

DOI: 10.26896/1028-6861-2019-85-2-73-78

АЛГОРИТМЫ КЛАССИФИКАЦИИ ДЛЯ КОНТРОЛЯ КАЧЕСТВА ПРОМЫШЛЕННО ПРОИЗВОДИМЫХ МИНЕРАЛЬНЫХ УДОБРЕНИЙ

© Дмитрий Валерьевич Юновидов^{1,2}, Валерий Васильевич Соколов^{1,2}, Алексей Сергеевич Бахвалов³

¹ Череповецкий государственный университет, г. Череповец, Россия; e-mail: Dm.Yunovidov@gmail.com

² АО «Научно-исследовательский институт по удобрениям и инсектофунгицидам им. проф. Я. В. Самойлова» (АО «НИУИФ»), г. Череповец, Россия.

³ АО «Научные приборы», г. Санкт-Петербург, Россия.

*Статья поступила 13 июня 2018 г. Поступила после доработки 14 ноября 2018 г.
Принята к публикации 20 декабря 2018 г.*

Проведено сравнение различных алгоритмов классификации для динамического описания качества промышленно производимых минеральных удобрений. Проанализировано пять брендов удобрений (NPK 16-16-8, NP(S) 12-40(10), NPK(S) 4-30-15(16), NPK(S) 0-20-20(5), МАФ (NP) 12-52). Сбор данных осуществляли с использованием визуального оптического и рентгенофлуоресцентного методов контроля. Описан способ выделения характеристических свойств промышленно производимых минеральных удобрений с построением матрицы «объекты-признаки». Приведены оценки качества определения физико-химических параметров удобрений (содержания N, P, K и общего содержания S, типа крупности запрессованных частиц и наличия предварительной сушки). Показана процедура нахождения оптимальных параметров каждого из исследованных алгоритмов классификации. Работа алгоритмов оценена с использованием F-меры (гармоническое среднее точности и полноты классификации). Обучение алгоритмов классификации и оценку качества их работы проводили по стратегии кросс-валидации на 10 наборах данных (фолдах) с отложенной выборкой (30 % от общего объема данных). Итоговое значение метрики качества рассчитывали как среднее по всем классам. Описано использование процедуры проекции на две главные компоненты для информативного представления исследуемых объектов на плоскости. В результате предложен комбинированный метод анализа для автоматизированного и информативного исследования производимой продукции.

Ключевые слова: многомерная классификация; проекция на главные компоненты; качество работы алгоритмов; энергодисперсионный рентгенофлуоресцентный анализ; визуальный оптический контроль поверхности.

CLASSIFICATION ALGORITHMS FOR QUALITY CONTROL OF INDUSTRIALLY MANUFACTURED MINERAL FERTILIZERS

© Dmitry V. Yunovidov^{1,2}, Valery V. Sokolov^{1,2}, Aleksey S. Bakhvalov³

¹ Cherepovets State University, Cherepovets, Russia; e-mail: Dm.Yunovidov@gmail.com

² JSC “The Research Institute for Fertilizers and Insectofungicides” (JSC “NIUIF”), Cherepovets, Russia.

³ JSC “Science Instruments”, St. Petersburg, Russia.

Received June 13, 2018. Revised November 14, 2018. Accepted December 20, 2018.

A comparison of different classification algorithms for dynamic description of the quality of industrially produced mineral fertilizers is presented. Five brands of fertilizers (NPK 16-16-8, NP(S) 12-40 (10), NPK(S) 4-30-15 (16), NPK(S) 0-20-20 (5), MAP (NP) 12-52) were analyzed using optical and X-ray fluorescence methods of control. A method of identifying the characteristic features of industrially produced mineral fertilizers using constructed “objects-characteristics” data array is described. Estimates of the quality of determination of physicochemical parameters of fertilizers (the content of N, P, K and total content of S, types of the particle size of the press-fitted granules and the pre-drying conditions) are given. The procedure of choosing optimal parameters for each considered classification algorithm is shown. The algorithms were estimated using the F-measure (harmonic mean of “precision” and “recall” of classifier). The training of the classification algorithms and assessment of the quality of their “work” were carried out according to the cross-validation strategy on 10 data sets (folds) with a delayed test (30% of the total data volume). The total value of the quality metric was calculated as the average for all classes. The use of the principal component analysis for informative representation of the studied objects on the plane is de-

scribed. Finally, a combined method of analysis for automated and informative study of industrial products was proposed.

