УДК: 004.42:519.23:574
doi: 10.31140/j.vestnikib.2016.3(197).4

ВЗАИМОДЕЙСТВИЕ EXCEL И СТАТИСТИЧЕСКОГО ПАКЕТА R ДЛЯ ОБРАБОТКИ ДАННЫХ В ЭКОЛОГИИ

А.Б. Новаковский
Федеральное государственное бюджетное учреждение науки
Институт биологии Коми научного центра Уральского отделения РАН, Сыктывкар
E-mail: novakovsky@ib.komisc.ru

Аннотация. В статье дается описание надстройки ExcelToR, разработанной для совместного использования Microsoft Excel и статистического пакета R. Microsoft Excel используется для ввода, хранения и подготовки данных для анализа, программа статистической обработки данных R – в качестве вычислительного «ядра». Базовые функции взаимодействия Excel и R через XML файлы стандартизованы и унифицированы, что позволяет в течение короткого времени реализовывать и/или модернизировать необходимые алгоритмы анализа и визуализации данных.
В настоящее время в модуле ExcelToR реализованы следующие алгоритмы обработки данных: кластерный анализ (Single-linkage clustering и Complete-linkage clustering, Ward's method, UPGMA), ординация методом главных компонент и неметрическое многомерное шкалирование, построение теплокарт.

Ключевые слова: статистический анализ, Excel, R, кластерный анализ, ординация, теплокарты

Термин «экология» впервые был введен Эрнстом Геккелем в 1866 г. для определения отрасли науки, изучающей взаимодействие живых организмов и их сообществ между собой и окружающей средой (Одум, 1986; Бигон, 1989). В современной экологии для решения этих фундаментальных задач широко применяют различные математические методы (Экологическая оценка..., 1956; Песенко, 1982; Legendre, 1998; Лебедева, 2002; Шитиков, 2003), в том числе основанные на статистическом анализе данных (Василевич, 1969; Нешатаев, 1971; Hill, 1994; Leps, 1999; McCune, 2002; Пузаченко, 2004). Активное внедрение вычислительной техники позволило автоматизировать и существенно ускорить процесс статистической обработки больших объемов данных. Для анализа используют множество различных программных продуктов: SPSS, Statistica, Past (Hammer, 2001), PC-ORD (McCune, 2002), Juice (Tichy, 2002), Twinspan (Hill, 1994; Legendre, 1998), Canoco (Jongman, 1987), Decorana (Hill, 1979) и т.п. На наш взгляд, одной из наиболее перспективных разработок в этой области является программа статистического анализа R (Seefeld, 2007; Borcard, 2011; Мастицкий, 2014; Kabacoff, 2015). Это свободно распространяемая программа с открытой лицензией (https://cran.r-project.org), обладающая большой гибкостью в плане выбора алгоритмов для анализа и отображения результатов. Кроме базовых функций программа позволяет использовать дополнительные расширения, в которых реализованы практически все виды статистических задач (на сегодняшний день таких расширений более 2000). Основным недостатком пакета R является сложность его использования. Программа R управляется при помощи командной строки. Это означает, что для любого действия (проведения статистического анализа, построения диаграммы, записи результатов работы в файл и т.п.) требуется вручную ввести необходимые команды или целые наборы команд (скрипты). По этой причине язык R при всех своих очевидных достоинствах редко используют в научных исследованиях.

Существует несколько разработок, упрощающих работу с программой R. Некоторые из них приближают интерфейс данной программы к общепринятым стандартам и позволяют частично автоматизировать процесс написания скриптов (RStudio, RCommander), другие интегрируют функционал R с табличными процессорами (RExcel, R and Calc) и даже Microsoft Word (SWord). Однако во всех указанных разработках использован английский язык и, кроме того, большинство из них являются платными. С учетом этого мы поставили цель создать свободно распространяемый программный продукт, ориентированный на русскоязычного пользователя.

Разработанная нами программа ExcelToR является надстройкой для Excel. Она обладает простым и понятным интерфейсом пользователя и позволяет объединить легкость ввода и хранения данных, характерную для табличных процессоров, с возможностями статистической обработки данных, предоставляемыми пакетом R. С использованием надстройки можно проводить ординацию и кластерный анализ данных, накладывать дополнительные экологические факторы на результирующие диаграммы в виде цветов и корреляционных векторов, представлять данные в виде теплокарт.

