УДК 004.9

Колотий А.В., аспирант
Институт космических исследований НАНУ-ГКАУ
andrew.k.911@gmail.com

Выбор информационных признаков для построения прогнозов урожайности озимой пшеницы для территории Украины

В статье проанализирована возможность использования спутниковых данных для прогнозирования урожайности озимой пшеницы в Украине. Приводятся оценки согласованности временных интервалов отбора релевантных информационных признаков среди предикторов различной природы, а также их информативности. В качестве предикторов используются нормализованный разностный вегетационный индекс NDVI, индекс здоровья растительности VHI и продукт FAPAR, характеризующий долю фотосинтетически активной солнечной радиации, поглощенной растительностью.

Ключевые слова: прогноз урожайности, регрессионная модель, информационная технология, информационные признаки, MODIS, NDVI, FAPAR, VHI.

Введение

В свете проблем продовольственной безопасности задача прогнозирования урожайности является актуальной не только для Украины, как одного их крупнейших производителей зерна, но и для международного сообщества [1]. Важным источником данных для решения этой задачи являются спутниковые наблюдения [2, 3]. Анализируя публикации, связанные с прогнозированием урожайности озимой пшеницы в Украине, следует констатировать факт, что большинство работ близкой тематики направлены на прогнозирование урожайности для всей территории страны [4] или для отдельных районов [5]. Проблеме прогнозирования урожайности на уровне отдельных областей, которые являются основными единицами административного деления, уделяется меньше внимания [6, 7]. В работах [5, 8] авторами предложен метод прогнозирования урожайности озимой пшеницы с использованием вегетационного индекса NDVI, построены регрессионные модели и разработаны информационные технологии для прогнозирования урожайности для каждой области, выделены основные информационные признаки, проведено оценивание полученных результатов на данных, не участвовавших в обучении эмпирических линейных регрессионных моделей.

В упомянутых работах предикторы прогнозных моделей выбираются на основе процедуры LOOCV из 16-дневных композитов NDVI, полученных по данным прибора MODIS.

Остаются открытыми вопросы информативности выбранных предикторов и достоверности полученных результатов.

Учитывая появление в свободном доступе временных рядов индексов и продуктов других спутников, возникает задача определения наилучших предикторов для регрессионных моделей прогнозирования урожайности, для решения которой автором будут использованы временные ряды индекса здоровья растительности VHI для территории Украины, полученные с прибора AVHRR, а также продукты FAPAR, построенные по данным Spot Vegetation .

Используемые данные

Коротко рассмотрим физический смысл вегетационных индексов, используемых для верификации полученных ранее результатов.

NDVI (полученный из продукта MOD13) - нормализованный разностный вегетационный индекс, один из наиболее распространенных и используемых индексов для решения задач, использующих количественные оценки растительного покрова.

Рассчитывается по формуле:

$$NDVI = \frac{NIR - RED}{NIR + RED}, \qquad (1)$$

где NIR - отражение в ближней инфракрасной области спектра, RED - отражение в красной области спектра. Согласно (1) плотность растительности (NDVI) в заданной точке изображения равна разности интенсивностей отраженного света в красном и инфракрасном диапазоне, деленной на сумму их интенсивностей.
VHI [9] (полученный с помощью AVHRR - Advance Very High Resolution Radiometer) - вегетационный индекс, основанный на отражении видимого света растительным покровом,

характеризующий здоровье культур. Данный индекс базируется на сочетании индекса VCI (Vegetation Condition Index), который характеризирует угнетённость растительного покрова и индекса температурного режима TCI (Temperature Condition Index), которые были предложены в 1995 году Ф. Коганом:

$$VHI = 0.5*VCI + 0.5*TCI, \quad (2)$$

$$VCI = 100*\frac{(NDVI - NDVI_{\min})}{(NDVI_{\max} - NDVI_{\min})}, \quad (3)$$

$$TCI = 100*(BT_{\max} - BT)/(BT_{\max} - BT_{\min}), \quad (4)$$

где BT, $BT_{\max}$, $BT_{\min}$ - усредненные сезонные значения яркостной температуры, ее абсолютный максимум и минимум соответственно, $NDVI$, $NDVI_{\max}$, $\mathrm{NDVI}_{\min}$ - усредненные значения индекса NDVI [10].

Более подробно применение вегетационного индекса NDVI и индекса здоровья VHI для прогнозирования урожайности описано в работе [11].