Keywords: multidimensional classification; principal component analysis; quality of algorithm; energy dispersive X-ray fluorescence analysis; visual optical control of the surface.

Современная промышленность минеральных удобрений постоянно стремится к упрощению и автоматизации контроля качества выпускаемой продукции [1, 2]. Производство нуждается в экспрессной и информативной системе контроля и сортировки готовой продукции как единого целого, объединяя физические и химические свойства в одно понятие — качество [3]. Для этой цели хорошо подходит метод энергодисперсионного (ЭД) рентгенофлуоресцентного анализа (РФА) за счет своей информативности, экспрессности и многоэлементности [4 – 7]. Несмотря на широкую распространенность методов многомерного анализа свойств различных объектов [3, 8 – 11], в промышленности минеральных удобрений данные подходы используют нечасто. В настоящей работе предложена программно-аппаратная установка, объединяющая ЭД РФА и оптическое распознавание поверхности исследуемых объектов, для сбора физико-химических параметров сложных фосфорсодержащих минеральных удобрений. На основании полученных данных построена матрица «объекты-признаки» и исследована работа различных алгоритмов многомерной классификации для мониторинга качества производимых удобрений и качества их пробоподготовки для рентгенофлуоресцентного анализа по таким параметрам, как содержание N, P, K и общей S, тип крупности запрессованных частиц и наличие предварительной сушки удобрений. Дополнительно изучена возможность непрямого определения азота с помощью предложенного подхода.

Методы и объекты исследования. Объектом исследования являются пять брендов сложных фосфорсодержащих минеральных удобрений с известным содержанием питательных элементов и серы (NPK 16-16-8, NP(S) 12-40(10), NPK(S) 4-30-15(16), NPK(S) 0-20-20(5), МАФ (NP) 12-52). Так, тип удобрения отражает наличие тех или иных макро- и микрокомпонентов (например, NP(S + S) + Zn содержит общую и элементную серу и цинк). Марка удобрения отражает содержание компонентов, указанных в типе (12-40(6 + 3) + 1). Бренд удобрения подразумевает тип и марку (например, NPK 16-16-8). В отечественной литературе синонимом бренда часто выступает понятие марки. Предложенная система сбора данных на основе комбинации ЭД РФА и оптического распознавания поверхности может работать как в промышленных, так и в лабораторных

условиях и требует определенной подготовки пробы к анализу. Каждую пробу готовили в виде «сэндвич-структуры»: прессовали на подложке из борной кислоты для трех типов крупности частиц: гранулы, порошок фракции менее 500 мкм и порошок фракции менее 100 мкм. Часть навесок порошковых проб удобрений до прессования сушили при 60° С до постоянной массы. Исследовали также гранулы, насыпанные в кювету с лавсановым дном (толщина дна кюветы — 4 мкм). Все полученные данные (изображения и спектры) заносили в общую матрицу данных «объекты-признаки». Согласно параметрам контроля технологического производства, желательно, чтобы время анализа не превышало 15 мин. Предложенный способ пробоподготовки удовлетворяет данному требованию.

Каждую пробу анализировали с использованием ЭД рентгенофлуоресцентного спектрометра «РЕАН», производства АО «Научные приборы» с молибденовой трубкой (25 кэВ, 100 мкА, 50 с) и делали фотографию поверхности. По фотографии рассчитывали среднюю яркость, количество и площадь определенных аномалий с использованием собственного алгоритма распознавания [11] (рис. 1).

