УДК 633.11:631.559

Р.И. Ибятов, доктор технических наук, профессор;
Ф.Ш. Шайхутдинов, доктор сельскохозяйственных наук, профессор;
А.А. Валиев, старший преподаватель,
ФГБОУ ВО «Казанский государственный аграрный университет»
(420015, Казань, ул.К. Маркса, 65; pim.kazgau@mail.ru)

АНАЛИЗ УРОЖАЙНОСТИ ЯРОВОЙ ПШЕНИЦЫ МЕТОДОМ ГЛАВНЫХ КОМПОНЕНТ

Работа посвящена применению факторного анализа для уменьшения размерности влияющих факторов на урожайность яровой пшеницы. Для проведения анализа между восьмью факторами наиболее влияющими на урожайность яровой пшеницы, была использована выборка за 32 года. Все данные выборки были предварительно нормализованы, центрованы и представлены в виде таблицы. Вычислительным путем были построены восемь главных компонент, определены факторные нагрузки. По полученным факторным нагрузкам было решено оставить четыре главных компонента, описывающих в сумме 84 процента общей дисперсии. Каждую главную компоненту представили в виде линейной комбинации факторных нагрузок и факторов. Использование главных компонент позволило понизить размерность исходных данных с восьми входных факторов до четырех. Полученная информация была представлена в пространстве главных компонент. Новые координаты опытных данных по урожайности яровой пшеницы вычислены полученными соотношениями. Для поиска скрытых связей между факторами исходные данные были представлены в графическом виде. В двухмерном пространстве количество диаграмм по четырем главным компонентам составляет шесть возможных вариантов, а в трехмерном пространстве четыре. Отдельные точки были поименованы с учетом их вариации по отдельным факторам. Приведена диаграмма данных в плоскости первой и четвертой главных компонент. Расположение точек указывает на то, что большие значения массы зерна связаны с высокими показателями содержания клейковины. Была построена исследовательская модель на базе нейронных сетей типа многослойный персептрон с одним входным, одним выходным и одним скрытым слоем. Нейронная сеть была предварительно обучена по исходным данным и проверена на адекватность.

Ключевые слова: *яровая пшеница, факторный анализ, главная компонента, визуализация данных, нейросетевая модель.*

R.I. Ibyatov, Doctor of Technical Sciences, professor;

F.Sh. Shaykhutdinov, Doctor of Agricultural Sciences, professor;
A.A. Valiev, senior lecturer,
FSBEI HE 'Kazan State Agricultural University'
(420015, Kazan, Karl Marks Str., 65; email: pim.kazgau@mail.ru)

THE ANALYSIS OF SPRING WHEAT PRODUCTIVITY BY THE METHOD OF PRINCIPAL COMPONENT

The work deals with the use of the factor analysis to reduce the importance of factors affecting spring wheat productivity. To carry out the analysis among eight most affecting on crop productivity factors we used the dataset of 32 years. All the dataset has been pre-normalized, being centered and presented in tabular form. Eight principal components were calculated, and the factor loadings were determined. According to the factor loadings it has been decided to take four principal components describing 84% of general dispersion. Each principal component has been presented as a linear combination of factor loadings and factors. The use of principal components allowed reducing the size of initial dataset from eight factors to four ones. The obtained information has been given in the space of principal components. The new coordinates of the experimental dataset on spring wheat productivity has been estimated by the received interdependences. The initial dataset has been given in a graphic form to search latent interdependences among factors. The number of diagrams in four principal components is six variables in a binary space and four ones in tree-dimensional space. It has been given a diagram of the data according to the first and the fourth principal component. The location of the points shows that the largest value of kernel weight is connected with high indexes of gluten content. It has been constructed a model on the basis of neural network (multiclass perceptron) with one input, one output and one latent layer. The neural network has been preliminary studied according to initial dataset and checked for adequacy.