Продукт FAPAR [12] (получен с помощью Spot Vegetation) - содержит результаты измерения поглощенной фотосинтетической радиации и фракции абсорбирования, выступает в качестве интегрального показателя состояния и здоровья растительного покрова. Данный продукт играет важную роль при определении первичной продуктивности фитосферы, может быть использован для количественного определения растительности.

Нередко измерения FAPAR происходят во время ясной или малооблачной погоды, что усложняет получение усредненных данных по указанному продукту, рекомендуется использовать измерения, сделанные в пасмурную погоду. FAPAR определяется долей, присущей фотосинтезу активной радиации (ФАР), поглощаемой листовым покровом, а иногда и почвой и определяется следующим уравнением:

$$FAPAR = \frac{(PAR_{\downarrow AC} - PAR_{\uparrow AC})(PAR_{\downarrow BC} - PAR_{\uparrow BC})}{PAR_{\downarrow AC}}, \quad (5)$$

где $PAR_{\downarrow AC}$ и $PAR_{\uparrow AC}$ - поступающий (сверху вниз) и отраженный (снизу вверх) ФАР через лиственный покров, и $PAR_{\downarrow BC}$ и $PAR_{\uparrow BC}$ - соответствующие условия под лиственным покровом.

Из формулы (5) можно получить соотношение (6) между отраженным ФАР под поверхностью листвы (ρ_{BC}) и над нею (ρ_{AC}):

$$FAPAR = [1 - \rho_{AC}(t)] - [1 - \rho_{BC}(t)](PAR_{\downarrow BC} / PAR_{\downarrow AC}) \quad (6)$$

где соотношение $PAR_{\downarrow BC} / PAR_{\downarrow AC} = 1 - FIPAR$, $FIPAR$ - часть ФАР, которая была задержана лиственным покровом.

Во многих измерениях FIPAR используют вместо FAPAR, так как этот продукт легче измеряется и он является полным аналогом продукта FAPAR [13].

Для временных рядов VHI и FAPAR в рамках кросс-валидационной процедуры LOOCV [14] (учитывая нехватку данных обучения) будут отобраны информационные признаки, обеспечивающие наименьшую ошибку прогноза, аналогично тому, как в [5, 8] это делалось для NDVI.

Будет проведена оценка временной согласованности отобранных информационных признаков для построения регрессионных моделей для каждой из областей Украины по различным спутниковым данным (NDVI, VHI, FAPAR).

Методы решения задачи

В качестве предикторов в рассматриваемых регрессионных моделях будем использовать значения индекса NDVI, усредненные на уровне областей по маске посевных территорий для каждого 16-дневного композита, полученные из продукта MOD13 прибора MODIS; значения индекса VHI, полученные с прибора AVHRR, и продукт FAPAR, построенный по данным Spot Vegetation, усредненные на уровне области для каждого еженедельного (VHI) или 10-дневного (FAPAR) композита по маске посевных территорий.

В виду небольшого объема выборки, используемой для обучения регрессионных моделей, для поиска наилучшего предиктора, обеспечивающего минимальное значение среднеквадратической ошибки *RMSE,* воспользуемся процедурой кросс-валидации (leave-one-out cross-validation - *LOOCV*) [14]:

$$RMSE = \sqrt{\frac{1}{n} \cdot \sum_i (P_i - O_i)^2}, \quad (7)$$

где P_i и O_i – прогнозируемое и наблюдаемое (по данным статистики) значения урожайности озимой пшеницы соответственно, n — число лет, данные которых используются для построения модели (например, если для идентификации параметров модели используются данные за 2000-2012 годы, то n=13).

Для более удобного сравнения результатов анализа временной согласованности выбранных информационных признаков по данным различной природы (NDVI, VHI и FAPAR) недели VHI и декады FAPAR приведем к единой временной шкале. Результаты этого сравнения представлены ниже.

Анализ результатов

Рассмотрим представленные в табл. 1-3 результаты поиска временных интервалов, которым соответствуют минимальное значение среднеквадратической ошибки (1) для VHI, NDVI

и FAPAR. Для более комфортного сравнения результатов подбора по еженедельным данным VHI, немного «огрубим» полученные результаты, приведя их к той же временной шкале, что и композиты NDVI.

Таблица 1. Недели, обеспечивающие минимальные ошибки прогноза для VHI.