Взаимодействие Excel и статистического пакета R осуществляется через XML (eXtensible Markup Language – расширяемый язык разметки) файлы. Посредством этих файлов из Excel в R передаются исходные данные для анализа и необходимая вспомогательная информация (размеры результирующей диаграммы, настройки для алгоритмов). Обратно передаются результаты обработки данных, пути к построенным диаграммам или же сообщения о произошедших во время выполнения в среде R ошибках (рис. 1).

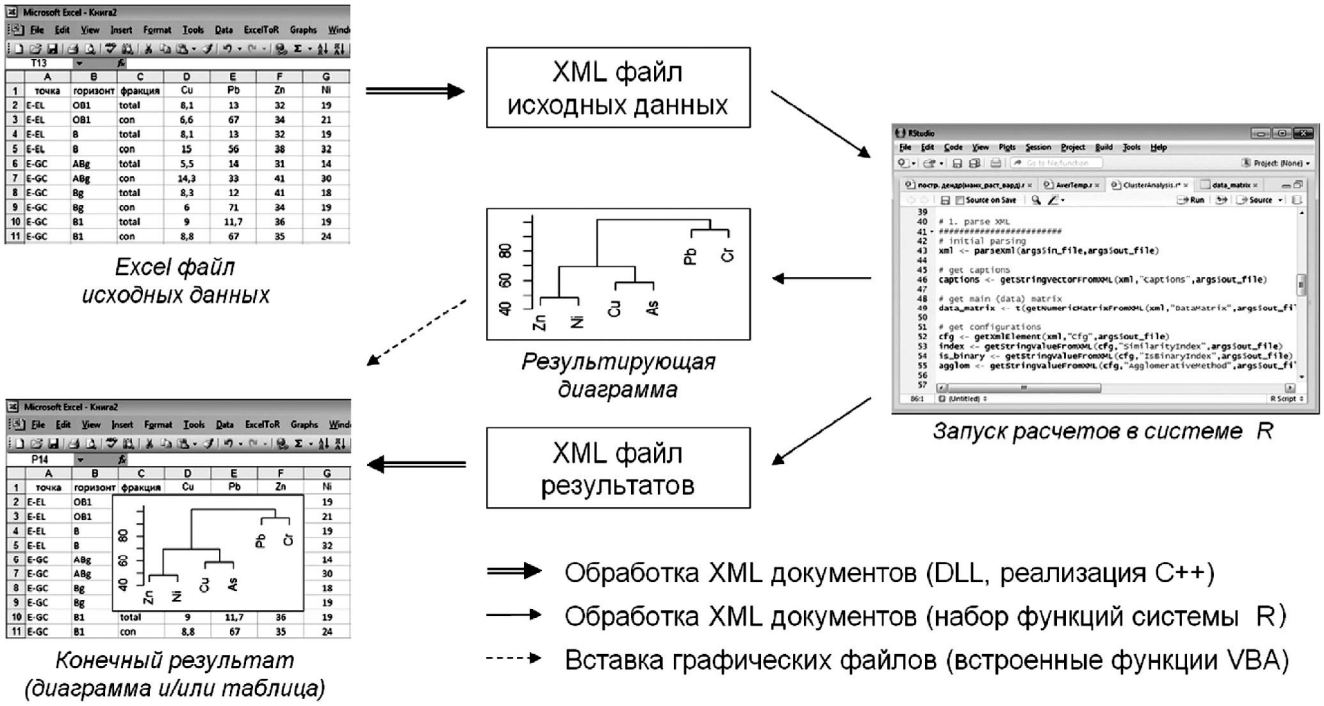


Рис. 1. Схема взаимодействия Excel с пакетом R через XML файлы, реализованная в надстройке ExcelToR.

Реализация алгоритмов обработки построена по блочной системе, т.е. каждый вид анализа содержится в отдельном R скрипте и не зависит от остальных файлов. За счет этого достигается необходимая гибкость и простота при модернизации каждого отдельного алгоритма или же добавлении нового.

