УДК 519.24

М.А. Салтыков
Дальневосточный государственный технический рыбохозяйственный университет,
690087, г. Владивосток, ул. Луговая, 52б

МЕТОДИЧЕСКИЕ РЕКОМЕНДАЦИИ ПРОВЕДЕНИЯ КЛАСТЕРНОГО АНАЛИЗА С ИСПОЛЬЗОВАНИЕМ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ В ИССЛЕДОВАНИЯХ РЫБНОЙ ПРОМЫШЛЕННОСТИ

Обсуждается применение статистических методов анализа многомерных массивов данных в исследованиях рыбной промышленности; пошагово разбирается процедура проведения кластеризации товарной продукции рыбопромышленного комплекса Дальневосточного федерального округа с использованием иерархического алгоритма построения дендограммы экспорта и импорта.

Целью публикации является рассмотрение алгоритма проведения кластерного анализа с использованием информационного продукта Statistika версия 6. Предметом исследования выступает информационное обеспечение и инструментарий проведения статистических исследований, реализованное в прикладном информационном программном продукте Statistika. Объектами исследования являются экспорт и импорт товарной рыбной продукции предприятий Дальнего Востока, реализуемые через таможенные органы Дальневосточного таможенного управления.

Результатом работы являются выделенные на основе метода иерархического кластерного анализа товарные кластеры экспорта и импорта рыбной продукции, а также методические рекомендации проведения кластерного анализа с использованием прикладного информационного программного обеспечения Statistika. Полученные результаты могут использоваться в более глубоком анализе экспорта и импорта рыбной продукции, в качестве методического обеспечения в исследованиях рынков рыбной продукции, производителей рыбной продукции, проведении сегментации потребителей рыбной продукции по различным признакам.

Ключевые слова: *информационные технологии, статистические методы, иерархический кластерный анализ, обработка данных, рыбная продукция, базы данных, классификация товарной номенклатуры, дендрограмма.*

M.A. Saltykov

METHODOLOGICAL RECOMMENDATIONS FOR CLUSTER ANALYSIS USING INFORMATION TECHNOLOGIES IN FISH INDUSTRY RESEARCH

The research discusses the use of statistical methods for the analysis of multidimensional data sets in fisheries research. The procedure of clustering of commercial products of the fishing complex of the Far East Federal District using a hierarchical algorithm and construction of an export and import dendogram is dealt with step by step. The purpose of the publication is to consider the algorithm of cluster analysis using the information product Statistika 6. The subject of the study is information support and tools for carrying out statistical research, implemented in the application information software product Statistika. The object of the study is the export and import of commercial fish products of industry of enterprises of the Far East, sold through customs authorities of the Far East Customs Administration.

The results of the work are commodity clusters of export and import of fish products identified on the basis of hierarchical cluster analysis method, as well as methodological recommenda-

tions for cluster analysis using Statistics application information software. The results can be used in more in-depth analysis of export and import of fish products, as a methodological support in research of fish products markets, fish products producers, carrying out segmentation of fish products consumers by various characteristics.

Key words: *information technology, statistical methods, hierarchical cluster analysis, data processing, fish products, databases, commodity nomenclature classification, dendrogram.*

Введение

Статистические методы исследования многомерных массивов данных позволяют обосновать и выбрать модель совокупности показателей, наиболее адекватно соответствующей исходным данным, объективно характеризующей совокупность исследуемых объектов.

В свою очередь применение информационных технологий обработки данных в исследованиях повышает производительность научного процесса и качество полученных результатов, отвечает современным тенденциям применения компьютерной техники в науке и образовании.

В зарубежных работах метод иерархического кластерного анализа широко применяется в исследованиях, связанных с вопросами связи между продуктивностью гидробионтов и экологическими, биотическими, абиотическими факторами. На основе статистических инструментов кластерного анализа получают модели прогнозирования продуктивности рыбы, данный метод применяется при обработке экспертных оценок, экологических факторов, таких как температура воды, концентрация хлорофилла в море и других параметров. Некоторые разработанные модели позволяют обеспечить более достоверный прогноз рыбного промысла, иногда метод применяется при систематизации и распределении рыбопромысловых участков.