Область	2010	2011	2012	2013
Винницкая	113	113	145	145
Волынская	113	113	113	113
Днепропетровская	161	161	161	161
Донецкая	113	113	113	113
Житомирская	113	113	113	145
Закарпатская	113	129	129	129
Запорожская	145	145	145	145
Ивано-Франковская	113	113	113	145
Киевская	113	113	113	113
Кировоградская	161	161	161	161
Луганская	113	113	113	113
Львовская	113	113	113	145
Николаевская	145	145	145	145
Одесская	145	129	145	145
Полтавская	113	113	113	161
Ровенская	113	113	113	145
Сумская	113	113	113	113
Тернопольская	113	113	113	113
Харьковская	113	113	113	113
Херсонская	145	145	145	145
Хмельницкая	113	113	113	145
Черкасская	113	113	113	113
Черновицкая	113	113	113	145
Черниговская	113	113	113	27
Республика Крым	129	129	129	129

В таблице 1 приведены номера недель, за которые следует брать значения VHI для соответствующей области для обеспечения минимальной ошибки прогноза. Результаты приведены для моделей, обученных на данных 2000-2009, 2000-2010, 2000-2011 и 2000-2012 годов.

Для большей части регионов Украины отобранные информативные предикторы для прогнозирования урожайности по индексу вегетационного здоровья VHI приходятся на 2-ю либо 3-ю декады апреля, временами на 1-ю декаду мая. Аналогичные результаты для нормализованного вегетационного индекса NDVI [5, 8] представлены в таблице 2.

Таблица 2. Дни года (DOY), обеспечивающие минимальные ошибки прогноза для NDVI.

Область	2010	2011	2012	2013
Винницкая	97	97	97	97
Волынская	145	97	97	145
Днепропетровская	145	129	145	145
Донецкая	129	129	129	129
Житомирская	97	97	97	97
Закарпатская	97	113	113	113
Запорожская	129	129	129	129
Ивано-Франковская	97	113	97	145
Киевская	97	97	97	97
Кировоградская	129	129	129	129
Луганская	129	129	129	129
Львовская	113	113	113	113
Николаевская	129	129	129	129
Одесская	129	129	129	129
Полтавская	129	129	129	129
Ровенская	97	97	97	97
Сумская	145	97	97	129
Тернопольская	97	113	97	113
Харьковская	129	129	129	129
Херсонская	129	129	129	129
Хмельницкая	97	97	97	113
Черкасская	129	129	129	129
Черновицкая	97	113	97	113
Черниговская	97	97	97	97
Республика Крым	129	129	129	129

Как видно из таблиц 1 и 2 для большей части областей Украины временные интервалы выбора информационных предикторов для построения регрессионных моделей совпадают.

Аналогичное исследование было проведено для продукта FAPAR, получаемого на основе данных Spot Vegetation – таблица 3.

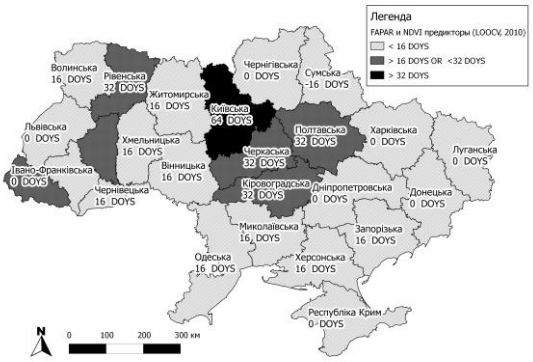
Таблица 3. Дни года (DOY), обеспечивающие минимальные ошибки прогноза для FAPAR.

Область	2010	2011	2012	2013
Винницкая	113	113	129	129
Волынская	113	97	97	129
Днепропетровская	145	145	145	145
Донецкая	129	129	129	129
Житомирская	113	97	97	129
Закарпатская	129	97	97	97
Запорожская	145	129	129	145
Ивано-Франковская	97	97	97	97
Киевская	113	97	97	97
Кировоградская	113	113	113	113
Луганская	129	129	129	129
Львовская	113	129	129	129

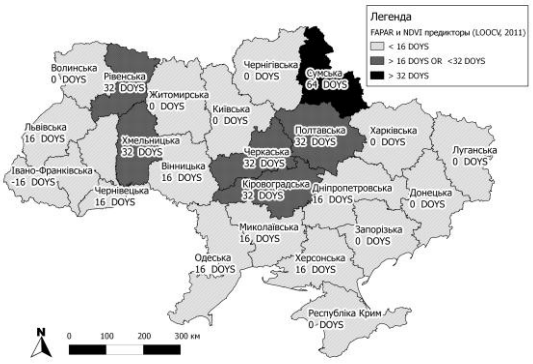
Николаевская	145	145	145	145
Одесская	145	145	145	145
Полтавская	113	113	113	113
Ровенская	129	129	113	113
Сумская	129	113	113	113
Тернопольская	129	129	129	129
Харьковская	129	129	129	129
Херсонская	145	145	145	145
Хмельницкая	113	129	113	113
Черкасская	113	113	113	113
Черновицкая	113	129	129	113
Черниговская	97	97	97	97
Республика Крым	129	129	129	145

Для большей части территории Украины временные интервалы выбора информационных признаков для рядов NDVI и FAPAR совпадают с точностью до ±16 дней.

