УДК 004.942

РЕГРЕССИОННЫЕ МОДЕЛИ ПРОГНОЗИРОВАНИЯ УРОЖАЙНОСТИ ОЗИМОЙ ПШЕНИЦЫ В УКРАИНЕ

А.В.Колотий
Институт космических исследований НАНУ и ГКАУ,

andrew.k.911@gmail.com

У статті вирішується задача оцінки відносної ефективності використання супутникових даних для прогнозування врожайності озимої пшениці в Україну на рівні окремих областей. Для ідентифікації параметрів моделей врожайності використовуються офіційні статистичні дані по врожайності озимої пшениці на рівні областей за період 2000-2009 рр., валідація моделей виконується на даних 2010 і 2011 року. Отримані результати показали, що при налаштуванні параметрів моделей на даних 2000-2009 років і 2000-2010 років і незалежному тестуванні моделей на даних 2010 і 2011 років відповідно, середньоквадратична помилка прогнозування становить приблизно 6 ц/га.
Ключові слова: прогноз врожайності, регресійна модель, інформаційна технологія, MODIS, NDVI.

In this paper we assess relative efficiency of using satellite data to winter wheat yield forecasting in Ukraine at oblast level, and compare efficiency of using regression and biophysical models to address this problem. For models identification we use official statistical data on winter wheat yield for 2000-2009, validation of models is done on independent data for 2010 and 2011. The achieved results showed that when training models for 2000-2009 and 2000-2010 years and validating for 2010 and 2011 respectively average root mean square error was approximately 0.6 t/ha.
Keywords: yield, the regression model, information technology, MODIS, NDVI.

В статье решается задача оценки относительной эффективности использования спутниковых данных для прогнозирования урожайности озимой пшеницы в Украине на уровне отдельных областей. Для идентификации параметров моделей урожайности используются официальные статистические данные по урожайности озимой пшеницы на уровне областей за период 2000-2009 гг., валидация моделей выполняется на данных 2010 и 2011 года. Полученные результаты показали, что при настройке параметров моделей на данных 2000-2009 годов и 2000-2010 годов и независимом тестировании моделей на данных 2010 и 2011 годов соответственно, среднеквадратическая ошибка прогнозирования составляет примерно 6 ц/га.
Ключевые слова: прогноз урожайности, регрессионная модель, информационная технология, MODIS, NDVI.

Введение

Современный этап развития систем мониторинга характеризуется активными процессами интеграции и глобализации. Наиболее известным из этих процессов является создание международной системы систем наблюдения Земли GEOSS, в рамках которой интегрируются данные дистанционного зондирования, данные in-situ и моделей для создания систем поддержки принятия управленческих решений. Одной из важных задач этой системы является решение проблемы продовольственной безопасности. Решению этой проблемы посвящена инициатива стран Большой двадцатки по созданию Глобальной системы сельскохозяйственного мониторинга GLAM. Задача

проекта состоит в объединении данных спутниковых и наземных наблюдений для мониторинга агрометеорологических параметров, роста сельскохозяйственных культур и динамики изменения водных ресурсов. Понятие агромониторинга включает два основных компонента: оценку посевных площадей и прогнозирование урожайности. Точность прогнозирования урожайности с заблаговременностью несколько месяцев до начала сбора урожая играет важную роль при решении задач глобального, национального и регионального уровней.

Украина является одним из крупнейших производителей сельскохозяйственных культур в мире. Согласно статистическим данным зарубежной сельскохозяйственной службы (FAS — Foreign Agricultural Service) Министерства сельского хозяйства США (USDA — U.S. Department of Agriculture) состоянием на 2011 год, Украина занимает 8-е место среди крупнейших экспортеров и 10-е — среди крупнейших производителей пшеницы в мире. Поэтому своевременный и точный прогноз урожайности в Украине является одним из ключевых элементов для поддержки принятия решений в политике продовольственной безопасности в мире.

Одним из наиболее часто используемых подходов к прогнозированию урожайности является использование эмпирических регрессионных моделей. Эмпирические модели связывают урожайность с некоторыми предикторами (например, характеристиками биомассы, получаемыми со спутников), и, как правило, не требуют больших объемов входных данных. Такие модели являются достаточно простыми в реализации, однако они недостаточно робастны. Эффективность таких моделей в значительной степени зависит от наличия и качества данных. Тем не менее, в работе [1] предложена обобщенная регрессионная модель для прогнозирования урожайности озимой пшеницы, которая была изначально построена и верифицирована для штата Канзас, США, а затем без адаптации использована для прогнозирования урожайности в Украине. Эта модель обеспечивает прогноз урожая озимой пшеницы в Украине с погрешностью 10% относительно официальной статистики.