Общими являются только модули взаимодействия с XML файлами, реализованные в виде DLL библиотек (чтение и запись XML файлов в/из Microsoft Excel) и наборов скриптов R (чтение и запись XML файлов в/из пакета R).

Для работы надстройки ExcelToR необходима установленная 32- или 64-разрядная версия Microsoft Excel из поставки Microsoft Office (2003, 2007, 2010) и свободно распространяемая программа статистической обработки R версии 3.2.2 или выше. Общий объем программы R, дополнительных статистических пакетов и самой надстройки ExcelToR – около 150 Мб на жестком диске.

Надстройка ExcelToR не является самостоятельной программой и не обладает собственным интерфейсом. Надстройка подключается к Microsoft Excel, а доступ к ее функциям осуществляется через дополнительный пункт меню Excel (рис. 2).

После выбора в основном меню пункта «ExcelToR» появляется главное меню надстройки (рис. 2), которое содержит полный список функций. На сегодняшний день в надстройке реализованы:

– метод главных компонент (Principal component analysis – PCA);

– неметрическое многомерное шкалирование (Nonmetric multidimensional scaling – NMS);

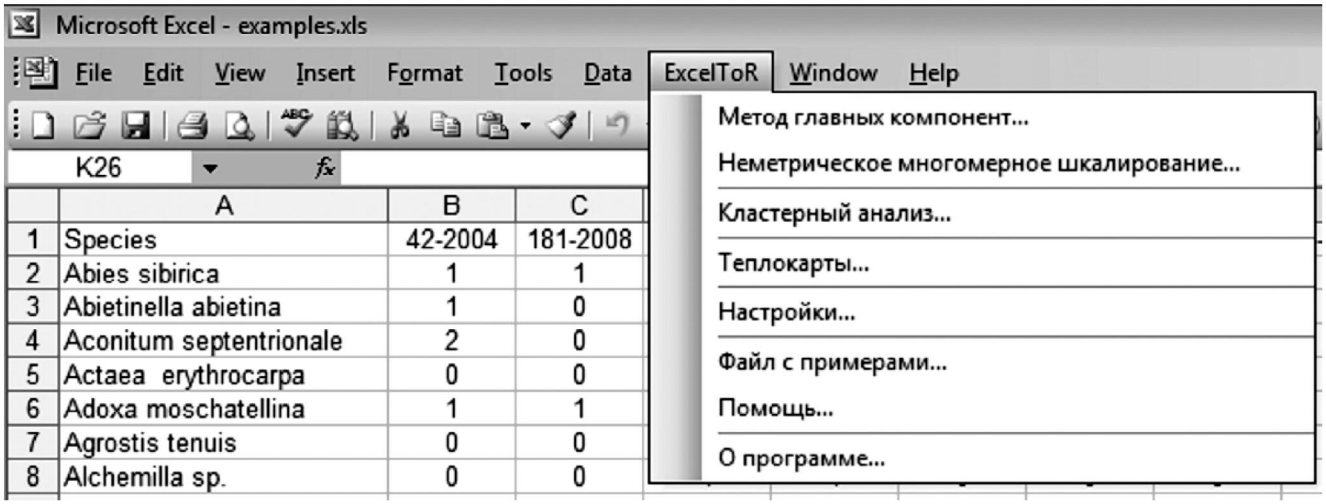
– кластерный анализ (cluster analysis);

– построение теплокарт (Heat maps).

В качестве исходных данных для анализа используют стандартную Excel таблицу. Столбцы таблицы представляют собой перечень анализируемых объектов, строки – их характеристики (или наоборот строки – объекты, столбцы – характеристики). В ячейках таблицы ставят числовые значения параметров. Для корректного определения областей с данными и подписями желательно придерживаться следующей схемы: названия объектов – одна верхняя строка, названия свойств – одна левая колонка (рис. 3). По умолчанию (если пользователь не выделил какую-либо область) в анализ включают все данные, расположенные на текущем листе Excel. При выделении пользователем области данных все вычисления ведутся только для нее.

