

Регресійні моделі прогнозування врожайності озимої пшениці на основі супутникових даних



Куссуль Н.М, Шелестов А.Ю., Скакун С.В.,
Куссуль О.М., Колотій А.В

Інститут космічних досліджень НАН України та
ДКА України
GEO-UA 2014
27-28 травня, Київ

Зміст



- Об'єкт дослідження
- Методологія
 - Емпіричні моделі
 - Біофізичні моделі
- Інформаційна технологія прогнозування врожайності
- Результати
 - Прогнози врожайності за супутниковими даними різної природи
 - Відбір інформаційних ознак
 - Верифікація результатів



Існуючі підходи



- **Моделі росту (біофізичні)**

- Симуляція росту культур для отримання значень таких біофізичних параметрів як валовий збір, величина біомаси, споживання води, і т.д.
- Переваги
 - Висока «фізичність» моделі та легкість інтерпретації
- Недоліки
 - Велика кількість вхідних параметрів;
 - Необхідність в коректній адаптації для кожного окремого поля та на загальнодержавному рівні(в поточній реалізації).

- **Кібернетичні (індукційні) моделі**

- Врожайність зернових пов'язується з певними вибраними предикторами (вегетаційні індекси (NDVI, VHI, FAPAR), метеорологічними спостереженнями, тощо и т.д.)
- Переваги
 - Наявність доступних джерел регулярних даних;
- Недоліки
 - Існують моделі для окремих полів або глобального рівня (з низьким розрізненням)
 - Відсутня технологія оцінки точності (верифікації)



Мета роботи



- **Україна**

- 8-ма серед найбільших країн-експортерів та 10-та серед виробників озимої пшениці в світі станом на 2011-й рік

- **Проблема:** при наявності рядів супутникових та наземних даних відсутня технологія надійного прогнозування врожайності, застосовна на рівні адміністративних одиниць та країни в цілому

- **Мета роботи:**

- *Підвищити достовірність та завчасність прогнозування врожайності озимої пшениці для території України шляхом подальшого розвитку регресійного підходу; створити автоматизовану інформаційну технологію прогнозування на основі супутникових даних різної природи.*



Алгоритм побудови моделей для областей



Підходи



- Три підходи

- Емпіричні регресійні моделі, побудовані із використанням даних **MODIS**;
- Емпіричні регресійні моделі, побудовані із використанням **метеорологічних** параметрів;
- Адаптована для території України модель **Crop Growth Monitoring System (CGMS)**, яка базується на симуляційній моделі **WOFOST** та метеорологічних параметрах

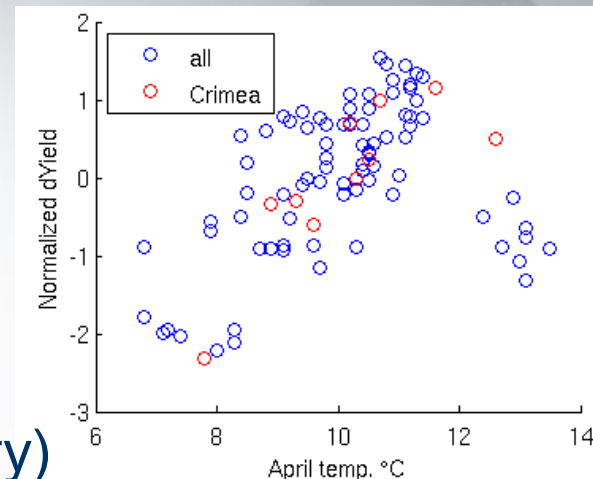


Прогнозування за метеорологічними параметрами



- Дослідження зв'язку між врожайністю (озима пшениця) та **основними метеорологічними параметрами**:

- Максимальна температура
- Мінімальна температура
- Середня температура
- Опади
- Запас вологи:
 - 0-20 см (продуктивний шар ґрунту)
 - Дані за вересень, жовтень, квітень, травень, червень
- Щомісячні **метеорологічні спостереження** з 180 станцій на території України за 2000-2011 рр.



- Методологія**

- Кореляційний аналіз
- Лінійна багатофакторна регресія
- Нелінійна багатофакторна регресія



Нелінійна регресія



- Нейронні мережі
- Ядерні методи:
Gaussian processes regression

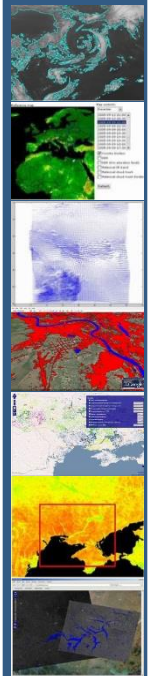
$$\left\{ (x_i, y_i), i = \overline{1, n} \right\}$$

$$m_* = M \left[f(x_*) | X, Y \right] = k_*^T \left[K(X, X) + \sigma^2 I \right]^{-1} Y$$

$$\sigma_*^2 = k(x_*, x_*) - k_*^T \left[K(X, X) + \sigma^2 I \right]^{-1} k_* \quad k(x, x') = \exp \left\{ -\frac{1}{2l^2} \|x - x'\|^2 \right\}$$

$$\bar{f}(x_*) = m_* = \sum_{i=1}^n \alpha_i k(x_i, x_*) \quad \alpha_i = \left[K(X, X) + \sigma^2 I \right]^{-1} Y$$

$$k_{ARD}(x, x') = \sigma_f^2 \exp \left\{ -\frac{1}{2} \sum_{j=1}^m \frac{1}{l_j^2} |x_j - x'_j|^2 \right\} + \sigma_n^2 \delta(x, x')$$