Keyword: *spring wheat, factor analysis, principal component, visualizing of dataset, neural net model.*

Введение. Продуктивность посевов яровой пшеницы является результатом сложного взаимодействия самых разных факторов внешней среды с генетически обусловленными биологическими особенностями выращиваемой культуры. Осуществить оптимизацию условий внешней среды применительно к генетической программе той или иной культуры можно лишь при целенаправленном управлении комплексом факторов [1-3]. С развитием агрохимии, агрометеорологии и вычислительной техники постоянно создавались теоретические и прикладные основы системы управления ростом и развитием растений в посевах с применением современных технических средств и методов обработки информации [4, 5].

Урожайность яровой пшеницы определяется большим количеством факторов. Каждый фактор имеет свое индивидуальное влияние и может быть исследован отдельно. Однако «одномерные методы», учитывающие доли индивидуальных влияний факторов на урожайность, часто оказываются недостаточными для полноценного анализа. Для анализа данных и поиска полезной информации по яровой пшенице можно воспользоваться современными интеллектуальными методами [6-10]. Интеллектуальные методы и подходы позволяют выявить скрытые (латентные) взаимосвязи между множеством входных переменных, что позволяет повысить эффективность моделей за счет отбрасывания малозначащих переменных, а также группировки данных, имеющих скрытые связи. Подобные задачи можно решать на основе метода главных компонент.

Материалы и методы. Закладку опытов проводили в 1982-2013 гг. на опытном поле ФГБОУ ВО «Казанский государственный аграрный университет» по общепринятым при изучении яровых зерновых культур методикам. Обработку данных исследований осуществляли с использованием программ Excel, Deductor Academic.

Объектом исследований были 10 сортов яровой пшеницы, внесенных в Государственный реестр селекционных достижений и допущенных к использованию в седьмом регионе Российской Федерации.

Анализ данных проводили методом главных компонент. Данный метод позволяет выбрать новую систему координат U_j, в которой определенная переменная, соответствующая одной из новых осей, имеет максимальную дисперсию. Каждая следующая координатная ось, то есть главная компонента, должна быть ортогональна предыдущим и «охватывать» как можно большую часть дисперсии.

Главная компонента U_j представляет собой некоторую линейную комбинацию восьми факторов:

$$U_j = \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_8 x_{8j}, j = 1,2,....32 \quad (1)$$

Неизвестные коэффициенты β_i, являются векторами из условия обеспечения наибольшей дисперсии с учетом их ортонормированности. Полученная таким образом условная задача оптимальности решается методом множителей Логранжа. Тогда определение коэффициентов линейной комбинации (1) сводится к нахождению собственных значений и собственных векторов соответствующей ковариационной матрице [6].

Погодные условия за указанный период различались по годам: засушливые и жаркие, сильно засушливые, удовлетворительно увлажненные, теплые оптимально увлажненные и теплые. Метеоусловия всех лет проведения исследования послужили

хорошим фоном для полной оценки влияния отдельных факторов при формировании урожая объекта исследования.

Почва опытного поля Казанского ГАУ – серая лесная среднесуглинистая. Содержание гумуса – 2,8-3,2 %, сумма поглощенных оснований – 26 мг-экв. на 100 г почвы, азота легкогидролизуемого – 8-11, подвижного фосфора по Кирсанову – 163-183, обменного калия по Кирсанову –109-149 мг на 1000 г почвы, pH солевой вытяжки 5,6-5,7.

Результаты. Для проведения анализа урожайности яровой пшеницы были использованы результаты наблюдений за урожайностью в течение 32 лет [8]. Известны средние значения по годам следующих восьми независимых факторов, оказывающих наибольшее влияние на урожайность пшеницы: влажность воздуха, эффективная температура за вегетацию, осадки, вегетационный период, содержание клейковины, масса тысячи зерен, масса зерна с одного колоса, длина стебля. Так как все исследуемые факторы имеют различную единицу измерения и различаются своими порядками, необходимо провести нормирования и центрирования данных. Для нормирования исходных данных они были разделены на свои стандартные отклонения. Затем проводили центрирование путем вычитания их средних значений от факторов. Преобразованные данные урожайности яровой пшеницы представлены в таблице 1.