Кластерный метод получил широкое применение в экономических и социологических исследованиях. В то же время в российских исследованиях в области экономики, маркетинга рыбной промышленности данный метод не получил широкого распространения, но имеет большой потенциал в области стратегического маркетинга рыбной продукции. Таким образом, целью данного исследования является рассмотрение алгоритма проведения кластерного анализа с использованием информационного продукта Statistika в исследованиях, связанных с экономикой и маркетингом рыбной промышленности.

Предмет исследования – информационное обеспечение, инструментарий проведения статистических исследований, реализованное в информационной платформе Statistika.

Объект исследования – товарная продукция рыбной промышленности предприятий Дальнего Востока, реализуемая через таможенные органы Дальневосточного таможенного управления в 2018 г.

Методы исследования

Название кластерный анализ происходит от английского слова cluster – гроздь, скопление. По мнению многих исследователей, первым стал применять кластерный анализ и описал его методологию в 1939 г. R.C. Tryon [1].

Основное назначение кластерного анализа – это расчленение совокупности исследуемых объектов и их признаков на условно однородные группы, т.е. кластеры. Методология кластерного анализа универсальна и применяется при решении многих задач, допустима при простой группировке, сформированной по признаку количественного сходства.

При применении кластерного анализа решается задача, при которой данные множества Y разделяются на множество A × n (a – является целым числом) кластеров, т.е. подмножеств Z_1, Z_2, Z_n, так, чтобы каждый объект A_i принадлежал только одному подмножеству

распределения. В свою очередь объекты, которые принадлежат одному кластеру, должны быть схожими и отличаться от объектов, принадлежащих разным кластерам.

Самой распространённой метрикой в кластерном анализе является евклидово расстояние, которое выступает геометрическим расстоянием многомерного пространства:

$$d_{(x,y)} = \sum_{i=1}^{m} ((x_i - y_i)^2)^{\frac{1}{2}}$$

где x и y – точки в n-мерном пространстве.

Результатом иерархического кластерного анализа является дендрограмма (от греческого dendron –«дерево»), которая наглядно характеризует и демонстрирует близость отдельных точек и кластеров, а также стадийность объединения кластеров.

В настоящее время встречается множество работ в области методологии кластерного анализа [2, 3] и др.

В ходе данного исследования рассматривались зарубежные работы в области биологии водных биологических ресурсов с применением кластерного анализа [5, 6] и др.

Результаты и их обсуждение

Выполним решение практической задачи по кластеризации в программе Statistika, являющейся известным статистическим информационным продуктом обработки данных компании StatSoft, Inc [7].

Для проведения исследования предварительно подготовим данные по экспорту и импорту рыбной продукции, используем данные Дальневосточного таможенного управления [8] и подготовим данные в таблице Excel. Для корректного импорта данных в пакет Statistika из исходных данных необходимо удалить неинформативные значения, такие как номер по порядку и страна экспорта и другие данные.

Затем заходим в пакет Statistika и выполняем команду открыть, выбираем тип файлов с форматом .xls, импортируем выбранные листы (рис. 1)