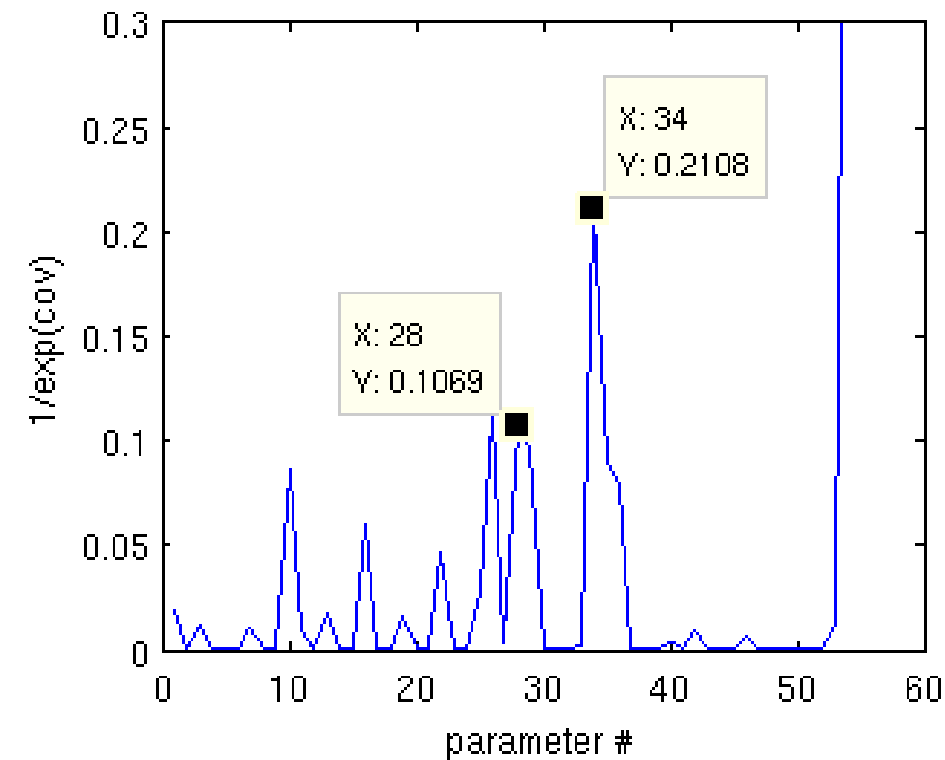


Відбір ознак

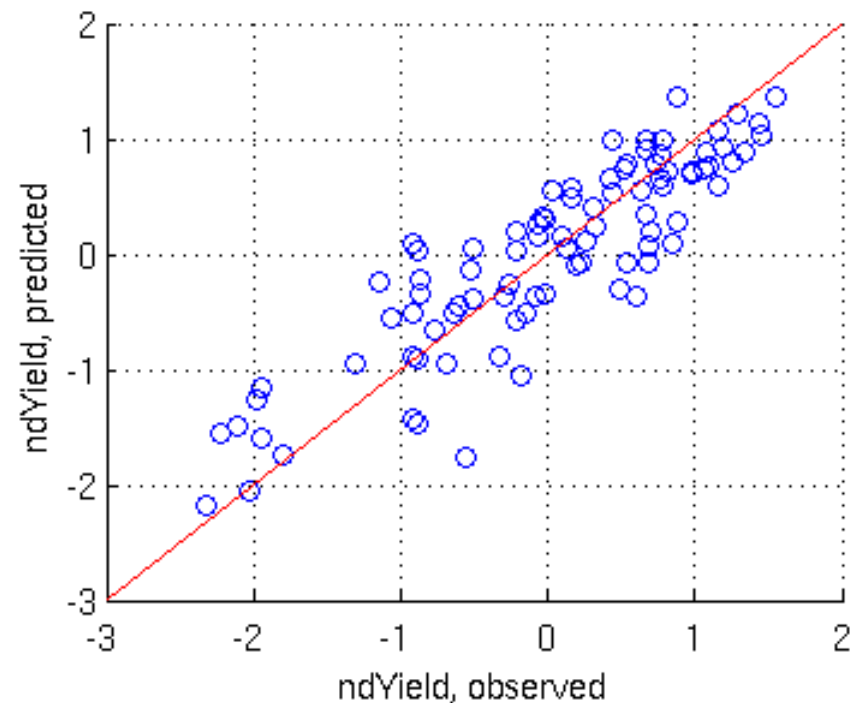


- Найбільш релевантні
– Середні температури (Травень, Жовтень, Квітень)

Parameter relevance



Leave-one-out CV, Set1



Ефективність, 2010 рік

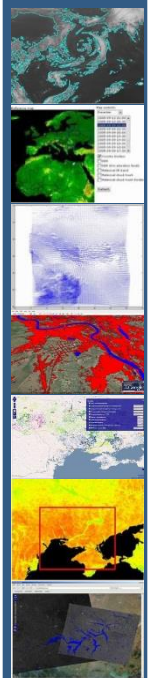


Область	Eff.
Дніпропетровська	2,999
Донецька	2,322
Запорізька	1,800
Кіровоградська	2,469
Луганська	1,845
Миколаївська	1,855
Одеська	2,511
Херсонська	2,759
АР Крим	3,592

За інші роки такої залежності не спостерігалось

Висновок:

Підхід не робастний (для кожного року - свої нелінійності)



Прогнозування за вегетаційними індексами



- **Клас моделей**

- Для побудови моделей використовується 1 предиктор (NDVI)
- Збільшення кількості предикторів не призводить до підвищення точності
- Критерієм оцінки якості моделей є мінімізація СКП (RMSE)

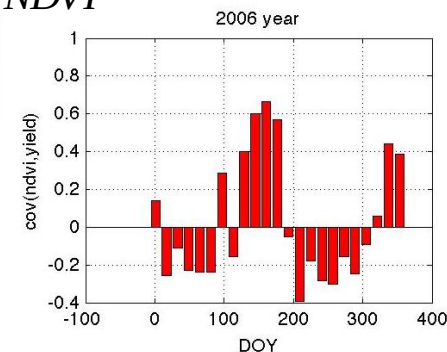
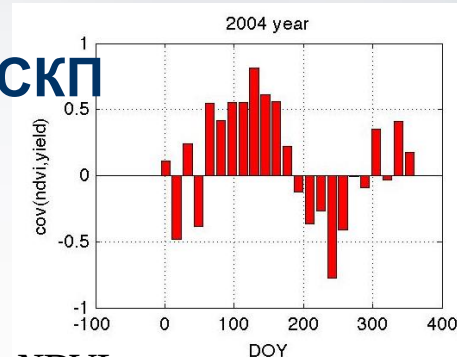
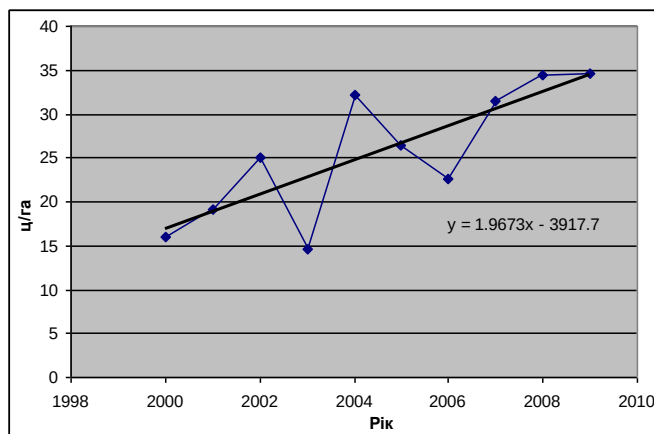
$$Y_i = T_i + dY_i$$

урожайність в год i

$$T_i = a_0 + a_1 * i$$

трендовий компонент

$$dY_i = Y_i - T_i = f(NDVI_i) = b_0 + b_1 * NDVI_i \equiv dY_i^{NDVI}$$

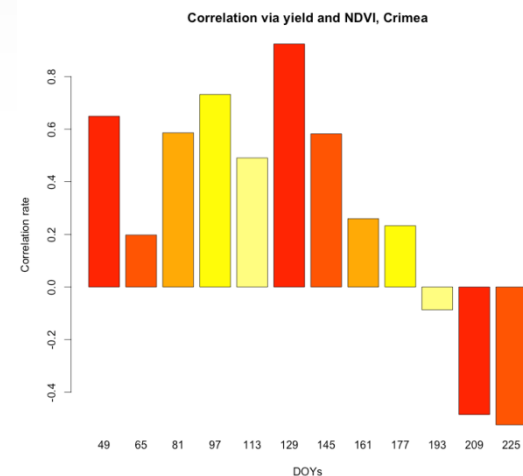
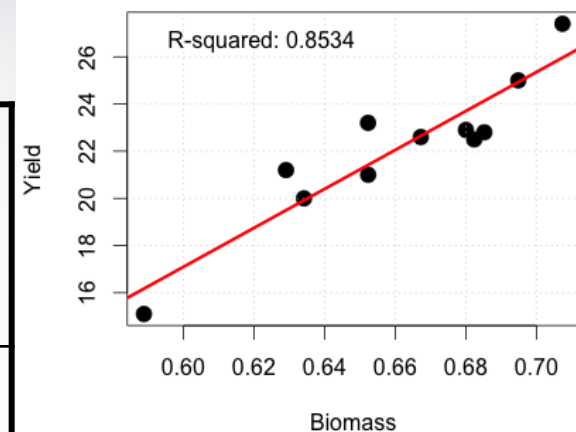