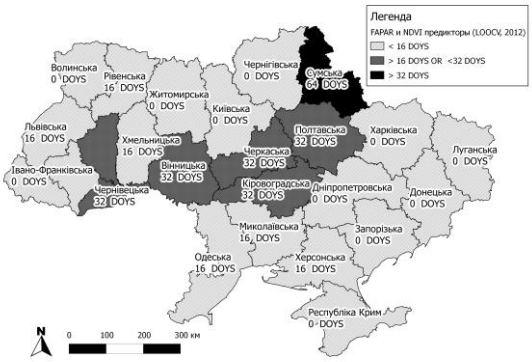
Более наглядно информация из таблиц 1-3 представлена на рис. 1 (а - г).



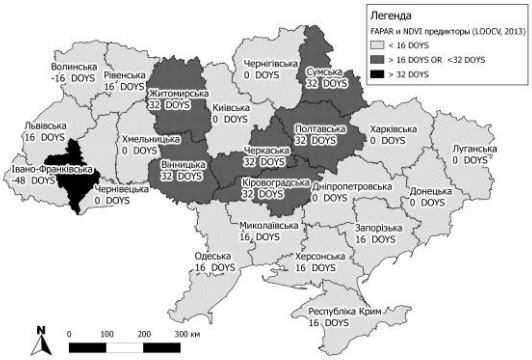
а) Обучение на данных 2000-2009 гг.



б) Обучение на данных 2000-2010 гг.



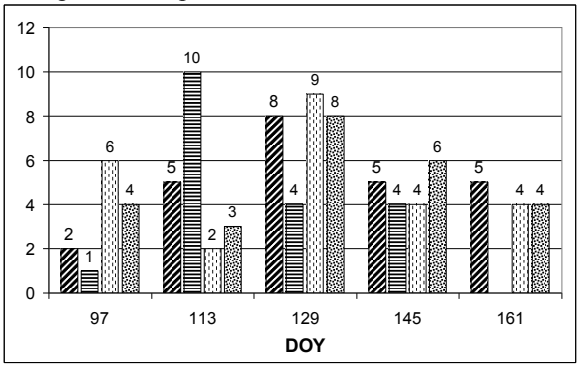
в) Обучение на данных 2000-2011 гг.



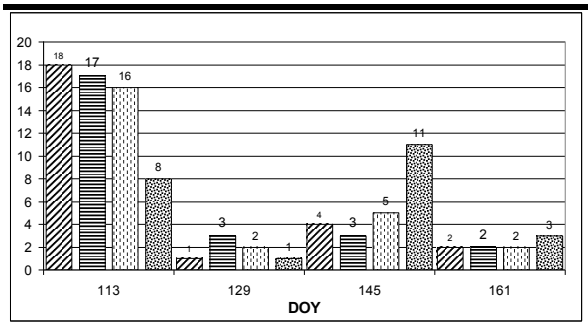
г) Обучение на данных 2000-2011 гг.

Рисунок 1 - Временная согласованность предикторов (FAPAR и NDVI)

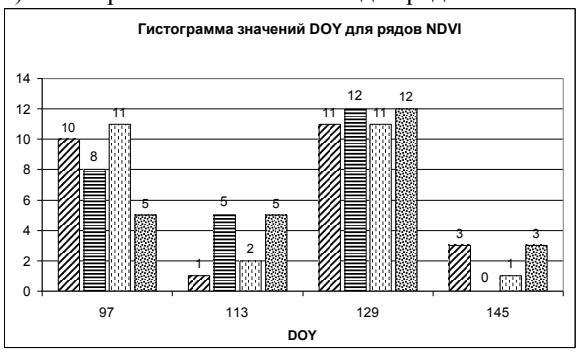
Распределение DOY для рядов FAPAR для областей Украины при разных обучающих выборках показано на рис. 2а. Для большей части областей Украины предиктором для построения регрессионной модели прогнозирования урожайности по продукту FAPAR является значение данного продукта за 113-й (последняя декада апреля) день года или 129-й (первая декада мая) день года. Аналогичные результаты были получены и по данным VHI (при условии выбора единственного информационного признака) – рис. 2б. На рис. 2в показаны аналогичные результаты для временных рядов NDVI.