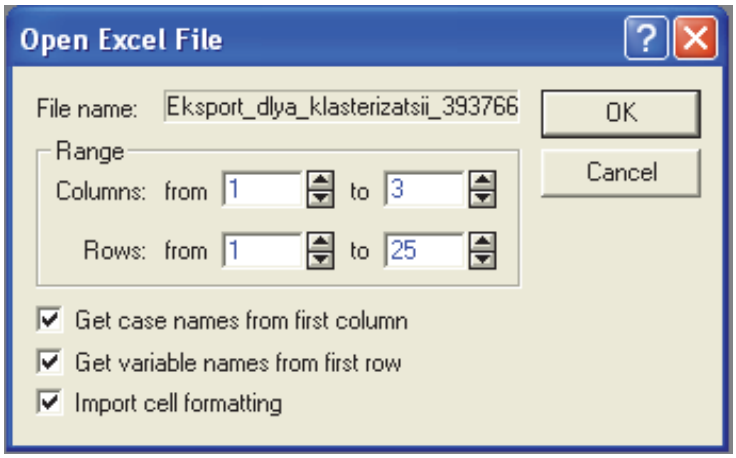


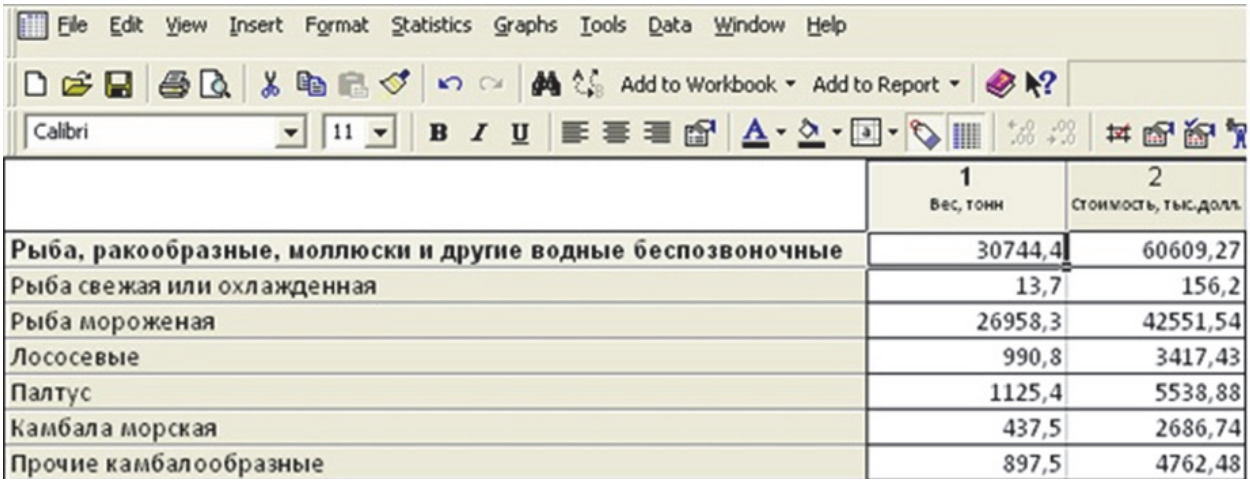
Рис. 1. Импорт данных из программы Excel в пакет STATISTICA
Fig. 1. Importing data from Excel into the STATISTICA package

Исходный файл данных содержит следующую информацию о товарах экспорта (рис. 2) и импорта (рис. 3):

- наименование товара;
- вес экспортируемого и импортируемого товара в тоннах – первая переменная;
- стоимость данного товара в тысячах долларов – вторая переменная.

	1 Вес, тонн	2 Стоимость, тыс.долл.
Рыба, ракообразные, моллюски и другие водные беспозвоночные	1536837,6	3280468,76
Рыба свежая или охлажденная	395	589,64
Рыба мороженая	1362488,3	2119010,1
Лососевые	196588,4	558209,58
Палтус	9488,9	53485,08
Камбала морская	12406,4	16694,04
Прочие камбалообразные	34297,7	44991,07
Сельдь	213556,7	115385,82
Треска	53938,1	159011,39
Минтай	698960,2	716290,15
Окунь морской	1035,8	2892,25
Печень, икра, молоки	52653,1	361627,46
Плавники, головы, хвосты и прочие пищевые рыбные субпродукты	6573,6	14757,59
Филе рыбное и прочее мясо рыбы (включая фарш)	65928,3	156226,85
Рыба сушеная, соленая, копченая, рыбная мука	363,9	822,84
Ракообразные	68337,4	889758,84
Крабы	56948,5	797779,08
Моллюски	27742,6	84635,05
Гребешки	6305,5	14306,61
Каракатицы, кальмары и осьминоги	12362,6	28801,49
Прочие моллюски	8963,4	41425,85
Водные беспозвоночные, кроме ракообразных и моллюсков	11581,9	29425,44
Голотурии	2967,7	5994,26
Морские ежи	7921,4	22606,34