Результаты работы



- Относительная эффективность и коэффициент детерминации моделей (NDVI) - усредненные в пределах агроклиматических зоны

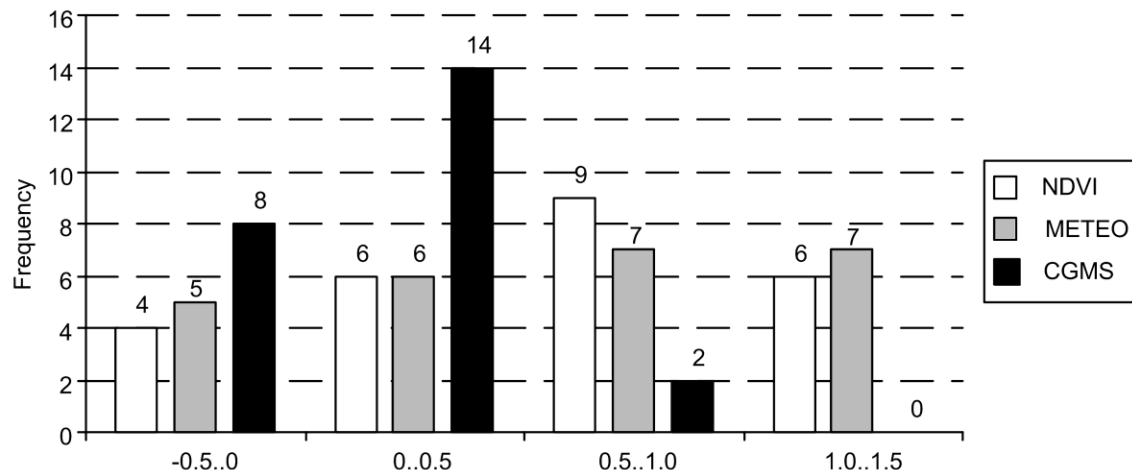
	Модель, построенная на данных 2000-2011		Модель, построенная на данных 2000-2012	
	Относ. эфф.	R ²	Относ. эфф.	R ²
Агроклиматическая зона				
Полесье	1,182	0,479	1,177	0,433
Лесостепь	1,576	0,667	1,532	0,680
Степь	1,883	0,804	1,894	0,796



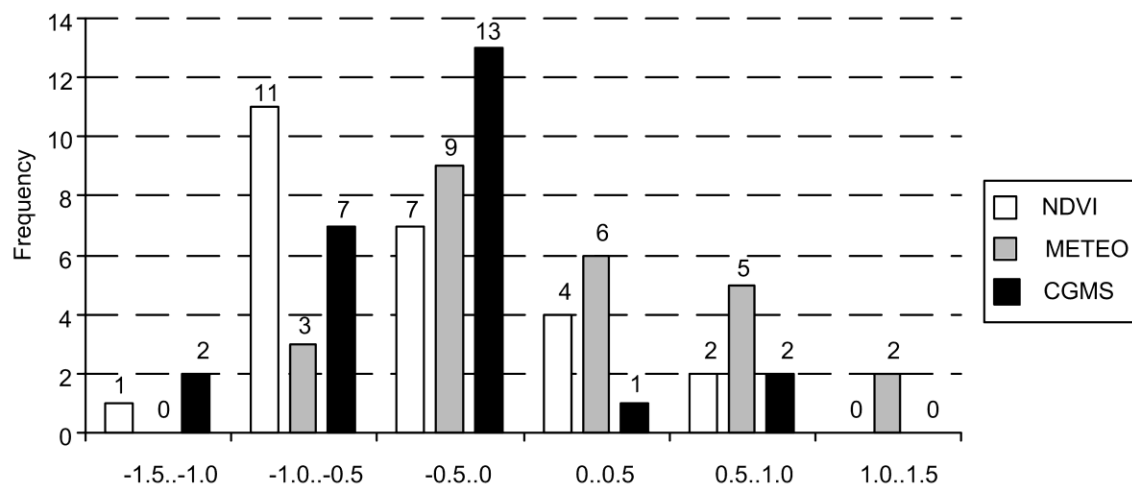
Порівняння результатів різних підходів



- Гістограма помилок прогнозу врожайності озимої пшениці 2010 (А) та 2011 рр. (В)



(A) Predicted minus observed winter wheat yield for 2010, t ha⁻¹



(B) Predicted minus observed winter wheat yield for 2011, t ha⁻¹

Порівняння результатів підходів, 2010-11 рр.

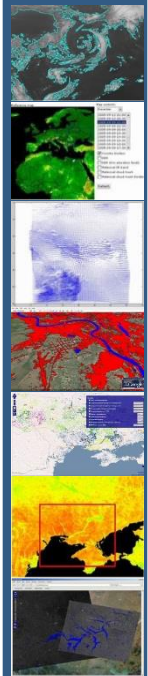


Minimum RMSE values of winter wheat yield forecasts for 2010 and 2011 produced by different models.

Model and time of forecast	RMSE for 2010, t ha ⁻¹	RMSE for 2011, t ha ⁻¹
MODIS NDVI (2000–2009), April–May	0.794	0.585
MODIS NDVI (2000–2010), April–May	–	0.625
Meteorological (2000–2009), April–May	0.779	0.565
CGMS (2000–2009), June	0.304	–
CGMS (2000–2010), June	–	0.509
Trend model (2000–2009)	1.346	0.944
Trend model (2000–2010)	–	1.050

Висновки:

- Модель на основі метеорологічних предикторів не є робасною
- Точність індуктивних та біофізичних моделей є приблизно однаковою
- Надалі будемо використовувати супутникові дані



Оптимізація супутникових предикторів прогностичної моделі



- **Продукт MOD13 (NDVI)** з роздільною здатністю 250 м за 2000-2013 рр.
- **Часові ряди FAPAR** (1 км) 2000-2013 рр.
- **Часові ряди VHI** (16 км), 2000-2013 рр.
- Карта ESA Global Land Cover (**GlobCover**) з роздільною здатністю 300 м за 2008 г., перемасштабована карта GlobCover з роздільною здатністю 1000 м
- Дані офіційної статистики по врожайності озимої пшениці для території України на обласному рівні за 2000-2012 рр.



Оцінка ефективності



- **Крос-валідація**

- «відкидаємо по одному» (**LOOCV**)
- Використовуємо одне спостереження із вихідної вибірки як тестову вибірку, решту – як навчальну

- **Критерій**

- RMSE (середньоквадратична похибка) на тестовій вибірці

P_i и O_i – прогноз та факт по врожайності озимої пшениці

$$RMSE = \sqrt{\frac{1}{n} \sum_i (P_i - O_i)^2}$$

- **Відносна ефективність**

$$\text{Rel. eff.} = \frac{V(Y_{\text{sample}})}{V(Y_{\text{satellite}})} = \frac{\frac{1}{n} \sum_i (dY_i)^2}{RMSE^2}$$