Одними из часто используемых алгоритмов анализа экологических данных являются методы ординации (Bray, 1957; Hill, 1979; Ter Braak, 1986, 1994; Jongman, 1987; Legendre, 1998; Leps, 1999; McCune, 2002; Пузаченко, 2004). Они позволяют в графическом виде отобразить существующие зависимости и расположить исследуемые объекты в пространстве основных (чаще всего двух-трех) влияющих факторов. С одной стороны, это позволяет отобразить структуру взаимосвязей внутри массива данных, существующие тренды или группы сходных объектов, с другой – выделить наиболее значимые факторы. В разработанной нами надстройке реализовано два метода ординации: метод главных компонент (Андерсон, 1963; Андрукович, 1973; Шитиков, 2003) и неметрическое многомерное шкалирование (She-

А



	A	B	C
1	Species	42-2004	181-2008
2	Abies sibirica	1	1
3	Abietinella abietina	1	0
4	Aconitum septentrionale	2	0
5	Actaea erythrocarpa	0	0
6	Adoxa moschatellina	1	1
7	Agrostis tenuis	0	0
8	Alchemilla sp.	0	0

Б

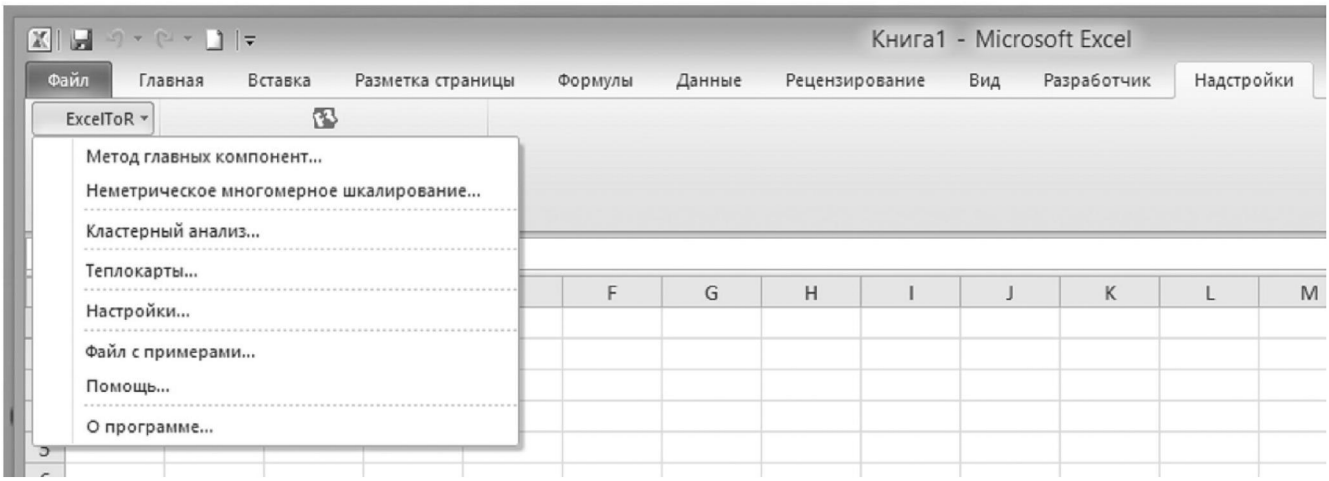
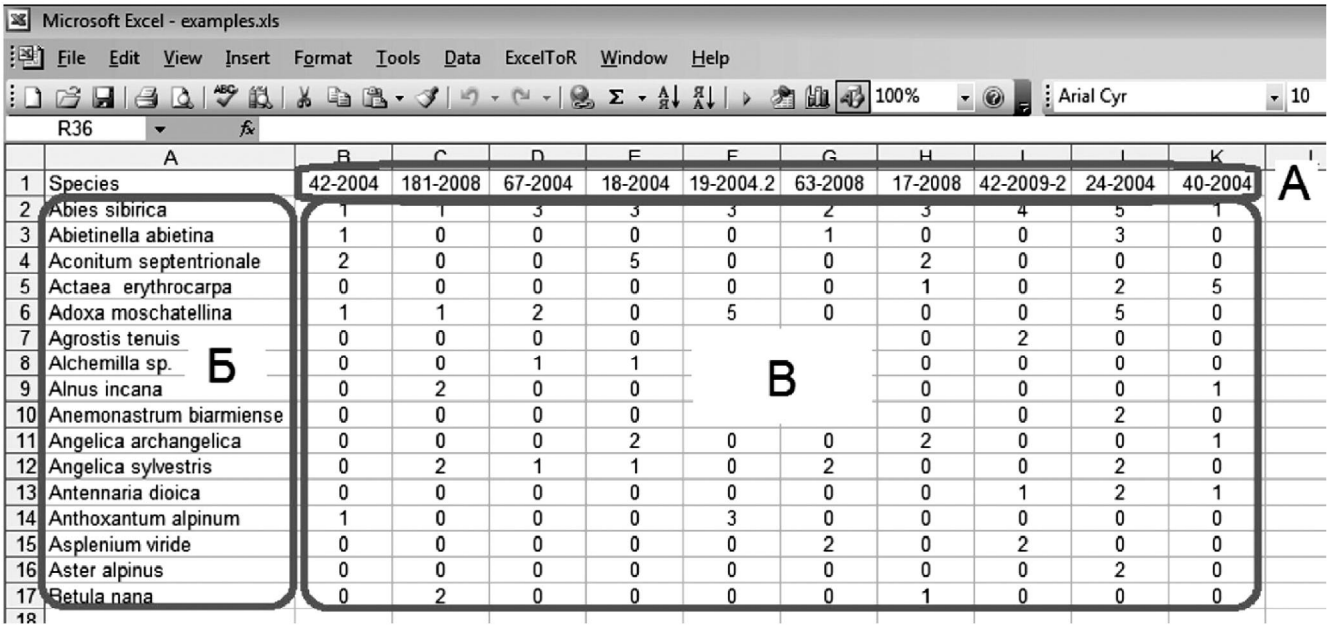