Спутниковые данные играют важную роль в прогнозировании урожайности, так как они являются источником своевременной и объективной информации для больших территорий. Спутниковые продукты, такие как лиственный индекс LAI (Leaf Area Index), вегетационные индексы NDVI (Normalized Difference Vegetation Index) и EVI (Enhanced Vegetation Index), метеорологические параметры, получаемые со спутников Meteosat и NOAA-AVHRR, можно ассимилировать в модели роста растительности для повышения их эффективности [2]. Их также можно использовать в качестве предикторов в эмпирических регрессионных моделях. В частности, в работах [3] описано применение вегетационного индекса NDVI и индекса здоровья растительности VHI (Vegetation Health Index) для прогнозирования урожайности.

1. Постановка задачи

Пшеница является одной из самых важных сельскохозяйственных культур в Украине. За последние пять лет площадь пшеницы составляет в среднем 48% от всех засеянных площадей и около 45% валового сбора всех зерновых культур.

Проблеме прогнозирования урожайности озимой пшеницы в Украине с использованием спутниковых данных посвящено достаточно много работ. Детальный анализ направлений использования аэрокосмических методов при решении сельскохозяйственных задач в Украине приведен в работе [4]. Однако известные работы были в основном направлены на прогнозирование урожайности в масштабе страны [5] или иллюстрацию возможностей прогнозирования для отдельных районов или полей [6]. Прогнозированию урожайности на уровне отдельных областей уделялось меньше внимания [7, 8].

Поэтому в данной статье ставится задача оценить относительную эффективность использования спутниковых данных для прогнозирования урожайности озимой пшеницы в Украине на уровне отдельных областей.

В качестве эмпирических моделей урожайности будут рассмотрены линейные регрессионные модели зависимости урожайности от 16-дневного композита индекса NDVI на основе данных MODIS с разрешением 250 м (MOD13).

Для идентификации параметров моделей урожайности будут использованы официальные статистические данные по урожайности озимой пшеницы на уровне областей за период 2000-2009 гг. Валидация моделей будет выполнена на данных 2010 и 2011 года.

2. Решение задачи прогнозирования урожайности

Линейные регрессионные модели будем строить для каждой области в отдельности. В качестве предиктора модели будем использовать значения индекса NDVI, полученные из продукта MOD13 и усредненные на уровне области для каждого 16-дневного композита по маске посевных территорий, построенной на основе карты классификации земного покрова ESA GLOBCOVER с пространственным разрешением 300 метров.

Будем учитывать тот факт, что за последнее десятилетие для урожайности озимой пшеницы наблюдается положительный линейный тренд для всех областей в связи с улучшением сельскохозяйственных технологий [3]. Таким образом, урожайность аппроксимируется следующим уравнением:

$$Y_i = T_i + dY_i, \qquad (1)$$

где Y_i – урожайность в год i, T_i – трендовый компонент, который связан с сельскохозяйственными технологиями, dY_i – случайный компонент, который

обусловлен метеорологическими условиями и состоянием посевов в текущем году i.

Трендовый компонент будем аппроксимировать с помощью линейной регрессии [3]:

$$T_i = a_0 + a_1 * i\,, \tag{2}$$

где i – год.

Построим регрессионную модель, связывающую отклонение урожайности от тренда со значением вегетационного индекса NDVI:

$$dY_i = Y_i - T_i = f\left(NDVI_i\right) = b_0 + b_1 * NDVI_i\,, \tag{3}$$

где $NDVI_i$ – 16-дневный композит значений индекса NDVI за некоторый период года i. В качестве предиктора в регрессионной модели (3) будем использовать 16-дневный композит NDVI за тот промежуток времени, который обеспечивает минимальное значение среднеквадратической ошибки *RMSE* на основе процедуры кросс-валидации (leave-one-out cross-validation - LOOCV) [9]:

$$RMSE = \sqrt{\frac{1}{n}\sum_i \left(P_i - O_i\right)^2}\,, \tag{4}$$

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 (3) используются данные за 2000-2009 годы, то n=10).