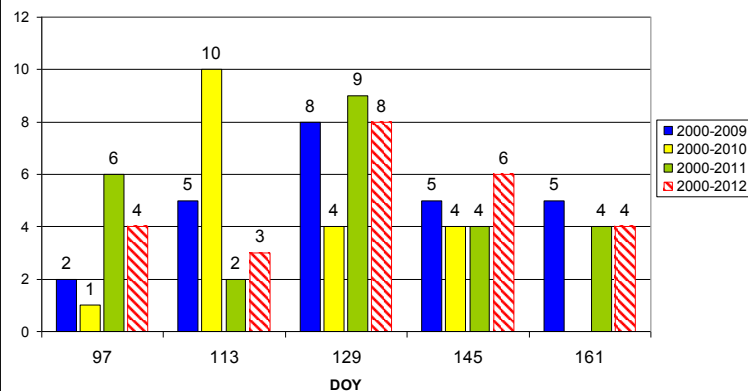
- dY_i - відхилення від тренду



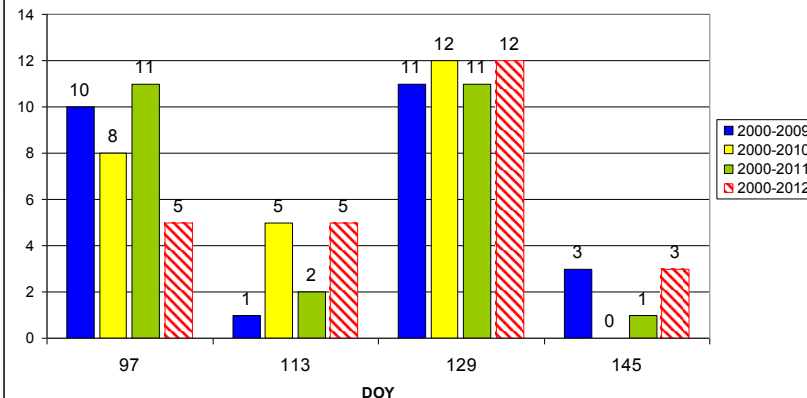
Верифікація вибору предикторів-1



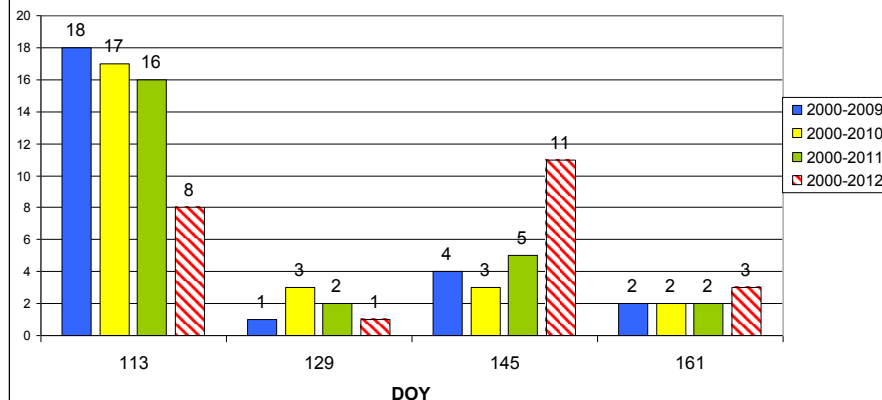
Гистограмма значений DOY для рядов FAPAR



Гистограмма значений DOY для рядов NDVI

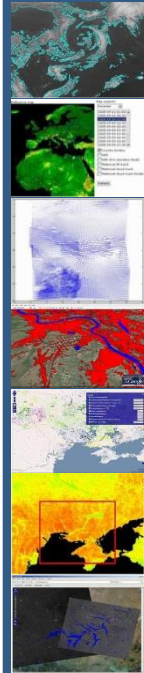
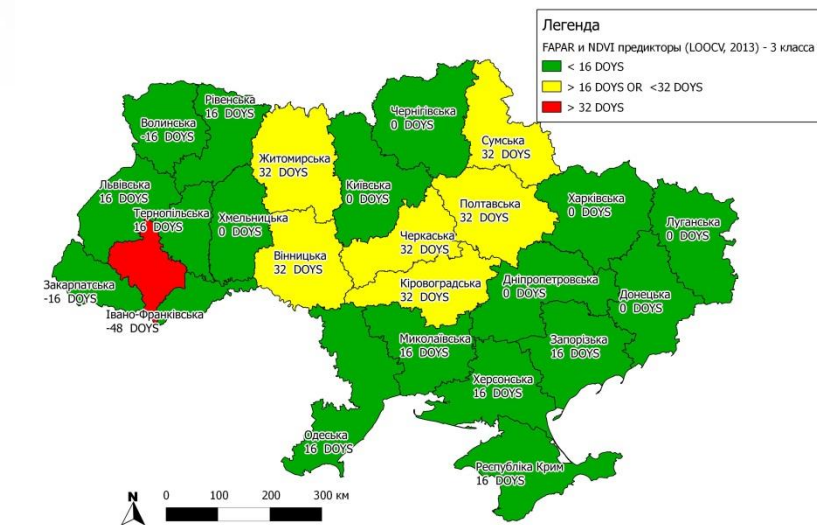
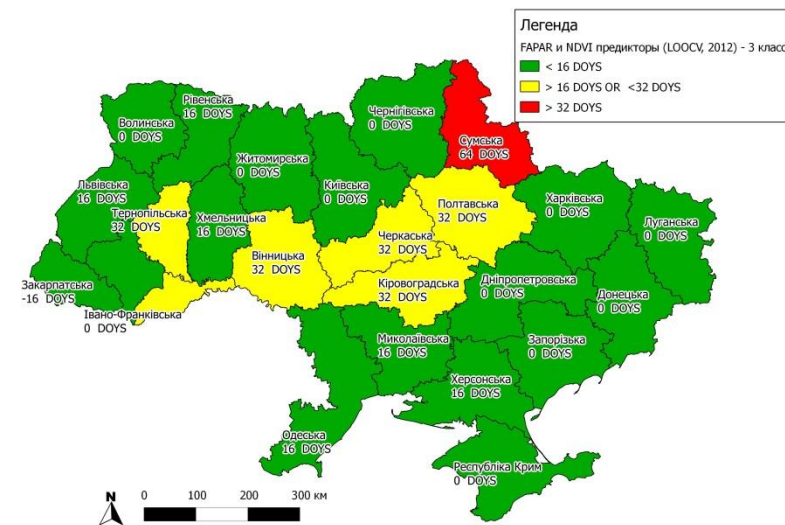
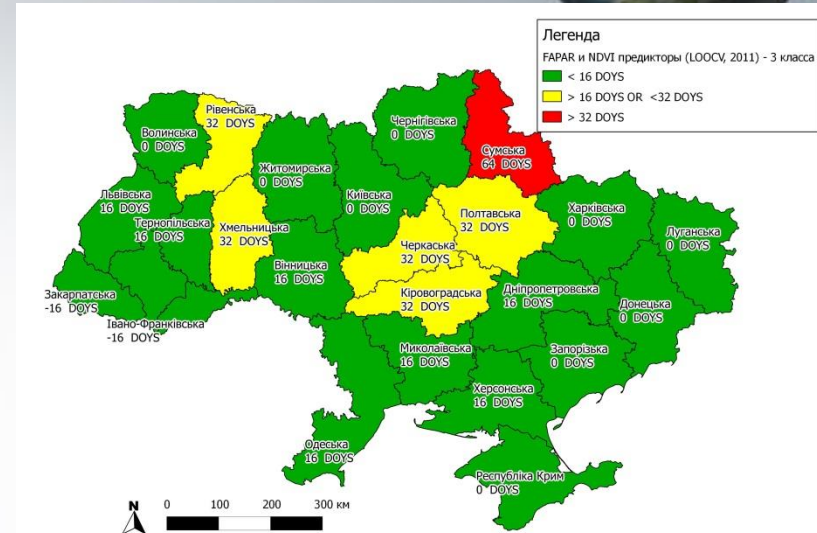
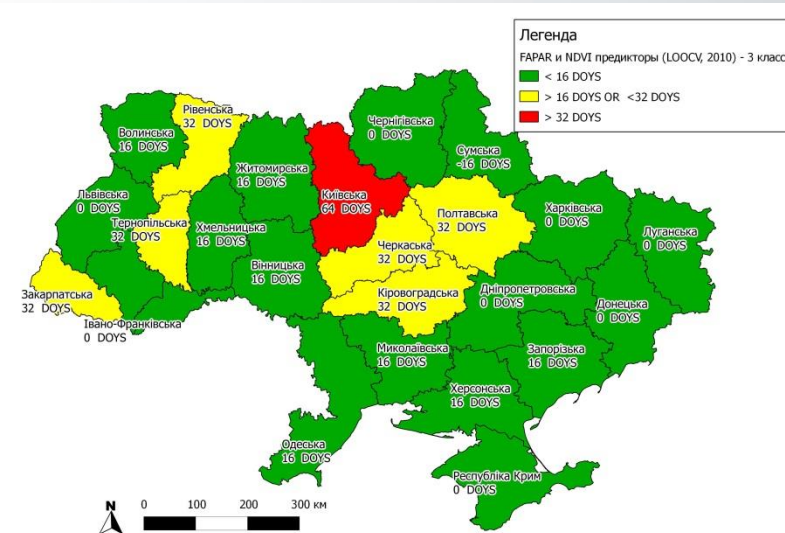