Из полученных спектров извлекали информацию об интенсивности характеристических линий всех обнаруженных элементов (в общем случае присутствовали Si, P, S, K, Cl, Ca, Fe, Sr), интенсивности когерентного и некогерентного рассеяния трубки и площади базовой линии. Для проведения дальнейшего статистического анализа строили матрицу «объекты-признаки» размером порядка 272 × 25.

В работе использовали различные методы многомерной классификации. Для каждого метода находили оптимальные параметры работы с использованием «перебора по сетке» (табл. 1). Все описанные подходы реализованы на языке программирования Python 2.7 [11].

Классификация используется для информативного отслеживания качества выпускаемого продукта и его пробоподготовки. Данная информация особенно актуальна при контроле перехода производства с одного бренда удобрений на другой. Точная и воспроизводимая классификация обеспечивает быстрый и информативный вывод о качестве исследуемой пробы. Для срав-

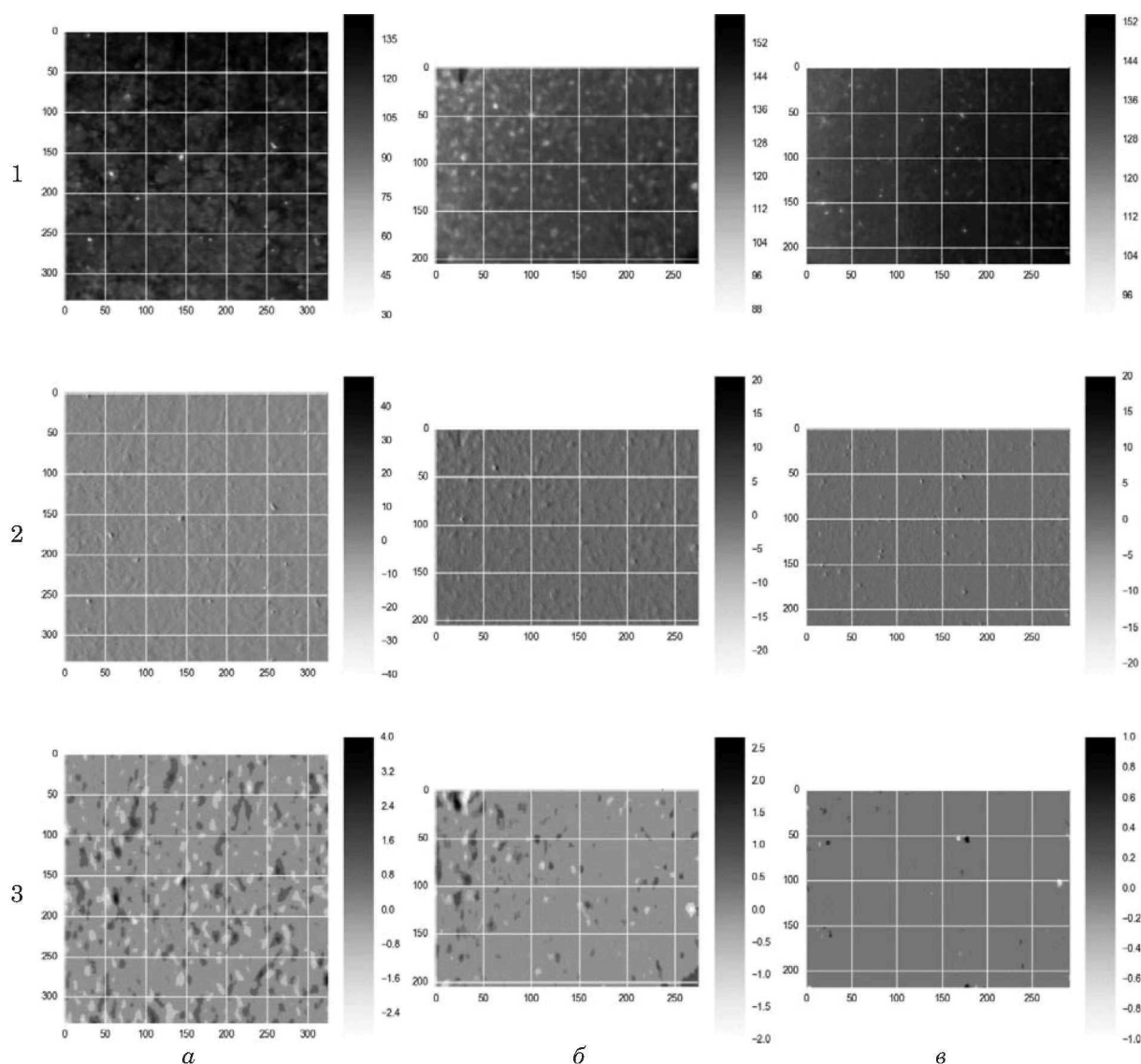


Рис. 1. Пример работы алгоритма: *а* — запрессованные гранулы; *б* — запрессованный порошок <500 мкм; *в* — запрессованный порошок <100 мкм (в 1-м ряду — исходные черно-белые изображения поверхности, во 2-м ряду — поверхность после дифференцирования, в 3-м ряду — карта поверхности после сглаживания медианным фильтром с аномалиями — областями высокой и низкой интенсивности пикселей)

нения различных алгоритмов классификации использовали Ф-метрику качества (табл. 2).

Точность, полноту и Ф-меру рассчитывают по формулам:

$$\text{Точность}(a, X) = \frac{\text{ВП}}{\text{ВП} + \text{НП}}, \quad (1)$$

$$\text{Полнота}(a, X) = \frac{\text{ВП}}{\text{ВП} + \text{НО}}, \quad (2)$$

$$\Phi = \frac{2 \times \text{Точность} \times \text{Полнота}}{\text{Точность} + \text{Полнота}}. \quad (3)$$

Обучение алгоритмов классификации и оценку качества их работы проводили по стратегии

кросс-валидации на 10 наборах данных (фолдах) с отложенной выборкой (30 % от общего объема данных). Итоговое значение метрики качества рассчитывали как среднее по всем классам.