1. Нормализованные и центрированные данные урожайности яровой пшеницы

№	Урожайность	Влажность воздуха	Эффективная t^0 за вегетацию	Осадки	Вегетационный период,	Содержание клейковины	Масса 1000 зерен	Масса зерна с одного колоса	Длина соломы
1	-1,40	-2,14	-1,12	-1,48	-2,23	-0,20	-1,47	-1,63	-0,51
2	-1,83	-2,14	2,18	-0,54	0,18	-0,34	-1,76	-2,07	-2,01
3	1,84	-1,81	1,98	-2,12	-0,79	0,12	-0,85	0,14	-1,71
4	0,55	-1,48	-0,28	-1,64	-1,75	-0,39	0,19	1,02	-0,51
5	0,04	-1,31	0,60	-0,60	-0,15	0,77	-2,29	-2,18	-1,11
6	0,07	-1,15	0,44	0,56	1,30	-0,58	0,43	0,03	1,58
7	0,64	-0,98	1,40	-0,40	-0,63	-0,09	-0,97	-0,86	-0,21
8	0,62	-0,65	-1,24	0,58	0,34	0,83	1,10	-0,19	1,28
9	0,95	-0,32	0,77	0,96	0,82	0,10	0,27	1,02	1,88
10	1,08	-0,16	-1,01	1,08	-0,79	1,45	0,19	-0,52	1,28
11	0,23	-0,16	-1,29	0,28	-1,43	-0,04	-1,26	-2,40	-1,11

12	0,88	-0,16	-0,20	-0,20	0,18	-0,09	-0,72	0,25	0,69
13	0,39	0,01	-0,69	0,86	-0,47	0,93	0,81	0,58	0,09
14	0,30	0,01	-0,31	-0,87	0,50	0,80	-0,19	0,36	0,09
15	0,49	0,17	-1,89	1,58	-1,27	0,69	-0,68	-0,19	1,88
16	0,20	0,17	0,54	-0,86	-0,63	-0,53	-0,35	0,69	0,09
17	0,42	0,17	-0,54	0,59	-0,15	-0,31	-0,52	0,36	1,28
18	0,02	0,17	1,36	-1,18	0,50	-0,07	-0,27	-0,63	-1,71
19	1,66	0,34	0,29	1,59	0,18	0,31	-0,02	-0,19	-0,51
20	1,79	0,34	0,96	-0,08	0,34	0,39	-0,19	-0,52	0,39
21	0,46	0,50	0,23	0,46	0,66	-4,31	0,19	1,35	0,69
22	1,31	0,50	-1,00	-0,13	-1,43	0,07	0,10	0,36	0,39
23	0,53	0,67	-0,43	0,61	0,98	-0,63	1,92	1,90	0,09
24	0,76	0,67	-0,23	-1,11	-0,47	0,64	0,27	0,80	-0,21
25	0,17	0,67	0,61	1,42	0,18	0,29	1,01	0,14	-0,21
26	0,83	1,00	-0,89	-1,05	0,01	0,39	2,25	1,02	-0,81
27	1,46	1,00	0,64	0,33	1,94	-0,90	1,55	-0,41	-0,51
28	1,72	1,00	-0,90	-0,23	0,01	0,18	0,27	0,03	0,09
29	0,49	1,16	-1,14	1,42	0,50	1,47	0,48	0,36	-0,51
30	1,41	1,33	-0,31	-0,30	2,26	0,72	-0,43	0,80	-0,51
31	-0,47	1,33	0,74	-0,71	0,34	-0,85	0,85	0,14	-0,21
32	0,77	1,33	0,69	1,12	0,98	-0,69	-0,02	0,58	0,69

Вычислительным путем были построены восемь главных компонент (ГК), определены факторные нагрузки и объясненная ими дисперсия:

- на первую главную компоненту (ГК1) оказывают существенное влияние следующие факторы: «масса 1000 зерен», «влажность воздуха» и «масса зерна с одного колоса», причем доля дисперсии, объясненная ГК1, равна приблизительно 37%;

- на вторую главную компоненту (ГК2) оказывают высокую нагрузку переменные «эффективная температура», «содержание клейковины» и «вегетационный период»; доля дисперсии, объясненная ГК2, равна приблизительно 21%;

- на третью главную компоненту (ГК3) – «осадки»; доля дисперсии, объясненная ГК3, равна приблизительно 13%;

- на четвертую главную компоненту (ГК4) – «содержание клейковины»; доля дисперсии, объясненная ГК4, равна приблизительно 12%.