Гистограмма значений DOY для рядов VHI



Для більшості областей обирається **16-17** тиждень (~**113** день) або **18-19** тиждень (~**129** день)

Верифікація вибору предикторів-1



Перевірка часової узгодженості вибору ознак

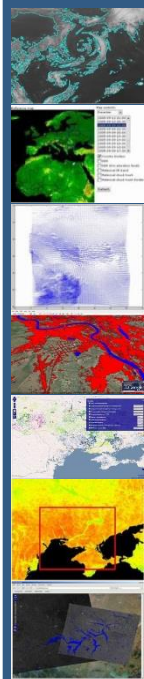


Алгоритм RandomForest

$T = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ –обучающая выборка,

Алгоритм *RandomForest*:

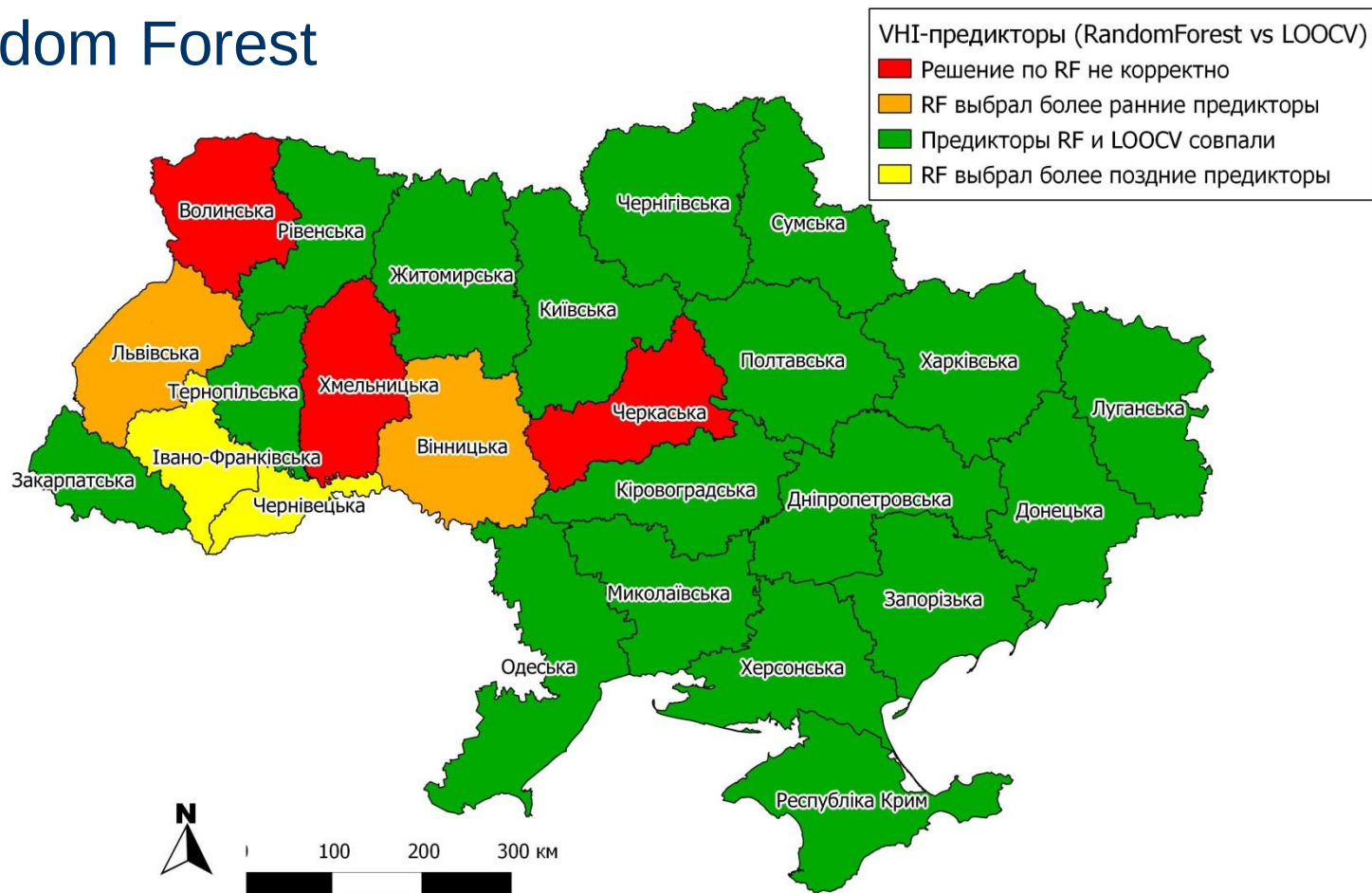
1. Пусть строим «лес» из S деревьев, формируются обучающие выборки $\{T_1, T_2, \dots, T_S\}$.
2. Используя m признаков, на выборках $\{T_1, T_2, \dots, T_S\}$ строится S классификационных деревьев $K_i : \forall i = 1..s \exists T_i \rightarrow K_i$
3. При принятии решения о классификации входа, опрашиваются все S деревьев ансамбля и принимается решение, поддержанное большинством классификаторов.



VHI: оцінка коректності вибору предикторів



- Методи пошуку:
 - LOOCV
 - Random Forest



Прогностичні моделі на основі різних супутникових даних



- Для побудови моделей використовується **1 предиктор**:

- $NDVI$,
- VHI ,
- $FAPAR$

- **RMSE - критерій якості** моделей

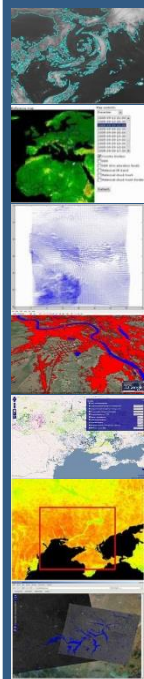
$$Y_i = T_i + dY_i \quad \text{врожайність в рік } i$$

$$T_i = a_0 + a_1 * i \quad \text{трендовий компонент}$$

$$dY_i = Y_i - T_i = f(NDVI_i) = b_0 + b_1 * NDVI_i \equiv dY_i^{NDVI}$$

$$dY_i^{VHI} = Y_i - T_i = f(VHI_i) = k_0 + k_1 * VHI_i$$

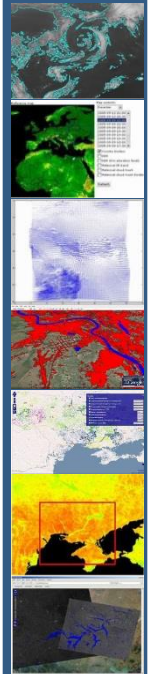
$$dY_i^{FAPAR} = Y_i - T_i = f(FAPAR_i) = h_0 + h_1 * FAPAR_i$$