Отдельно стоит отметить минимальную подготовку данных: классы разбалансированы и данные не стандартизованы.

Для информативного обобщенного представления всех исследуемых объектов использовали программную свертку всех признаков каждого из объектов в пространство двух координат (двух главных компонент). Использовали линейный метод проекции дисперсии на две главные компоненты (PCA). Для качественной работы метода данные предварительно стандартизовали на среднее и дисперсию (Z-преобразование). Коли-

чество главных компонент обусловлено дальнейшим представлением общей картины в наиболее понятной и интерпретируемой форме — на плоскости (каждая компонента использована как соответствующая координата).

Результаты и обсуждение. Проанализировано пять типов удобрений с известным содержанием питательных элементов и серы (NPK 16-16-8, NP(S) 12-40(10), NPK(S) 4-30-15(16), NPK(S) 0-20-20(5), МАФ (NP) 12-52). С исполь-

зованием полученной информации создана единая база данных «объекты-признаки» и проведена классификации объектов по основным показателям качества (табл. 3). Для построения регрессионных уравнений использовали сведения о содержании исследуемых элементов в удобрениях. В случае NP(S + S) + Zn удобрения исследовали общее содержание серы (особенность выбранного метода анализа). Полученные данные размещены в открытом доступе [11].

Таблица 1. Алгоритмы обработки данных и их параметры

Алгоритм	Параметр	Значение параметра
Линейный классификатор с градиентным спуском	Функция потерь (loss)	hinge, log, squared_loss, modified_huber
	Количество итераций градиентного спуска (n_iter)	1000 – 15 000
Линейный классификатор с градиентным спуском с L1 или L2 регуляризацией	loss	hinge, log, squared_loss, modified_huber
	n_iter	1000 – 15 000
	Коэффициент перед регуляризацией (alpha)	0,0001 – 1
Классификатор «Случайного Леса»	Количество решающих деревьев (n_estimators)	2 – 100
	Ограничение на максимальное количество свойств (max_features)	None, sqrt, log2
	Ограничение на максимальную глубину решающего дерева (max_depth)	None, 2 – 13
	Бутстрап (bootstrap)	True, False
	Балансировка классов (class_weight)	Balanced, None