Доля дисперсии первых четырех компонент в сумме дает приблизительно 84 %.

Остальные главные компоненты в сумме содержат 16 % дисперсии, их влияние не значительно и поэтому они могут быть не учтены [8].

На каждую главную компоненту приходятся различные доли факторных нагрузок. Для первых четырех главных компонент были получены следующие линейные комбинации:

$$U_1 = -0{,}78*x_1 + 0{,}37*x_2 - 0{,}61*x_3 - 0{,}53*x_4 + 0{,}09*x_5 - 0{,}79*x_6 + 0{,}74*x_7 - 0{,}6*x_8, \quad (2)$$

$$U_2 = -0{,}11*x_1 - 0{,}78*x_2 + 0{,}34*x_3 - 0{,}57*x_4 + 0{,}58*x_5 - 0{,}15*x_6 - 0{,}24*x_7 + 0{,}41*x_8, \quad (3)$$

$$U_3 = 0{,}35*x_1 - 0{,}29*x_2 - 0{,}61*x_3 - 0{,}43*x_4 - 0{,}02*x_5 + 0{,}33*x_6 + 0{,}44*x_7 - 0{,}2*x_8, \quad (4)$$

$$U_4 = -0{,}33*x_1 + 0{,}005*x_2 + 0{,}05*x_3 - 0{,}26*x_4 - 0{,}69*x_5 - 0{,}16*x_6 + 0{,}16*x_7 + 0{,}51*x_8. \quad (5)$$

В результате применения главных компонент в качестве координатных осей размерность исходных данных понизилась с восьми до четырех входных факторов.

Для дальнейшего анализа данных и построения прогнозирующей модели исходную информацию представим графически в пространстве главных компонент. Такая визуализация данных дает возможность обнаружения скрытых связей между факторами. Новые координаты опытных данных по урожайности яровой пшеницы были вычислены соотношениями (2) – (5). Результаты приведены в виде таблицы 2.

2. Исходные данные в координатах главных компонент.

№	Урожайность	U_1	U_2	U_3	U_4
1	-1,40	-2,89	-3,07	-0,01	3,19
2	-1,83	-2,65	1,46	-0,32	-1,73
3	1,84	-1,57	1,41	0,36	-0,47
4	0,55	-1,91	1,92	-0,16	0,04
5	0,04	-2,03	-1,21	-0,36	-1,6
6	0,07	-4,6	-1,2	0,82	0,04
7	0,64	-0,5	1,59	1,11	0,4
8	0,62	-0,9	3,09	-0,49	-0,14
9	0,95	-1,52	4,04	-0,57	0,92
10	1,08	-0,3	0,05	1,43	-0,61
11	0,23	-2,48	-0,49	2,27	-1,26
12	0,88	0,05	0,11	0,47	-0,59
13	0,39	-0,72	0,2	-1,12	-0,58
14	0,30	-1,22	-1,91	0,52	-0,2
15	0,49	0,59	-0,77	0,82	0,63

16	0,20	-2,46	-2,5	-0,74	-0,77
17	0,42	2,35	0,1	2,41	1
18	0,02	6	2,17	1,02	1,07
19	1,66	3,67	2,23	-0,44	-0,28
20	1,79	0,3	-0,46	-0,74	-0,34
21	0,46	-1,25	0,6	0,5	-0,46
22	1,31	-1,28	1,02	-0,4	1,02
23	0,53	-2,87	-1,12	-1,14	0,28
24	0,76	-2	-0,29	-0,67	-0,64
25	0,17	0,04	0,29	-0,17	0,57
26	0,83	-2,59	-0,17	-1,01	0,95
27	1,46	-1,35	-0,53	-1,21	1,21
28	1,72	2,55	-2,31	0,09	-1,13
29	0,49	5,48	-2,14	1,12	-0,11
30	1,41	5,86	0,44	-1,3	-0,64
31	-0,47	3,39	-0,55	-0,57	0,44
32	0,77	6,82	-1,99	-1,53	-0,22