Точність моделей (помилки) на різних даних



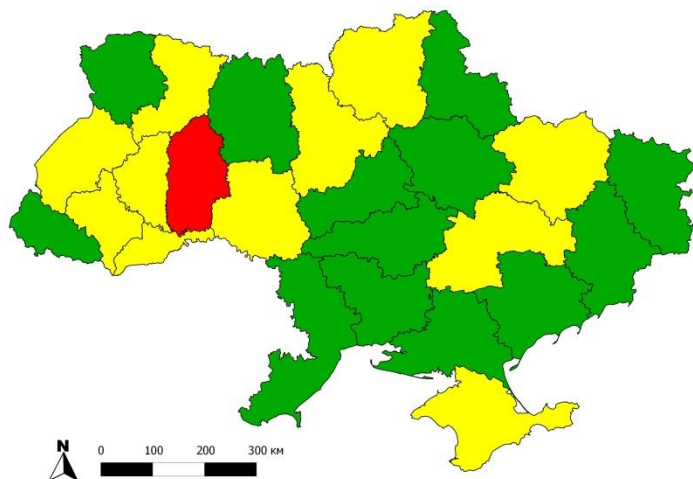
		2010	2011	2012	2013
<i>NDVI</i>	RMSE	8.2	6.2	6.8	5.8
	average	6.8	-3.7	-3.4	2
<i>VHI</i>	RMSE	6.3	5.1	7.0	5.9
	average	5.5	-3.8	-3.6	-2.1
<i>FAPAR</i>	RMSE	8.9	5.2	5.6	4.1
	average	7.6	-2.1	-0.5	0.8