С целью уменьшения влияния «выбросов» для идентификации параметров модели b_0 и b_1 в регрессионной модели будем использовать робастную линейную регрессию [10].

Для оценки эффективности предложенной модели будем использовать два критерия: поведение модели на независимых данных за 2010-2011 годы и относительную эффективность, оцененную согласно LOOCV-процедуре. Относительная эффективность — это отношение дисперсии исходных данных (в нашем случае - урожайности озимой пшеницы по данным официальной статистики) к дисперсии урожайности, спрогнозированной на основе спутниковых данных:

$$relleff = \frac{V\left(Y_{sample}\right)}{V\left(Y_{satellite}\right)} = \frac{\frac{1}{n}\sum_i \left(dY_i\right)^2}{RMSE^2}\,. \tag{5}$$

Относительная эффективность показывает, во сколько раз использования данных спутниковых наблюдений позволяет уменьшить ошибку прогнозирования урожайности по сравнению с моделью тренда.

3. Полученные результаты

Рассмотрим результаты, полученные для моделей, и оценим эффективность прогнозирования на независимых данных 2010 и 2011 годов, уделив внимание вопросу адекватности полученных моделей.

Определим тренд урожайности озимой пшеницы для каждой из областей и определим соответствующие параметры в уравнении (2). Коэффициенты угла наклона линии тренда урожайности для разных областей, вычисленные на основе статистических данных за 2000-2009 и 2000-2010 годы, представлены в табл. 1.

Таблица 1.

Коэффициент угла наклона линии тренда (a1)

Область	**Коэффициент тренда за 2000-2009 (a_1), ц/га/год**	**Коэффициент тренда за 2000-2010 (a_1), ц/га/год**
Винницкая	1,718	1,397
Волынская	0,307	0,155
Днепропетровская	0,382	0,254
Донецкая	0,67	0,615
Житомирская	1,08	0,795
Закарпатская	0,618	0,051
Запорожская	0,801	0,578
Ивано-Франковская	0,867	0,578
Киевская	1,161	0,533
Кировоградская	0,448	0,351
Луганская	1,03	0,741
Львовская	0,932	0,629
Николаевская	0,521	0,588
Одесская	0,191	0,276
Полтавская	1,821	1,215
Ровненская	0,545	0,489
Сумская	1,467	0,881
Тернопольская	1,544	1,065
Харьковская	1,236	0,519
Херсонская	0,57	0,465
Хмельницкая	0,945	0,796
Черкасская	1,707	1,321
Черновицкая	1,476	1,126
Черниговская	1,967	1,295
Автономная республика Крым	0,575	0,384

Как видно из табл. 1, добавление одной точки наблюдений (урожайности за 2010 год) существенно влияет на угол наклона линии тренда, поскольку для построения регрессии имеется слишком мало данных.

В качестве предиктора линейной регрессионной модели ($NDVI_i$, в уравнении (3)) для каждой области был выбран 16-дневный композит за тот период (обозначим его DOY — Day Of the Year), для которого корреляция NDVI с урожайностью является максимальной (с учетом процедуры кроссвалидации). На рис. 2 представлена гистограмма значений DOY, обеспечивающих минимизацию среднеквадратической ошибки в уравнении (4).

