), 488–500. https://doi.org/10.1080/02827581.2016.1259425

V. V. Myroniuk, A. M. Bilous, P. P. Diachuk (2019). Predicting forest stand parameters using the k-NN approach. *UKRAINIAN JOURNAL OF FOREST AND WOOD SCIENCE, 10(2): 51-63. http://dx.doi.org/10.31548/forest2019.02.051.*

Remote sensing (RS) data is a key source of auxiliary information on forest stand parameters used in forest inventory. Among the modern techniques that allow coupling ground truth and RS derived information, the nonparametric k-Nearest Neighbors (k-NN) approach has been widely used for imputing missing values in forest inventory data. In order to evaluate the effectiveness of the approach using different predictors and input parameters, we used experimental data collected in local forest inventory conducted in middle part of Kyiv region (Ukraine). The data consist of 141 sample plots distributed systematically within the forested area of about 56 km^2. Sample trees were selected using two different approaches – using sampling with probability proportional to the tree sizes and stands density. As an auxiliary dataset, we incorporated three sources of spectral information that represents different types of RS: 1) the seasonal mosaics of Landsat 8 OLI imagery; 2) one scene of SPOT 7 image; 3) time series of PlanetScope observations. All spectral data were orthorectified and radiometrically calibrated to top of atmosphere reflectance. We concluded that the accuracy of prediction of basal areas for all types of RS is higher for angle-counting plots. We refer this fact to the technique of tree selection since trees of bigger sizes are likely selected by relascope. These trees reflect parameters of forest overstory scanned by RS sensors. Among used satellite imagery, dense time series of Earth observations showed better agreement between the reference and predicted values of basal area per hectare. Thus, we concluded that pixel size is less important in case of predicting forest parameters compared to the temporal resolution of RS data. According to our findings, the k = 1 implementation of the k-NN method works better when preservation of covariance between variables is important while increasing the k value reduces their ranges. It was found that random forest was the most accurate nearest neighbors search routine for the k-NN method.

Keywords: *sample-based forest inventory, fixed-area plots, angle-count sampling, classification.*

Отримано : 18.11.2018 р.