Рис. 2. Исходные статистические данные экспорта рыбной продукции для кластеризации
Fig. 2. Source statistics of export of fish products for clustering



	1 Вес, тонн	2 Стоимость, тыс.долл.
Рыба, ракообразные, моллюски и другие водные беспозвоночные	30744,4	60609,27
Рыба свежая или охлажденная	13,7	156,2
Рыба мороженая	26958,3	42551,54
Лососевые	990,8	3417,43
Палтус	1125,4	5538,88
Камбала морская	437,5	2686,74
Прочие камбалообразные	897,5	4762,48

Рис. 3. Исходные статистические данные импорта рыбной продукции для кластеризации
Fig. 3. Source statistics of imports of fish products for clustering

Проведем стандартизацию данных (в меню Data необходимо выбрать пункт Standardize и выделить все переменные) с целью исключения влияния различных типов шкал, в которых представлены переменные (рис. 4, 5).

Data: Стандартизованные данные экспорта* (2v by 24c)

	1 Вес, тонн	2 Стоимость, тыс.долл.
Рыба, ракообразные, моллюски и другие водные беспозвоночные	3,2	3,7
Рыба свежая или охлажденная	-0,4	-0,5
Рыба мороженая	2,8	2,2
Лососевые	0,0	0,2
Палтус	-0,4	-0,4
Камбала морская	-0,4	-0,5
Прочие камбалообразные	-0,4	-0,5
Сельдь	0,1	-0,4
Треска	-0,3	-0,3
Минтай	1,2	0,4
Окунь морской	-0,4	-0,5
Печень, икра, молоки	-0,3	-0,0
Плавники, головы, хвосты и прочие пищевые рыбные субпродукты	-0,4	-0,5
ле рыбное и прочее мясо рыбы (включая фарш), свежие, охлажде	-0,3	-0,3
Рыба сушеная, соленая, копченая, рыбная мука	-0,4	-0,5
Ракообразные	-0,3	0,6
Крабы	-0,3	0,5
Моллюски	-0,4	-0,4
Гребешки	-0,4	-0,5
Каракатицы, кальмары и осьминоги	-0,4	-0,5

Рис. 4. Стандартизация данных экспорта рыбной продукции
Fig. 4. Standardization of fish export data

Data: Стандартизованые данные импорта* (2v by 7c)

	1 Вес, тонн	2 Стоимость, тыс.долл.
Рыба, ракообразные, моллюски и другие водные беспозвоночные	1,59599372	1,799072
Рыба свежая или охлажденная	-0,632744902	-0,700796998
Рыба мороженая	1,32140748	1,05234467
Лососевые	-0,561880896	-0,565937878
Палтус	-0,552119055	-0,478211199
Камбала морская	-0,602008881	-0,596153536
Прочие камбалообразные	-0,568647462	-0,510317067

Рис. 5. Стандартизация данных импорта рыбной продукции
Fig. 5. Standardization of data on imports of fish products

Для проведения кластерного анализа используем модуль Cluster Analysis (Кластерный анализ) с использованием переключателя модулей Statistika Multivariate Exploratory Techniques (рис. 6). Затем в появившемся окне (Clustering method) из предложенных методов Joining (tree clustering) (древовидная кластеризация): K-means clustering – Кластеризация

методом К-средних и Two-way joining – Двухходовое объединение – необходимо выбрать пункт Joining (tree clustering).