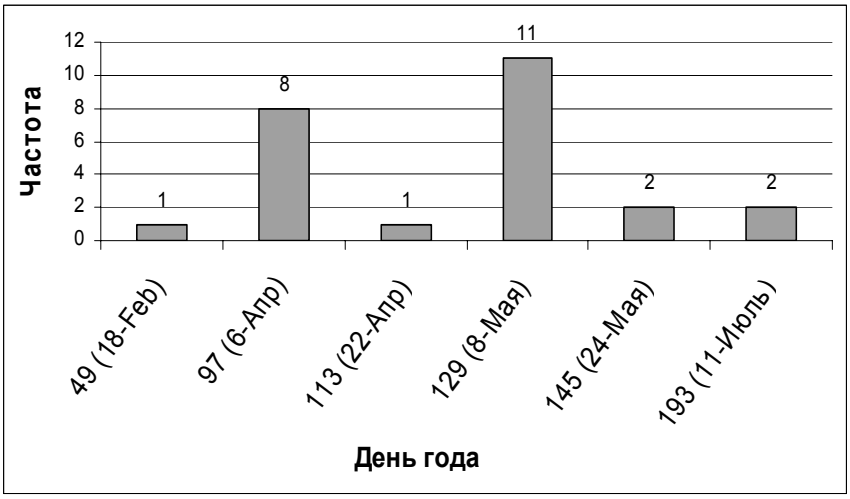
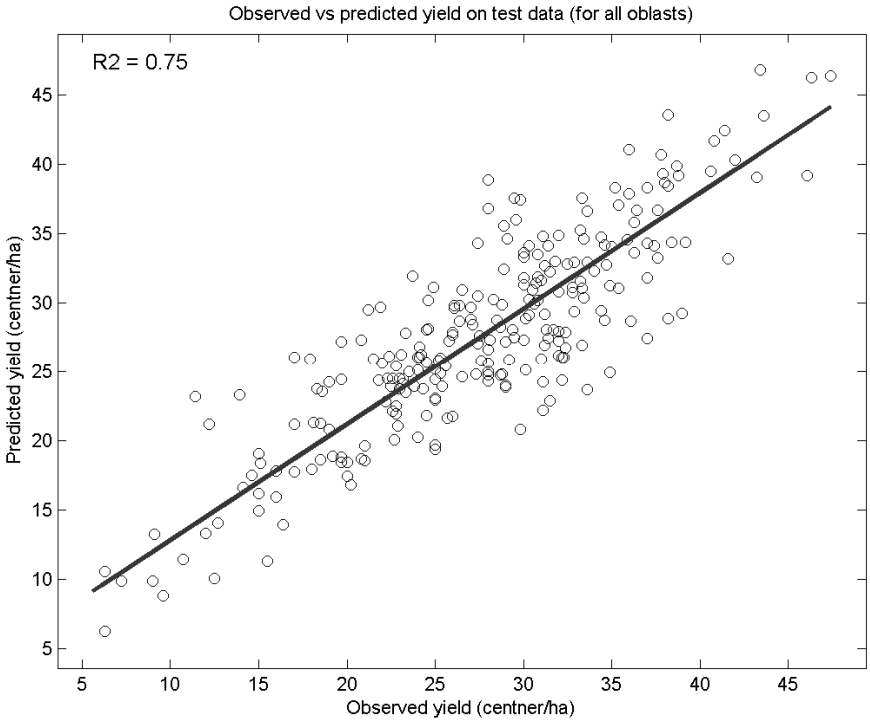


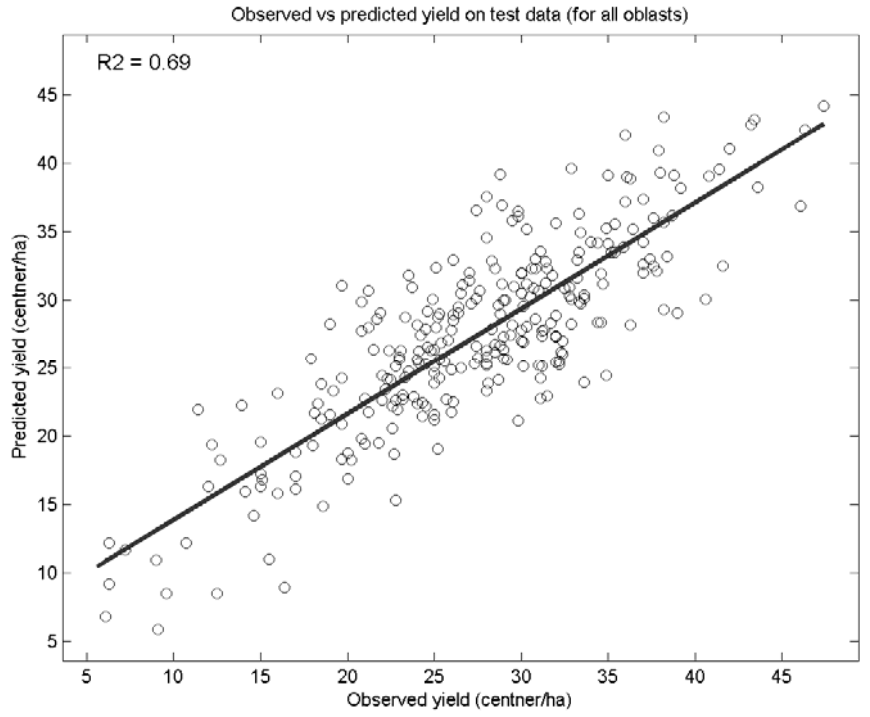
Рис. 2. Гистограмма значений DOY, используемых в качестве предиктора в регрессионной модели (3)

Как видно из рис. 2, наиболее информативными являются значения NDVI, полученные за 16-дневный период, начиная со 129 дня года (с 8 по 25 мая) и с 97 дня года (с 6 по 22 апреля), в зависимости от области, то есть за 2-3 месяца до сбора урожая.