Современные программные средства и вычислительная техника позволяют представить дискретную информацию в двумерном или трехмерном пространствах и наблюдать за расположением точек под разными углами обзора. В двумерном пространстве количество диаграмм по четырем главным компонентам составляет шесть возможных вариантов, а в трехмерном пространстве – четыре.

Поиск полезной информации с помощью подобных диаграмм представляет собой достаточно сложную и в некоторых случаях невыполнимую задачу. При визуальном анализе необходимо обратить внимание на характер расположения точек относительно главных компонент, расслоения и образования групп. При этом отдельные точки с учетом их вариации могут быть поименованы по каждому фактору. Следует отметить, что не каждая диаграмма содержит полезную информацию. Для выявления скрытых связей между факторами приходится исследовать различные возможные варианты диаграмм. На рисунке 1 представлена диаграмма, где прослеживается связь между первым и четвертым факторами.

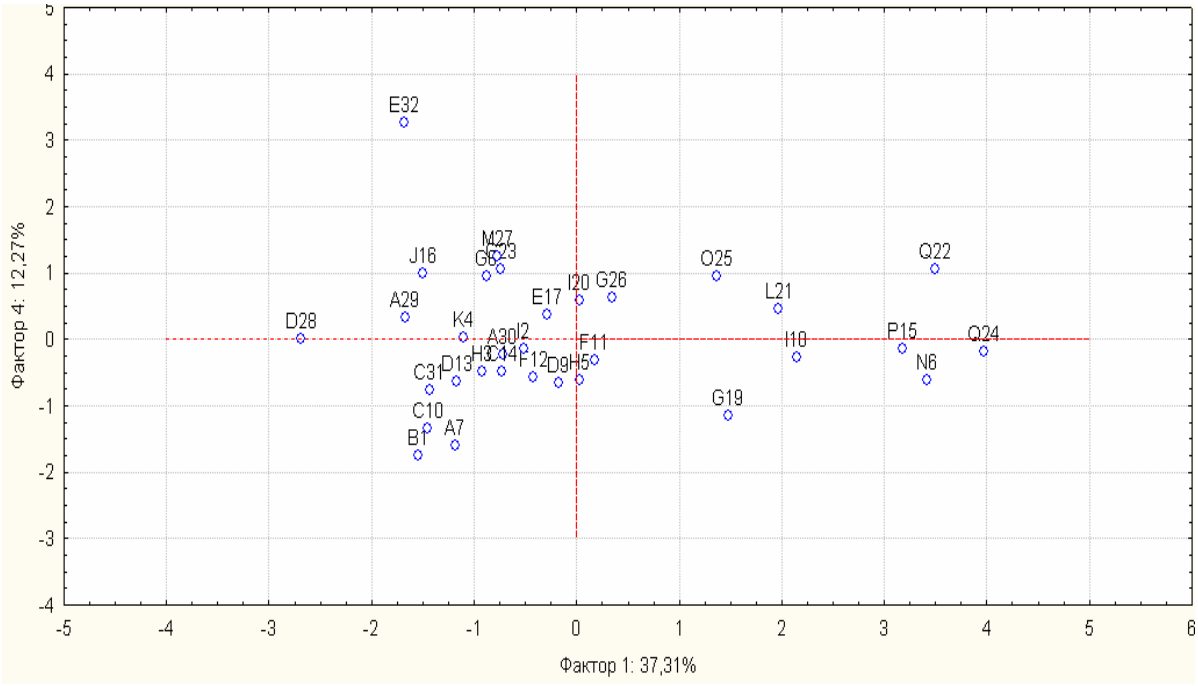

Рис. 1. Графическое представление данных в плоскости ГК1 и ГК4

Основная факторная нагрузка на первой главной компоненте приходится на фактор «масса 1000 зерен», а на четвертой главной компоненте – «модержание клейковины». На графике «масса 1000 зерен» обозначена латинскими буквами, а «модержание клейковины» –цифрами. Предварительный анализ показывает, что большие значения массы зерна связаны с высокими показателями содержания клейковины.