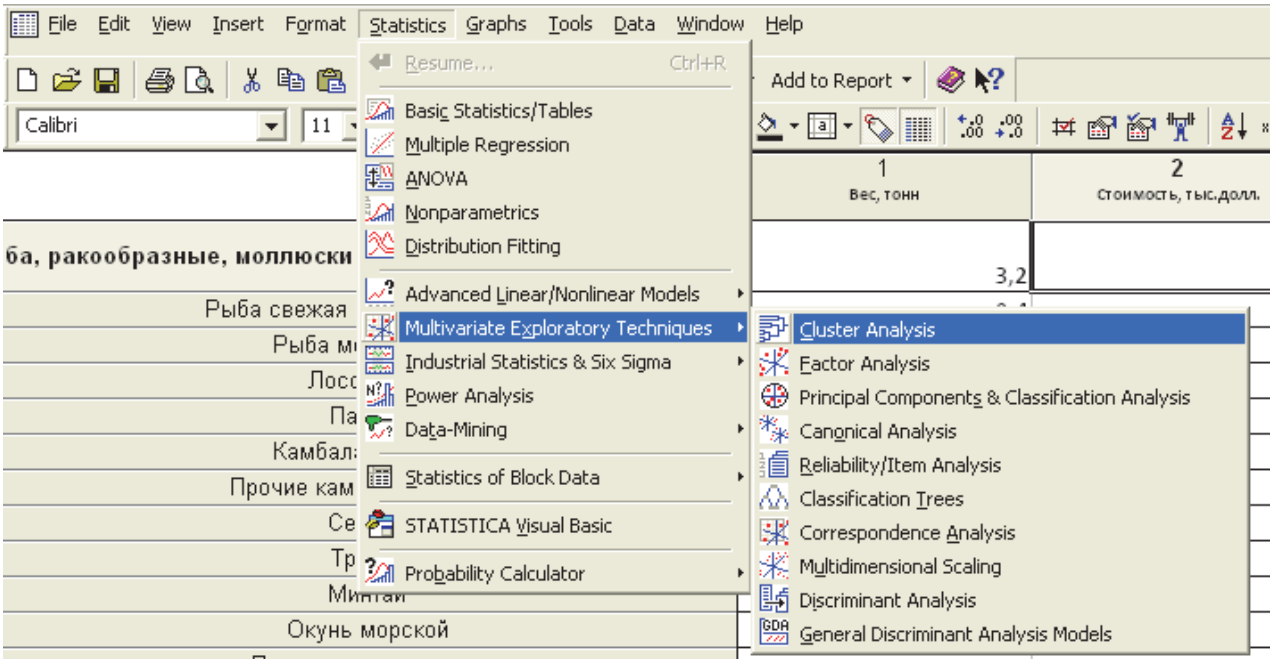
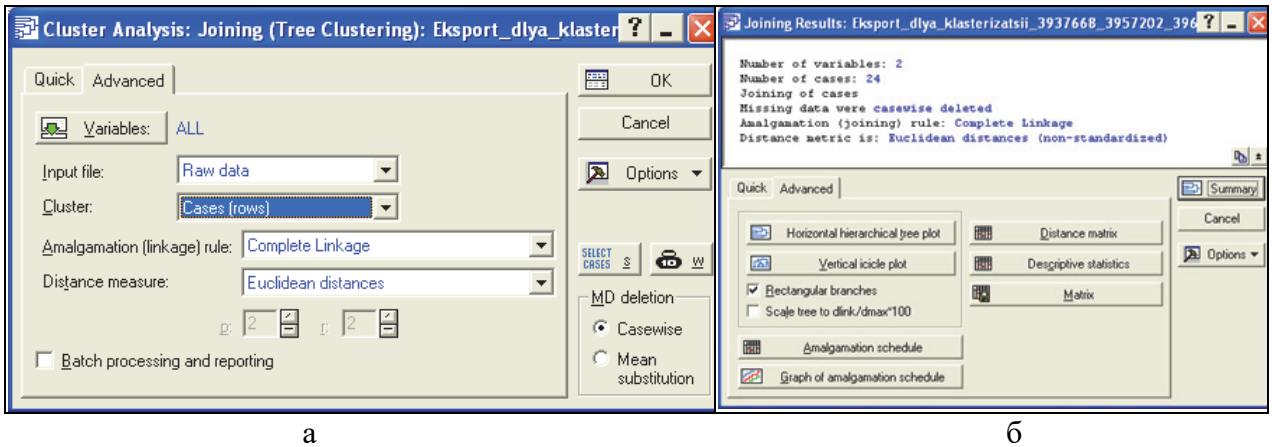