На рис. 3 представлено сравнение фактической и прогнозируемой урожайности озимой пшеницы в рамках LOOCV-процедуры (для моделей 2000-2009 гг. и 2000-2010 гг.). Для каждой области строится модель на основе данных за все года, кроме одного. На основе построенной модели, выполняется прогноз на этот год, который сравнивается с фактическим значением. Таким образом, при использовании данных за 2000-2009 гг. для каждой области имеется 10 сравнений (учитывая, что количество областей 25, всего 250 точек), а при использовании данных за 2000-2010 гг. — 11 (275 точек). Эти точки представлены на рис. 3а и 3б соответственно. Коэффициент детерминации в обоих случаях оказался достаточно большим (0,75 и 0,69), что соответствует о робастности модели на данных обучения, как за 2000-2009 гг., так и за 2000-2010 гг.



(а)



(б)

Рис. 3. Прогнозируемая и фактическая урожайность озимой пшеницы для всех областей в рамках LOOCV-процедуры при построении модели за 2000-2009 гг. (а) и за 2000-2010 гг. (б)

На рис. 4 представлено сравнение результатов прогнозирования урожайности озимой пшеницы по предложенному методу на 2012 год с фактической урожайностью по данным экспресс-статистики.

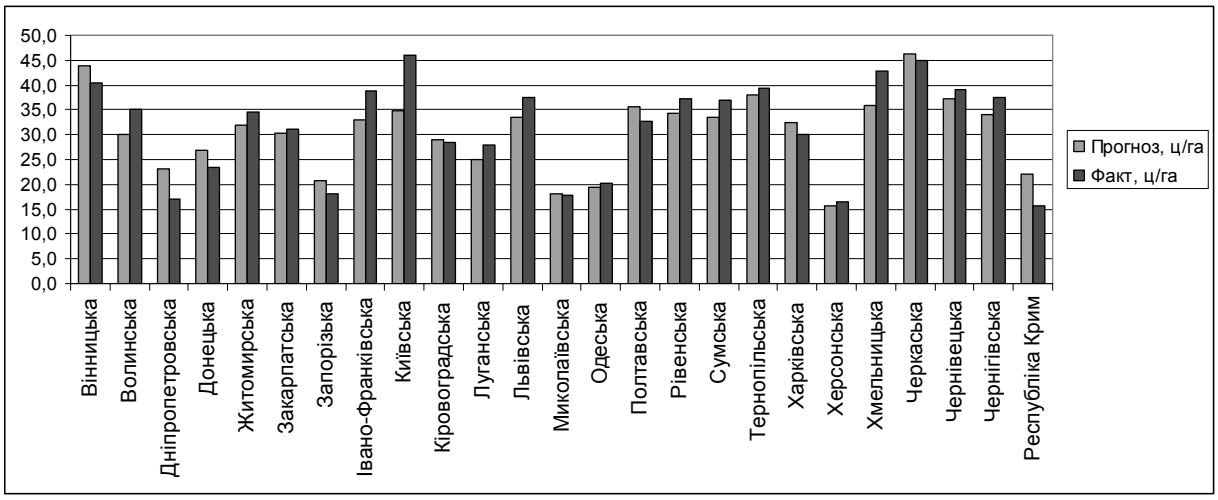


Рис. 4. Спрогнозированная (от 02.06.12) и фактическая (экспресс-статистика) урожайность озимой пшеницы для всех областей Украины на 2012 год

Таблица 2.