Дальнейшие исследования были направленны на построение прогнозирующей нейросетевой модели. Искусственная нейронная сеть представляет собой набор взаимодействующих блоков в вычислительном устройстве, служащих для обработки данных. Для конструирования модели был использован пакет Deductor Academic. Построена исследовательская модель на базе нейронных сетей типа многослойный персептрон с одним входным одним выходным, и одним скрытым слоем[9,10]. Граф нейросети представлен на рисунке 2. Нейронная сеть была предварительно обучена по исходным данным таблицы 2 и протестирована на тестовом множестве. В процессе обучения нейросети ее параметры настраиваются таким образом, что сигнал на выходе приближен к фактическому. Таким образом, задавая входные данные U_1,U_2,U_3,U_4, можно получить прогнозируемый результат.

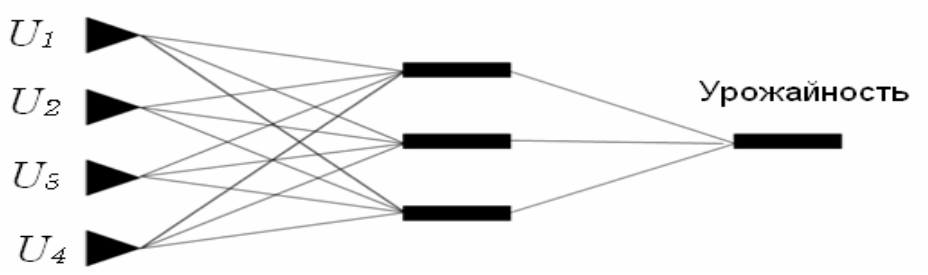

Рис. 2. Граф нейросети

Адекватность нейросетевой модели была проверена с помощью критерия Фишера. Расчетное значение данного критерия для 32 наблюдений и 4 факторов U_1, U_2, U_3, U_4 равно 34,4. Табличное значение при допустимом уровне ошибки 0,05 составляет $F_т$=2,05. Поскольку расчетное значения критерия Фишера больше $F_т$, то можно утверждать, что построенная нейросетевая модель адекватно отражает исходные данные и может быть использована для прогнозирования урожайности яровой пшеницы.

Выводы. Применение метода главных компонент дало возможность сократить размерность задачи об урожайности яровой пшеницы с восьми до четырех переменных, что облегчает поиск возможных скрытых связей между исходными факторами. В частности в диаграмме первого и четвертого главных компонент прослеживается связь между большими значениями массы зерна с высоким показателем клейковины. Сконструирована и настроена нейросетевая модель, которая позволяет прогнозировать урожайность яровой пшеницы по четырем главным компонентам.

Литература

1. Абдрахманов, Р.Х. Некоторые проблемы анализа и управления процессом формирования урожайности / Р.Х. Абдрахманов.– Оренбург, 1998. – 448с.
2. Жученко, А.А. Адаптивное растениеводство (Эколого-генетические основы)– А.А, Жученко. – М.: Изд-во «Агрорус», 2009. –172с.
3. Зиганшин, А.А. Современные технологии и программирование урожайности /А.А. Зиганшин. – Казань: Изд-во Казанского университета, 2001. – 172с.
4. Сержанов, И.М. Яровая пшеница в северной части лесостепи Поволжья / И.М. Сержанов, Ф.Ш. Шайхутдинов. – Казань, 2013. – 234с.
5. Таланов, И.П. Оптимизация приемов формирования высокопродуктивных ценозов яровой пшеницы / И.П. Таланов. – Казань, 2003. –174с.
6. Болч, Б. Многомерные статистические методы для экономики / Б. Болч, К.Дж. Хуань .–М.:Статистика, 1979.– 317с.
7. Валиев, А.А. Современные методы и подходы обработки информации по урожайности яровой пшеницы / А.А.Валиев, Р.И. Ибятов, Ф.Ш. Шайхутдинов // Вестник Казанского государственного аграрного университета. –2016. – № 3 (41).– С. 9-14.
8. Ибятов, Р.И. Факторный анализ данных влияющих на урожайность пшеницы / Р.И. Ибятов, Ф.Ш. Шайхутдинов, А.А. Валиев // Сборник трудов Материалы международной научно-практической конференции «Аграрная наука XXI века. Актуальные исследования и перспективы». – Казань: Изд-во Казанского ГАУ, 2016. – С. 77-79.