Рис. 6. Запуск модуля Кластерный анализ (Cluster Analysis)
Fig. 6. Starting the Cluster Analysis Module

На следующем этапе необходимо установить необходимые параметры, как показано на рис. 7, а, установить значение All в меню Variables. В поле Cluster необходимо установить значение Cases (rows). В качестве правила объединения отметим Complete Linkage (Метод полной связи), в качестве метрики расстояний – Euclidean Distances (евклидово расстояние). Далее выбираем пункт построить вертикальную дендрограмму (Verticalicicleplot) на рис. 7, б.

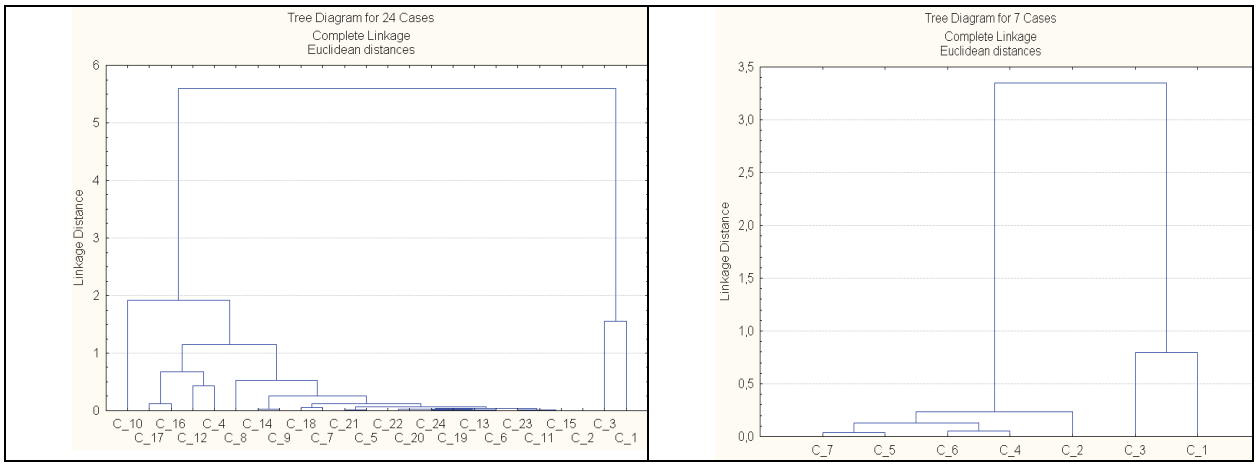


а б

Рис. 7. Параметры выполнения кластерного анализа
Fig. 7. Cluster Analysis Options

На рис. 8 представлены результаты кластеризации товаров экспорта и импорта. Из дендрограммы товаров экспорта (рис. 8, а) выделяется несколько классов:

1. C_1 и C_3 «Рыба мороженая» – наиболее прибыльные товары с большим объемом экспорта.
2. C_{10} «Минтай», C_{17} «Крабы», C_{16} «Ракообразные», C_{12} «Печень, икра, молоки», C_4 «Лососевые» – средние товары по объемам экспорта и суммарной стоимости.
3. Все остальные – низкоприбыльные товары с небольшими объемами экспорта.