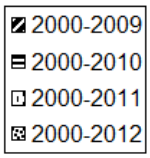
а) Гистограмма значений DOY для продукта FAPAR



б) Гистограмма значений DOY для рядов VHI



в) Гистограмма значений DOY для рядов NDVI



г) Легенда

Рисунок 2 - Распределение DOY для различных спутниковых данных

Выводы

В рамках проведенного исследования была проанализирована возможность использования различных источников спутниковых данных для прогнозирования урожайности сельскохозяйственных культур (на примере озимой пшеницы). Были рассмотрены временные ряды индекса вегетационного здоровья (VHI), и показатель поглощения солнечной радиации поверхностью листа (FAPAR) и оценена временная согласованность.

Как показали результаты проведенного исследования, при использовании различных источников информации для построения прогнозов время для построения прогноза выпадает на конец апреля – начало мая (113-й и 129-й день года для рядов FAPAR, преимущественно 113-й для рядов VHI, 97-й и 113-й день для рядов NDVI).

Информационные признаки, определенные в исследованиях [5, 8], в целом соответствуют результатам, полученным для временных рядов VHI и продукта FAPAR. Все это позволяет утверждать, что достаточно надежный прогноз урожайности для большей части областей Украины может быть построен состоянием на начало – конец мая, что является достаточно заблаговременным для принятия решений по обеспечению продовольственной безопасности на общегосударственном уровне.

Кроме того, автору кажется целесообразным не ограничиваться одной лишь LOOCV-процедурой, но также воспользоваться современными методами машинного обучения для поиска релевантных информационных признаков (предикторов) на примере Random Forest алгоритма [15]. Для больших объемов входных данных такие алгоритмы целесообразно использовать на высокопроизводительных системах [16, 17, 18, 19, 20]

Список литературы

1. Галлего Х., Кравченко А.Н., Куссуль Н.Н., Скакун С.В., Шелестов А.Ю., Грипич Ю.В.. Анализ эффективности различных подходов для классификации посевов на основе спутниковой и наземной информации // Проблемы управления и информатики. – 2012. – №3. – С. 123-134.
2. Kussul N., Shelestov A., Skakun S., Kravchenko O., Moloshnii B. Crop state and area estimation in Ukraine based on remote and in-situ observations //
Int. J. on Information Models and Analyses. – 2012. –
vol. 1, no. 3. – P. 251-259.
3. Kussul, N, Skakun, S, Shelestov, A, Kravchenko, O, Gallego, J & Kussul, O. Crop area estimation in Ukraine using satellite data within the MARS project // IEEE International Geoscience and Remote Sensing Symposium, IEEE, Munich, Germany. — 2012. — P. 3756-3759.
4. Kogan F., Menzhulin G., Shamshurina N., Pavlovsky A. New regression models for prediction of grain yield anomalies from satellite-based vegetation health indices // In "Use of Satellite and In-situ Data to Improve Sustainability" (Eds.) F. Kogan, A. Powell, O. Fedorov, Berlin: Springer-Verlag, 2011. — P.105–112.
5. Куссуль Н.Н., Кравченко А.Н., Скакун С.В., Адаменко Т.И., Шелестов А.Ю., Колотий А.В., Грипич Ю.А. Регрессионные модели оценки урожайности сельскохозяйственных культур по данным MODIS // Сборник научных статей "Современные проблемы дистанционного зондирования Земли из космоса". — 2012. — Том 9, №1. — С. 95–107.
6. Gallego, J., Kravchenko, A.N., Kussul, N.N., Skakun, S.V., Shelestov, A.Yu. Efficiency assessment of