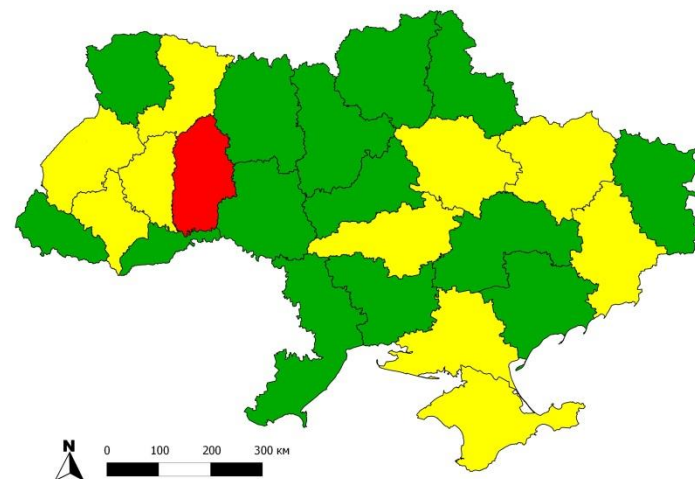
Помилки прогнозу за 2011-й рік



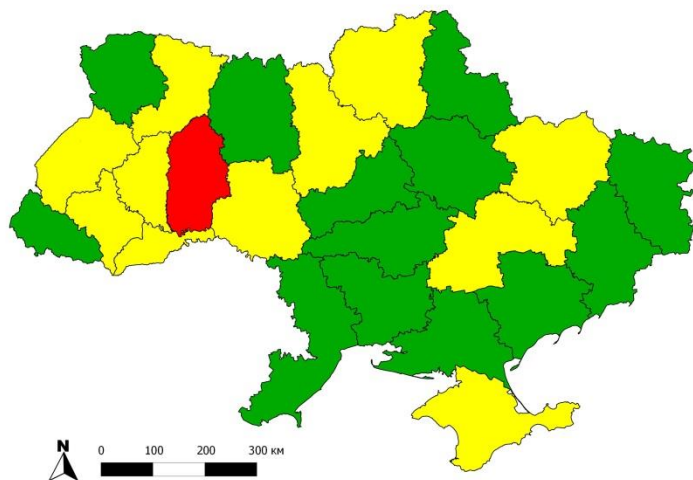
FAPAR



VHI



NDVI



Легенда

Ошибка прогноза, 2011 год

■ < 5 ц/га

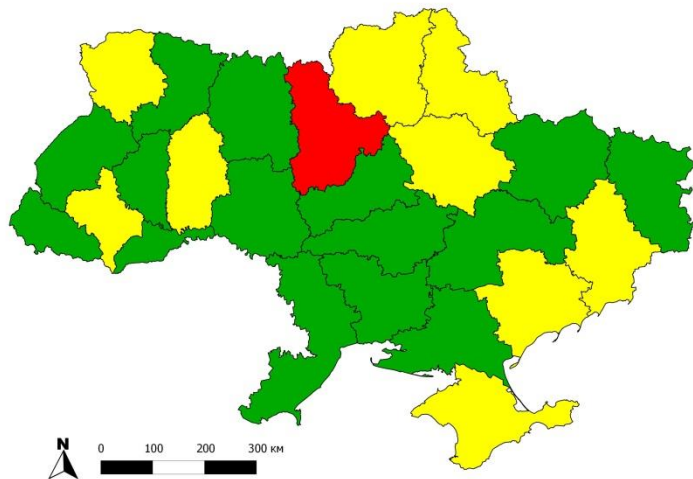
■ < 10 ц/га AND > 5 ц/га

■ > 10 ц/га

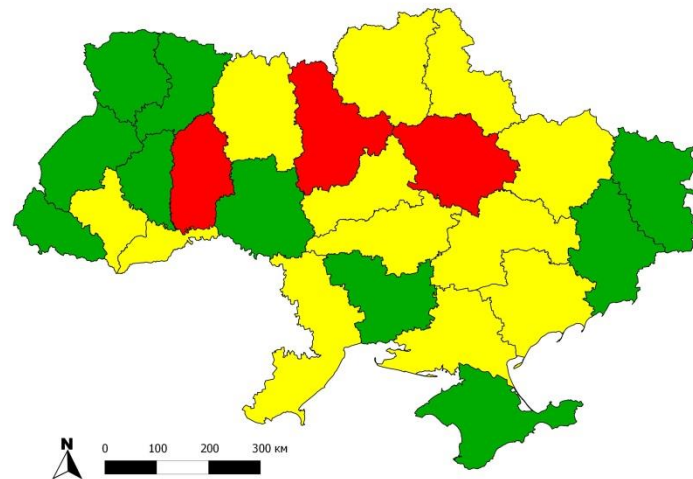
Помилки прогнозу за 2012-й рік



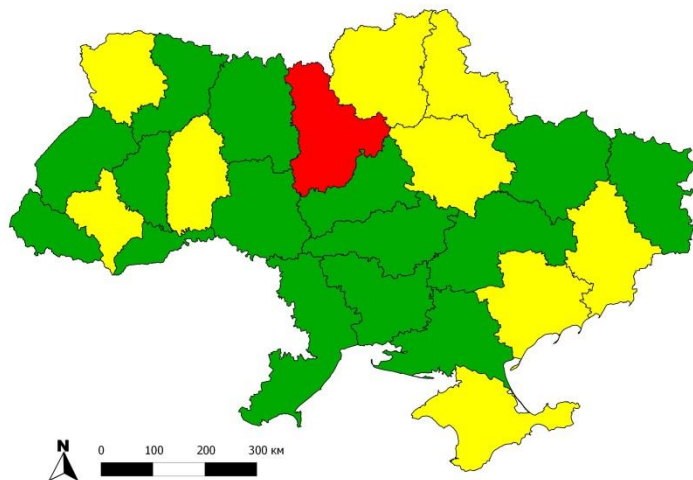
FAPAR



VHI



NDVI



Легенда

Ошибка прогноза, 2012 год

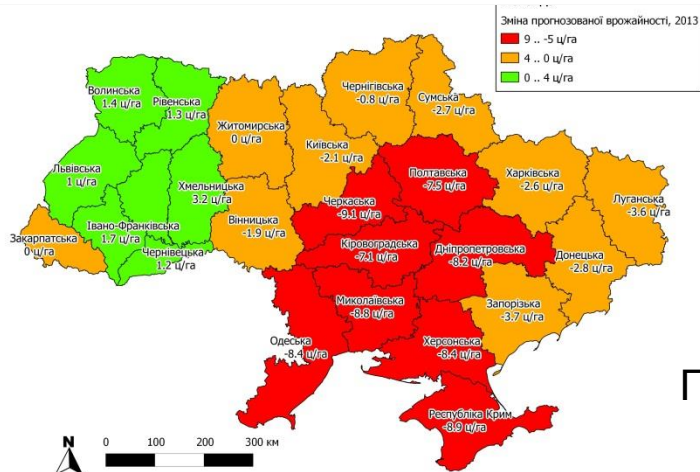
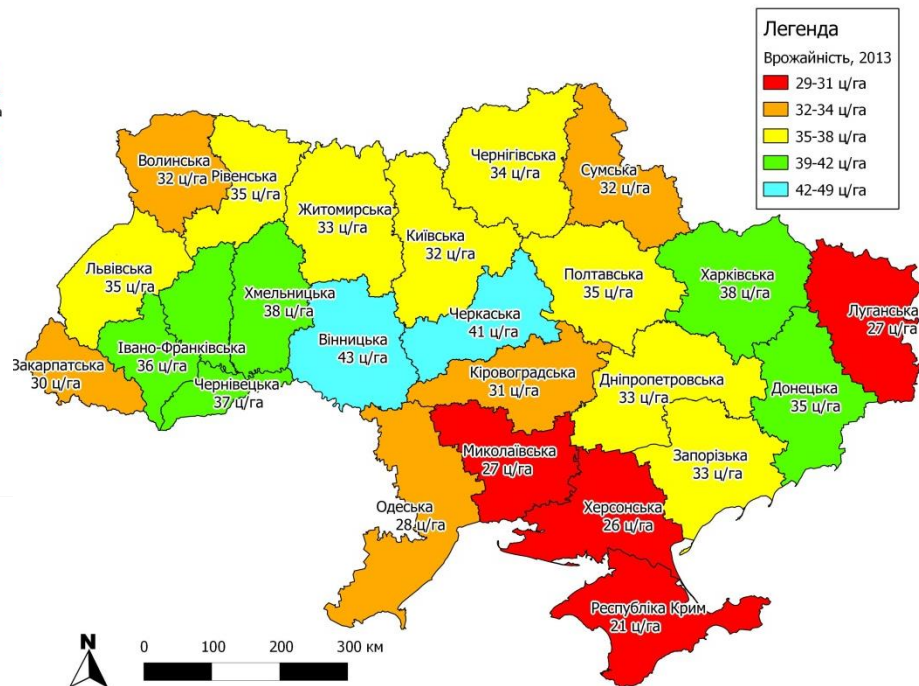
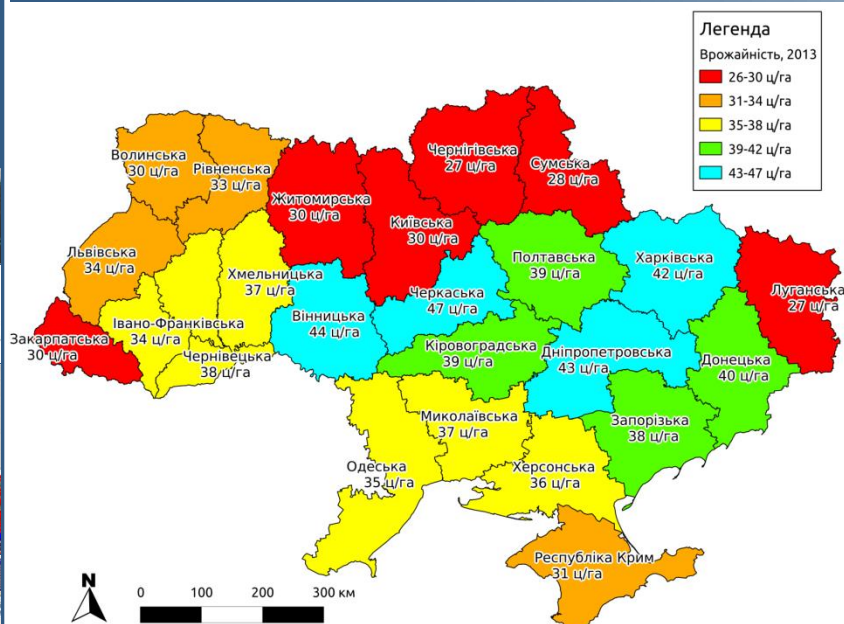
■ < 5 ц/га

■ < 10 ц/га AND > 5 ц/га

■ > 10 ц/га



Оперативні прогнози, 2013 р.