Таблица 2. Тип классификационных ошибок

Решение алгоритма на наборе данных X , $a(X)$	Истинные значения для набора данных y	
	1 (принадлежат к данному классу)	–1 (не принадлежат к данному классу)
1 (положительное — отнесено к верному классу)	верное положительное (ВП)	неверное положительное (НП)
–1 (отрицательное — не отнесено)	неверное отрицательное (НО)	верное отрицательное (ВО)

Таблица 3. Качество классификации

Диапазон значений	Содержание, % масс.				Тип крупности*	Влага**
	N	P	K	S общая		
	[0; 16]	[15; 52]	[0; 20]	[0; 20]		[0; 1]
Ф-мера, % (СКО, %)***						
Линейная	99,8 (0,7)	99 (2)	99 (2)	98 (3)	55 (7)	60 (9)
Линейная с L1	99,8 (0,7)	99 (2)	99 (2)	98 (3)	56 (5)	65 (5)
Линейная с L2	99 (1)	99 (2)	99 (2)	96 (4)	35 (11)	51 (10)
«Случайный Лес»	100 (—)****	100 (—)****	99,8 (0,6)	99 (3)	98 (2)	74 (6)

Наиболее важный признак для алгоритма «Случайного Леса» (% объясненной дисперсии)

Cl (15,50)	Ca (14,77)	K (19,95)	Ca (26,12)	Количество аномалий (25,28)	P (26,69)
------------	------------	-----------	------------	-----------------------------	-----------

* Гранулы, таблетки запрессованных гранул и порошков фракции менее 500 мкм и менее 100 мкм.

** Бинарное свойство: 0 — отсутствие предварительного высушивания удобрений, 1 — предварительное высушивание до постоянной массы при 60 °С.

*** Среднеквадратическое отклонение Ф-меры при оценке по стратегии кросс-валидации.

**** На всех десяти отложенных выборках при кросс-валидации выпадало 100 % для всех классов, и СКО не оценивали.

Полученные результаты по выбранным метрикам качества говорят о высоком качестве проведенной классификации (Ф-мера больше 96 %). Более того, созданная база данных позволяет проводить автоматическое распознавание бракованной продукции, обеспечивая более эффективный переход производства с изготовления одного бренда удобрений на другой. Так, с использованием проекции на две главные компоненты по всей совокупности выделенных свойств построена карта всех исследованных удобрений (рис. 2).

Проекция на главные компоненты является очень информативным и полезным методом анализа исследуемых объектов. Так, используя например, евклидову метрику для определения расстояния между новым исследуемым объектом и центрами соответствующих классов, можно количественно оценить близость их свойств (а значит, и качества как единого понятия).

Интересно заметить, что исходные свойства удобрений являются «хорошими» данными (т.е. они содержат мало шумов и хорошо разделимы по классам — имеют явно выраженные свойства для классификации). Стоит также отметить, что для классификации по химическому составу интенсивности линий определяемых элементов зачастую оказываются не так значимы, как интенсивности линий более тяжелых атомов. Это обусловлено сложным составом матрицы исследуемых удобрений, различием их типов и широким диапазоном определяемых содержаний компонентов. При учете матричных влияний классификатор настраивается на наиболее значимо влияющие элементы — обладающие большим коэффициентом поглощения рентгеновского излучения или являющиеся показателями технологического процесса (как, например, кремний или кальций). Тем самым удается нивелировать матричное влияние и обеспечить универсальность классификации для широкого круга объектов, в том числе, произведенных на разных предприятиях по разным технологическим схемам.