9. Новикова, С.В. Нейросетевые методы поиска скрытых связей в многомерных данных / С.В. Новикова, Р.И. Ибятов, А.А. Валиев, Э.Ш. Кремлева // Сборник трудов международной научной конференци «Математические методы в технике и технологиях». – Саратов: Изд-во Сарат. гос.техн. ун-та имени Ю.А. Гагарина, 2014. – С. 128-131.

10. Новикова, С.В. Нейросеть обратного распространения ошибки для анализа выброса свинца в атмосферу от вида и количества транспорта / С.В. Новикова, Р.И. Ибятов, А.А.Валиев, Э.Ш. Кремлева // Сборник трудов Материалы международной научно-практической конференции «Научное сопровождение агропромышленного комплекса: теория, практика, перспективы». – Казань: Изд-во Казанского ГАУ, 2015. – С. 269-271.

Literature

1. Abdrakhmanov, R.Kh. Some problems of the analysis and management of the process of productivity formation / R.Kh. Abdrakhmanov.– Orenburg, 1998. – 448p.

2. Zhuchenko, A.A. Adaptive plant-growing (ecologic and genetic basis) / A.A. Zhuchenko.– M.: Publ. “Agrorus”, 2009. – 172p.

3. Ziganshin, A.A. Modern technologies and programming of productivity / A.A. Ziganshin. Publ. of Kazan SAU, 2001. – 172p.

4. Serzhanov, I.M. Spring wheat in the northern part of forestry steppe of Povolzhie / I.M. Serzhanov, F.Sh. Shaykhutdinov. – Kazan, 2013. – 234p.

5. Talanov, I.P. Optimization of methods of highly productive coenosis of spring whea t/ I.P. Talanov. – Kazan, 2003.– 174p.

6. Bolch, B. Multi-dimensional methods for economics / B. Bolch, K.Dzh. Khuan.–M.: Statistics, 1979.– 317p.

7. Valiev, A.A. Modern methods and approaches of information processing on the yield of spring wheat / A.A. Valiev, R.I. Ibyatov, F.Sh. Shaykhutdinov // Newsletter of Kazan State Agrarian University. – 2016. – № 3 (41). – PP. 9-14.

8. Ibyatov, R.I. Factor analysis of the data effecting wheat productivity / R.I. Ibyatov, F.Sh. Shaykhutdinov, A.A. Valiev // Collection of the works of the International Science-practical conference ‘Agrarian Science of XXI century. Urgent research and prospects’. – Kazan: Publ. of Kazan SAU, 2016. – PP. 77-79.

9. Novikova, S.V.Neural methods for finding hidden relationships in multidimensional data / S.V. Novikova, R.I. Ibyatov, A.A.Valiev, E.Sh. Kremleva // Collection of the works of the International Science-practical conference «Mathematic methods in technique and technologies». – Saratov: Publ. of Saratov State technical University after Yu.A. Gagarin, 2014. – PP. 128-131.

10. Novikova, S.V. Neural network of reverse distribution of the mistake for the analysis of lead emission into the air from kind and quantity of traffic / S.V. Novikova, R.I. Ibyatov, A.A. Valiev, E.Sh. Kremleva // Collection of the works of the International Science-practical

conference «Scientific accompaniment of agro industrial complex: theory, practice, prospects». – Kazan: Publ. of Kazan SAU, 2015. – PP. 269-271.