different approaches to crop classification based on satellite and ground observations // Journal of Automation and Information Sciences . — 2012. — vol. 44, no. 5. — P. 67-80.
7. Kussul N.N., Sokolov B.V., Zyelyk Y.I., Zelentsov V.A., Skakun S.V., Shelestov A.Y. Disaster Risk Assessment Based on Heterogeneous Geospatial Information // Journal of Automation and Information Sciences. — 2010 , Volume 42, Issue 12, pp. 32-45.
8. Kogan, F., Kussul, N., Adamenko, T., Skakun, S., Kravchenko O., Kryvobok O., Shelestov A., Kolotii A., Kussul O. & Lavrenyuk A. Winter wheat yield forecasting in Ukraine based on Earth observation, meteorological data and biophysical models // International Journal of Applied Earth Observation and Geoinformation — 2013. — vol. 23 — P. 192-203.
9. http://www.star.nesdis.noaa.gov/smcd/emb/vci/VH
10. C. Bhuiyan. Various Drought Indices For Monitoring Drought Condition In Aravalli Terrain Of India.// International Journal of Applied Earth Observation and Geoinformation 8 — 2006. — P. 289-302
11. Kogan F., Salazar L., Roytman L. Forecasting crop production using satellite-based vegetation health indices in Kansas, USA // International Journal of Remote Sensing. — 2012. — 33, N 9. — P. 2798–2814
12. http://www.geoland2.eu/portal/service/ShowServiceInfo.do?serviceId=BB808F80
13. Gower, Stith T.; Kucharik, Chris J. and Norman, John M. Direct and indirect estimation of leaf area index, fAPAR, and net primary production of terrestrial ecosystems. // Remote Sensing of Environment. – 1999. – vol. 70, issue 1. – P.29-51. Reprint #4023.. P. 36
14. Jansen, M.J.W. Validation of CGMS // Workshop for Central and Eastern Europe on agrometeorological models: theory and applications in the MARS project, Ispra, Italy, 1994 (Eds.) J.F. Dallemand, P. Vossen, Luxembourg: Office for Off. Publ. of the EU, 1994. — P. 159-1130.
15. Leo Breiman. Random Forests // Machine Learning . — 2001. — vol. 45, issue 1. — P. 5-32.
16. Kussul N., Shelestov A., Skakun S. Grid technologies for satellite data processing and management within international disaster monitoring projects // Grid and Cloud Database Management, 2011. P. 279-305.
17. Kussul N., Shelestov A., Skakun S, Kravchenko O. High-performance intelligent computations for environmental and disaster monitoring// Int. J. Information Technologies & Knowledge — 2009. — 3, 135-156.
18. Kussul N., Shelestov A., Skakun S., Li G., Kussul O. The Wide Area Grid Testbed for Flood Monitoring Using Earth Observation Data // IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing – 2012.- vol. 5, no. 6. – P. 1746-1751.
19. Kussul N., Mandl D., Moe K., Mund J.P., Post J., Shelestov A., Skakun S., Szarzynski J., Van Langenhove G., Handy M. Interoperable Infrastructure for Flood Monitoring: SensorWeb, Grid and Cloud // IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 2012 vol. 5. No. 6. P. 1740-1745.
20. AN Kravchenko, NN Kussul, EA Lupian, VP Savorsky, L Hluchy, AY Shelestov Water resource quality monitoring using heterogeneous data and high-performance computations // Cybernetics and Systems Analysis. — 2008. — 44 (4). – P.616-624

Надійшла до редакції 10.09.2013

А.В. КОЛОТІЙ
Інститут космічних досліджень НАНУ-ДКАУ;

ВИБІР ІНФОРМАЦІЙНИХ ОЗНАК ДЛЯ ПОБУДОВИ ПРОГНОЗІВ ВРОЖАЙНОСТІ ОЗИМОЇ ПШЕНИЦІ ДЛЯ ТЕРИТОРІЇ УКРАЇНИ

У статті проаналізовано можливість використання супутникових даних для прогнозування врожайності озимої пшениці в Україні. Наводяться оцінки узгодженості часових інтервалів відбору релевантних інформаційних ознак серед предикторів різної природи, а також їх інформативності. В якості предикторів використовуються нормалізований різницевий вегетаційний індекс NDVI, індекс здоров'я рослинності VHI і продукт FAPAR, що характеризує частку фотосинтетично активної сонячної радіації, поглинутої рослинністю.

Ключові слова: прогноз врожайності, регресійна модель, інформаційна технологія, інформаційні ознаки, MODIS, NDVI, FAPAR, VHI.

A.V. KOLOTII
Space Research Institute NASU-NSAU

This article aims at the assessment of the possibility of use satellite data to predict the yield of winter wheat in Ukraine. In article we provide estimates of the coherence for time intervals for relevant features of different nature, as well as their informational quality. Among estimated predictors there are normalized difference vegetation index NDVI, vegetation health index VHI and FAPAR product, describing the part of solar photosynthetically active radiation absorbed by vegetation.

Keywords: yield forecast, the regression model, information technology, features, MODIS, NDVI, FAPAR, VHI.