В результате были достигнуты необходимые показатели работы алгоритмов по выбранным метрикам качества без предварительной обработки данных (более 96 % по Ф-мере). Наилучший результат показал алгоритм «Случайного Леса», что вполне ожидаемо, поскольку алгоритм является нелинейным и достаточно устойчив к качеству данных и наличию выбросов. Поэтому даже на таких неявных и зашумленных данных, как содержание влаги в удобрениях (обычно не более 1,5 % масс.), «Случайный Лес» показывает высокие результаты (более 70 % по Ф-мере на необработанных данных).

Как следует из представленных данных (см. табл. 3), исследованные алгоритмы позволяют

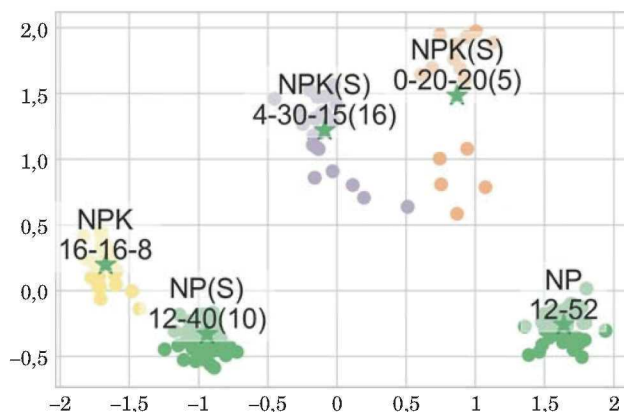


Рис. 2. Проекция дисперсий свойств на две главные компоненты (отложены по осям) для всех исследованных объектов

косвенно определить содержание азота и наличие предварительной сушки удобрений методом ЭД РФА. Хотя точность подхода не оценивали внутри одного класса удобрений, полученные результаты важны для информативного представления промышленных процессов и классификации удобрений по типу, марке и качеству пробоподготовки к РФА (наличие влаги оказывает влияние на качество определения легких элементов).

Таким образом, предложенный комбинированный метод анализа включает ЭД РФА, визуальное оптическое исследование поверхности, собственный алгоритм распознавания изображения и автоматическую многомерную классификацию данных. Разработанная техника позволяет проводить быстрый, дешевый и чрезвычайно информативный анализ промышленно выпускаемой продукции. Использование разработанной схемы обеспечивает экономию ресурсов промышленного комплекса и расширяет область применения ЭД РФА, позволяя определять азот и наличие предварительной сушки минеральных удобрений. Показано, что алгоритм «Случайного Леса», как правило, является более точным и полным при использовании на необработанных данных. Все предложенные алгоритмы и подходы автоматизированы и обеспечивают быструю оценку качества промышленно выпускаемой продукции и ее пробоподготовки. Рассмотренные методы обеспечивают универсальный контроль выбранных параметров качества в широких диапазонах определяемых свойств.

ЛИТЕРАТУРА

1. Remes A., Saloheimo K., Jämsä-Jounela S. L. Effect of speed and accuracy of on-line elemental analysis on flotation control performance / *Miner. Eng.* 2007. Vol. 20. N 11. P. 1055 – 1066.