Относительная эффективность и коэффициенты детерминации регрессионных моделей, усредненных по агроклиматическим зонам

	Модель, построенная на данных 2000-2009 годов		**Модель, построенная на данных 2000-2010 годов**	
Агроклиматическая зона	**Отн. эфф.**	**R^2**	**Отн. эфф.**	**R^2**
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

Как видно из табл. 2, наиболее эффективным является использование спутниковых данных для прогноза урожайности озимой пшеницы в степной и лесостепной зонах. Построенные для этих зон модели достаточно надежны, о чем свидетельствует высокий коэффициент детерминации. В Полесье применение спутниковых данных менее эффективно. Это связано с тем, что в этой зоне сельское хозяйство менее развито по сравнению с другими зонами. Эта зона характеризуется более сложным ландшафтом, наличием лесных массивов, природных лугов и заброшенных полей, и меньшими площадями озимых культур. Поэтому усреднение значений NDVI по маске посевных территорий представляет собой усреднение по смешанным пикселям, что не в полной мере отображает состояние озимых культур. Таким образом, качество

прогноза по спутниковым данным для Полесья не намного выше качества прогноза по тренду.

В то же время, из табл. 2 видно, что построенные модели достаточно робастны, поскольку добавление данных 2010 года не значительно влияет на коэффициенты детерминации и относительную эффективность модели.

4. Оценка адекватности полученных моделей

Адекватность построенных моделей проверялась с помощью процедуры кросс-валидации LOOCV. Данная процедура является классическим методом машинного обучения проверки адекватности моделей, который основан на использовании независимых наборов данных [11]. Дополнительно использовались классические статистические методы – оценивание F-критерия Фишера и оценка значимости коэффициентов регрессии [11]. Для всех областей, кроме Волынской, Закарпатской и Львовской все коэффициенты регрессии являются значимыми, а сами модели – адекватны по F-критерию Фишера.

5. Выводы

Таким образом, в работе рассмотрен вопрос прогнозирования урожайности с использованием линейных регрессионных моделей на основе спутниковых данных. При настройке параметров моделей на данных 2000-2009 годов и 2000-2010 годов и независимом тестировании моделей на данных 2010 и 2011 годов соответственно среднеквадратическая ошибка прогнозирования составляет примерно 6 ц/га. Стоит также отметить, что модели с использованием спутниковых данных оказались достаточно робастны: добавление данных за 2010 год не значительно влияет на коэффициенты детерминации и относительную эффективность модели. Также был выполнен прогноз урожайности на 2012 год от 2 июня 2012 года, результаты его сравнения в данными экспресс-статистики по урожайности 2012 года также свидетельствует о робастности предложенного метода. Разработанные алгоритмы могут быть реализованы на базе высокопроизводительных систем [12 -13].

Литература

1. Becker-Reshef I., Vermote E., Lindeman M., Justice C. A generalized regression-based model for forecasting winter wheat yields in Kansas and Ukraine using MODIS data // Remote Sensing of Environment. — 2010. — **114**, N 6. — P. 1312–1323.
2. de Wit A.J.W., van Diepen C.A. Crop growth modelling and crop yield forecasting using satellite-derived meteorological inputs // International Journal of Applied Earth Observation and Geoinformation. — 2008. — **10**. — P. 414–425.

3. 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.
4. Лялько В.И., Сахацкий А.И., Жолобак Г.М., Попов М.А. Некоторые направления использования аэрокосмических методов при решении сельскохозяйственных задач в Украине // Сборник научных статей "Современные проблемы дистанционного зондирования Земли из космоса". — 2010. — Том 7, №1. — С. 19-28.
5. 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.
6. 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.
7. 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.
8. Gallego, J. and Kravchenko, A.N. and Kussul, N.N. and Skakun, S.V. and 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.
9. 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. 159-170.
10. Street J.O., Carroll R.J., Ruppert D. A Note on Computing Robust Regression Estimates via Iteratively Reweighted Least Squares // The American Statistician. — 1988. — **42**. — P. 152–154.
11. Дрейпер Р., Смит Г. Прикладной регрессионный анализ: В 2-х кн. Кн. 1// Пер. с англ.— 2-е изд.— М: Финансы и статистика, 1986.— 366 с., Кн. 2// Пер. с англ.— 2-е изд.— М: Финансы и статистика, 1987.— 351 с.
12. Shelestov, A.Yu. and Kussul, N.N. and Skakun, S.V. Grid technologies in monitoring systems based on satellite data // Journal of Automation and Information Sciences. — 2006. — vol. 38, no. 3. — P. 69-80.
13. Kussul, N. and Shelestov, A. and Skakun, S. Grid and sensor web technologies for environmental monitoring // Earth Science Informatics. — 2009. — vol. 2, no. 1-2. — P. 37-51.