а б

Рис. 8. Дендограммы кластеров экспортной (а) и импортной (б) рыбной продукции
Fig. 8. Dendograms of clusters of export (a) and import (б) fish products

Из дендрограммы товаров импорта (рис. 8, б) можно выделить следующие классы:
1. C_1 и C_3 «Рыба мороженая» – кластер аналогичен товарам экспорта.
2. C_2 «Рыба живая» – товар с самым низким показателем массы импорта и стоимости.
3. C_4 «Филе рыбное и прочее мясо рыбы (включая фарш), свежие, охлажденные или мороженые» и C_6 «Ракообразные».
4. C_5 «Рыба сушеная, соленая, копченая, рыбная мука» и C_7 «Каракатицы, кальмары и осьминоги».

Выводы

В ходе данного исследования был рассмотрен подробный алгоритм выполнения кластеризации экспорта и импорта рыбной продукции с использованием информационного продукта Statistika версия 6. Исследование позволило выявить наиболее схожие объекты – кластеры по переменным – вес экспортируемого и импортируемого товара в тоннах и стоимость данного товара в тысячах долларов. Необходимо отметить, что при увеличении числа переменных результат кластеризации может значительно измениться. Поэтому на достоверность проводимого исследования оказывает значительное влияние предварительная подготовка данных к анализу.

При дальнейших исследованиях полученные данные могут применяться при разработке стратегических маркетинговых мероприятий, в исследованиях рынков рыбной продукции, структуры и кластеризации производителей рыбной продукции по различным переменным, сегментации потребителей водных биологических ресурсов и более детального исследования их предпочтений.

Список литературы

1. Tryon R.C. Cluster analysis. London: Ann Arbor Edwards Bros, 1939. 139 p.
2. Filippone M., Camastra F., Masulli F. Rovetta S. A Survey of Kernel and Spectral Methods for Clustering // Pattern Recognition. 2008. Vol. 41, № 1. P. 176–190.

3. Mirkin B. Core Concepts in Data Analysis: Summarization, Correlation, Visualization. Springer, 2010.

4. Donald A. Jackson, Steven C. Walker, Mark S. Poos Cluster Analysis of Fish Community Data: «New» Tools for Determining Meaningful Groupings of Sites and Species Assemblages. American Fisheries Society Symposium 73:503–527, 2010 [Электронный ресурс]. URL: http://jackson.eeb.utoronto.ca/files/2012/10/2010_Jackson_Walker_Poos_StreamFishCommunities.pdf (дата обращения: 03.11.2019).

5. Frimpong, E. A., Angermeier P. L. Traitbased approaches in the analysis of stream fish communities. Pages 109–136 in K. B. Gido and D. A. Jackson, editors. Community ecology of stream fishes: concepts, approaches, and techniques 2010 [Электронный ресурс]. URL: http://www.fishtraits.info/static/pdf/Using.pdf (дата обращения: 03.11.2019).

6. Yuan H.C., Tan M.X., Gu Y.T. A Model for Fishery Forecast Based on Cluster Analysis and Nonlinear Regression. International Conference on Artificial Intelligence and Industrial Engineering (AIIE 2015) [Электронный ресурс]. URL: https://www.atlantis-press.com/proceedings/aiie-15/22156 doi.org/10.2991/aiie-15.2015.113 (дата обращения: 03.11.2019).

7. STATISTICA Features Overview [Электронный ресурс]. URL: http://www.statsoft.com/Products/STATISTICA-Features (дата обращения: 03.11.2019).

8. Дальневосточное таможенное управление. Справочные и аналитические материалы. [Электронный ресурс]. URL: http://dvtu.customs.ru/statistic (дата обращения: 26.10.2019).

Сведения об авторе: Салтыков Максим Александрович, кандидат экономических наук, доцент, e-mail: saltykov_ma@mail.ru.