2. Юновидов Д. В., Соколов В. В., Бахвалова Е. В., Донских В. А. Разработка стандартного образца апатитового концентрата. Эффективный контроль однородности с помощью рентгенофлуоресцентных методов анализа / ГИАБ. 2016. № 7. С. 131 – 144.
3. Offroy M., Duponchel L. Topological data analysis: a promising big data exploration tool in biology, analytical chemistry and physical chemistry / *Anal. Chim. Acta*. 2016. Vol. 910. N 3. P. 1 – 11.
4. Sharma A. et al. Characterizing soils via portable X-ray fluorescence spectrometer: 3. Soil reaction (pH) / *Geoderma*. 2014. Vol. 232 – 234. P. 141 – 147.
5. Юновидов Д. В., Соколов В. В., Бахвалов А. С. Метод оценки влияния стадий пробоподготовки NPKS удобрений на результаты рентгенофлуоресцентного анализа по спектру пробы / *Заводская лаборатория. Диагностика материалов*. 2017. Т. 83. № 9. С. 15 – 21.
6. Angeyo K. H. et al. Feasibility for direct rapid energy dispersive X-ray fluorescence (EDXRF) and scattering analysis of complex matrix liquids by partial least squares / *Appl. Radiat. Isot.* 2012. Vol. 70. N 11. P. 2596 – 2601.
7. Bueno Guerra M. B. et al. Comparison of analytical performance of benchtop and handheld energy dispersive X-ray fluorescence systems for the direct analysis of plant materials / *J. Anal. At. Spectrom.* 2014. Vol. 29. N 9. P. 1667 – 1674.
8. Moran K. Big Data in Chemical Toxicity Research: The Use of High-Throughput Screening Assays To Identify Potential Toxicants / *Chem. Res. Toxicol.* 2014. Vol. 27. N 10. P. 1643 – 1651.
9. Kaniu M. I. et al. Energy dispersive X-ray fluorescence and scattering assessment of soil quality via partial least squares and artificial neural networks analytical modeling approaches / *Talanta*. 2012. Vol. 98. P. 236 – 240.
10. Kunz M. R., Kalivas J. H., Andries E. Model updating for spectral calibration maintenance and transfer using 1-norm variants of tikhonov regularization / *Anal. Chem.* 2010. Vol. 82. N 9. P. 3642 – 3649.
11. <https://github.com/DimYun/DSpectra> (дата обращения 13.06.2018).
- tion control performance / *Miner. Eng.* 2007. Vol. 20. N 11. P. 1055 – 1066.
2. Yunovidov D. V., Sokolov V. V., Bakhvalov A. S., Don-skikh V. A. Develop of reference sample of apatite concentrate. The effective control of homogeneity with X-ray fluorescence / *GIAБ*. 2016. Vol. 7. P. 131 – 144 [in Russian].
3. Offroy M., Duponchel L. Topological data analysis: a promising big data exploration tool in biology, analytical chemistry and physical chemistry / *Anal. Chim. Acta*. 2016. Vol. 910. N 3. P. 1 – 11.
4. Sharma A. et al. Characterizing soils via portable X-ray fluorescence spectrometer: 3. Soil reaction (pH) / *Geoderma*. 2014. Vol. 232 – 234. P. 141 – 147.
5. Yunovidov D. V., Sokolov V. V., Bakhvalov A. S. The Use of the Sample Spectrum for Assessing the Impact of Different Stages of the NPKS Fertilizer Preparation on the Results of X-Ray Fluorescence Analysis / *Zavod. Lab. Diagn. Mater.* 2017. Vol. 83. N 9. P. 15 – 21 [in Russian].
6. Angeyo K. H. et al. Feasibility for direct rapid energy dispersive X-ray fluorescence (EDXRF) and scattering analysis of complex matrix liquids by partial least squares / *Appl. Radiat. Isot.* 2012. Vol. 70. N 11. P. 2596 – 2601.
7. Bueno Guerra M. B. et al. Comparison of analytical performance of benchtop and handheld energy dispersive X-ray fluorescence systems for the direct analysis of plant materials / *J. Anal. At. Spectrom.* 2014. Vol. 29. N 9. P. 1667 – 1674.
8. Moran K. Big Data in Chemical Toxicity Research: The Use of High-Throughput Screening Assays To Identify Potential Toxicants / *Chem. Res. Toxicol.* 2014. Vol. 27. N 10. P. 1643 – 1651.
9. Kaniu M. I. et al. Energy dispersive X-ray fluorescence and scattering assessment of soil quality via partial least squares and artificial neural networks analytical modeling approaches / *Talanta*. 2012. Vol. 98. P. 236 – 240.
10. Kunz M. R., Kalivas J. H., Andries E. Model updating for spectral calibration maintenance and transfer using 1-norm variants of tikhonov regularization / *Anal. Chem.* 2010. Vol. 82. N 9. P. 3642 – 3649.
11. <https://github.com/DimYun/DSpectra> (accessed June 13, 2018).

REFERENCES