Посуха, зниження прогнозованої врожайності

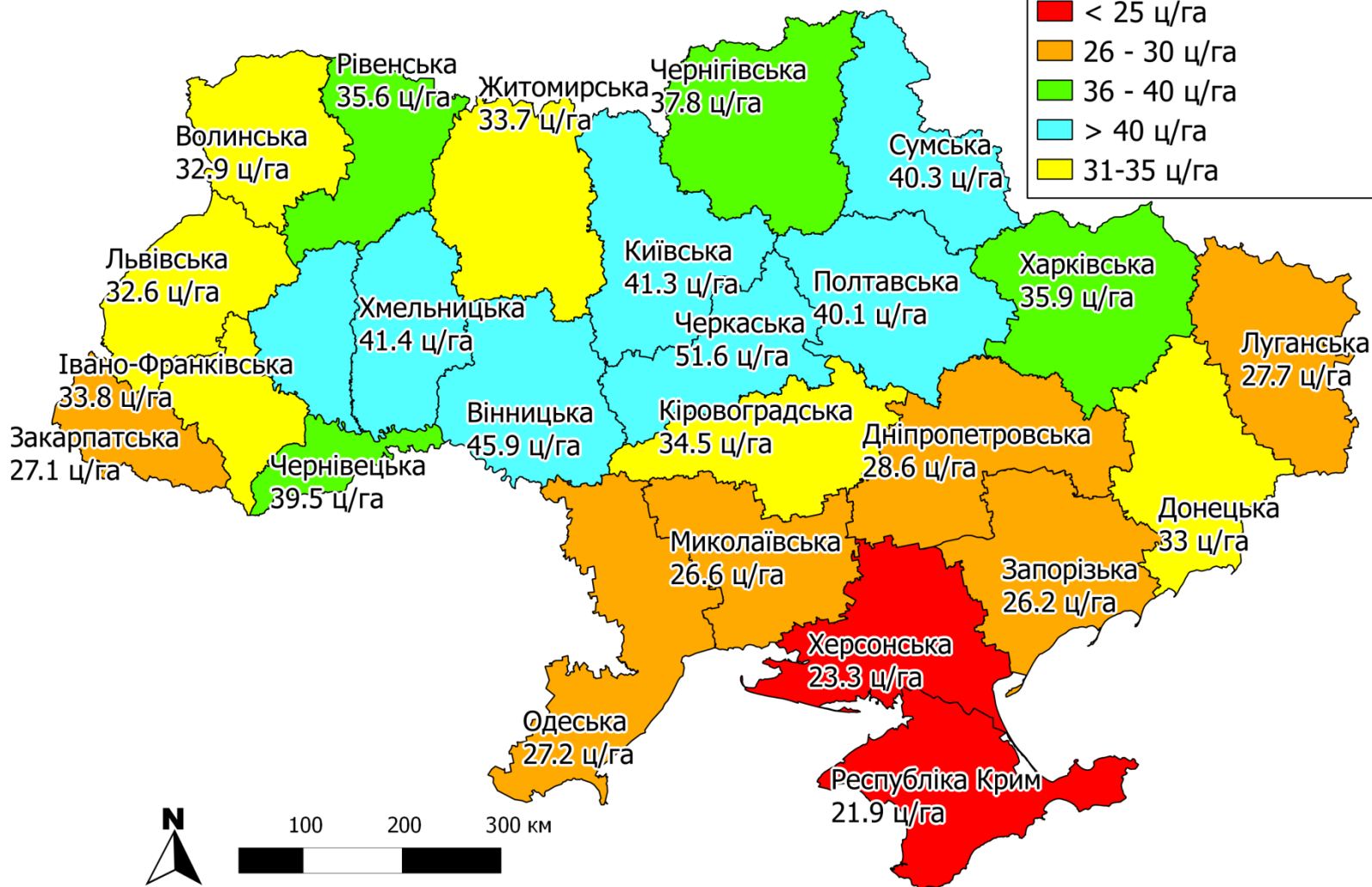
Оперативний прогноз, 2014



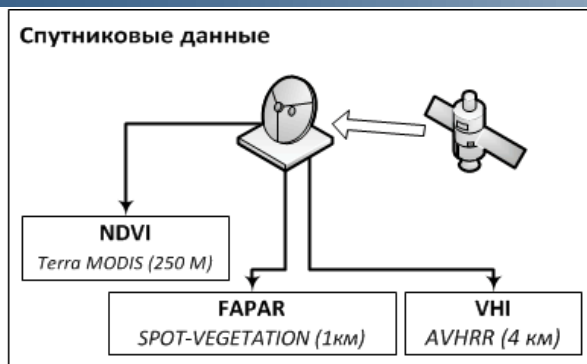
Легенда

Врожайність, 2014 (попередня)

- < 25 ц/га
- 26 - 30 ц/га
- 36 - 40 ц/га
- > 40 ц/га
- 31-35 ц/га



Автоматизована інформаційна технологія прогнозування врожайності

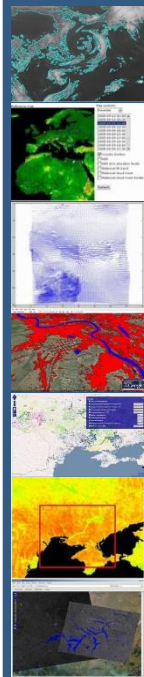
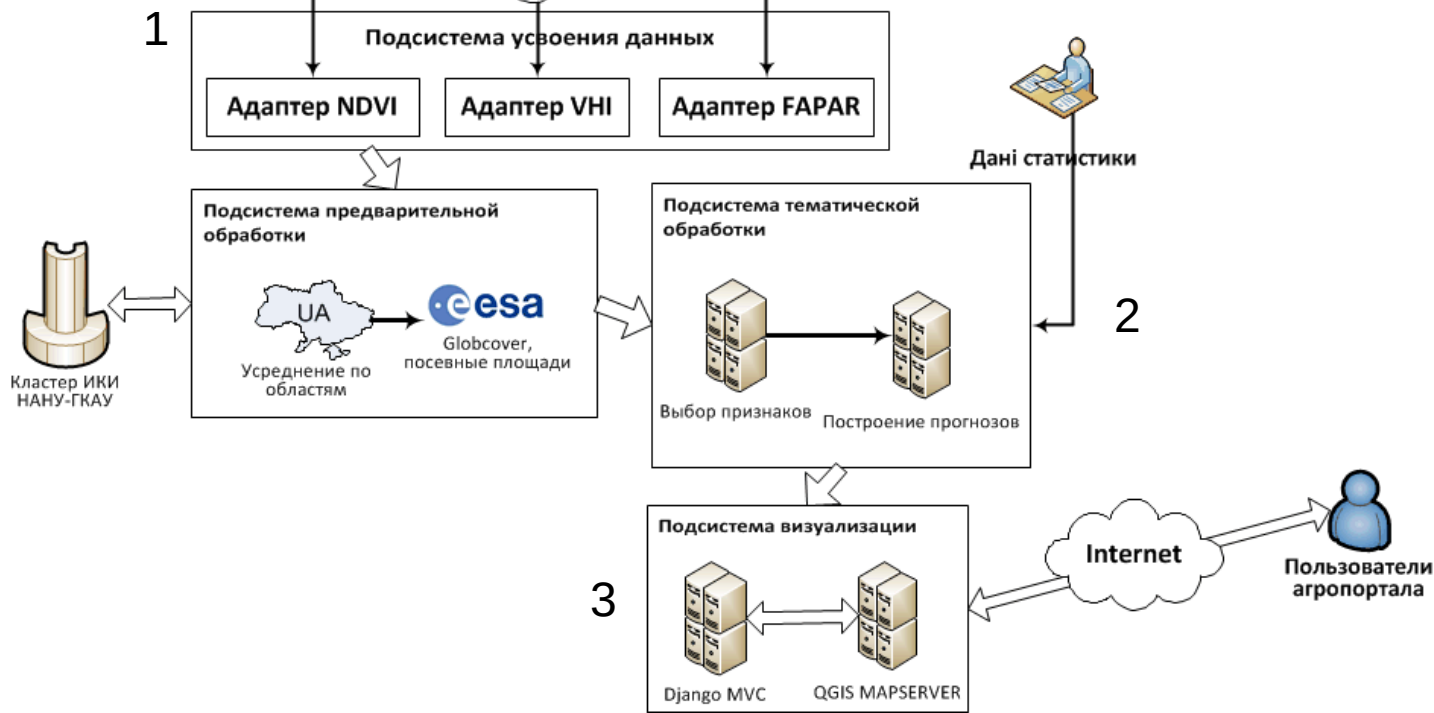


Три основных уровня:

- 1) уровень получения данных;
- 2) уровень бизнес-логики;
- 3) уровень пользовательского интерфейса;

Подсистемы:

подсистема усвоение данных;
подсистема предварительной обработки;
подсистема тематической обработки;
подсистема визуализации.



Висновки



- Прогнозні моделі за метеоданими не робастні, а **точність біофізичних та регресійних моделей порівняна**
- Точність прогнозування за часовими рядами FAPAR и VHI вища, ніж за NDVI.
Середньоквадратична помилка прогнозу до 7 ц/га, а середня помилка прогнозу по всіх областях – близько 3 ц/га.
- Інформативними ознаками для території України є значення індексів NDVI та VHI, а також супутникового продукту FAPAR на **кінець квітня – початок травня. Завчасність прогнозу - 2 місяці до збирання врожаю.**
- Розроблена регіональна модель та автоматизована інформаційна технологія прогнозування врожайності на рівні областей, реалізована у вигляді сервісу.





Дякую за увагу!