Рис. 2. Расположение пункта меню ExcelToR в Microsoft Excel. А – для версии Microsoft Excel 2003 (и более ранних), Б – для версии Microsoft Excel 2007 (и более поздних).



	A	B	C	D	E	F	G	H	I	J	K
1	Species	42-2004	181-2008	67-2004	18-2004	19-2004.2	63-2008	17-2008	42-2009-2	24-2004	40-2004
2	Abies sibirica	1	1	3	3	3	2	3	4	5	1
3	Abietinella abietina	1	0	0	0	0	1	0	0	3	0
4	Aconitum septentrionale	2	0	0	5	0	0	2	0	0	0
5	Actaea erythrocarpa	0	0	0	0	0	0	1	0	2	5
6	Adoxa moschatellina	1	1	2	0	5	0	0	0	5	0
7	Agrostis tenuis	0	0	0	0			0	2	0	0
8	Alchemilla sp.	0	0	1	1			0	0	0	0
9	Alnus incana	0	2	0	0			0	0	0	1
10	Anemonastrum biarmiense	0	0	0	0			0	0	2	0
11	Angelica archangelica	0	0	0	2	0	0	2	0	0	1
12	Angelica sylvestris	0	2	1	1	0	2	0	0	2	0
13	Antennaria dioica	0	0	0	0	0	0	0	1	2	1
14	Anthoxantum alpinum	1	0	0	0	3	0	0	0	0	0
15	Asplenium viride	0	0	0	0	0	2	0	2	0	0
16	Aster alpinus	0	0	0	0	0	0	0	0	2	0
17	Betula nana	0	2	0	0	0	0	1	0	0	0

Рис. 3. Таблица исходных данных. А – область подписей анализируемых объектов, Б – список свойств, В – цифровые данные.

pard, 1962; Kruskal, 1964; Clarke, 1993; Cox, 2001). Соответствующие диалоговые окна приведены на рис. 4 и 5.

Метод главных компонент является одним из наиболее распространенных методов снижения размерности. В математических терминах PCA относится к линейным методам, т.е. построение новых осей основано на линейной комбинации старых. Преобразование строится таким образом, чтобы среднеквадратичное расстояние между точками (анализируемыми объектами) было максимальным. Оси нового подпространства есть собственные вектора корреляционной (или ковариационной) матрицы, построенной на основе входной выборки, а соответствующие собственным векторам собственные значения характеризуют дисперсию входных данных (Dunteman, 1989; Зиновьев, 2000). Отбрасывание осей с минимальной дисперсией позволяет существенно снижать размерность пространства представления данных при одновременном сохранении максимальной части заложенной в них информации.

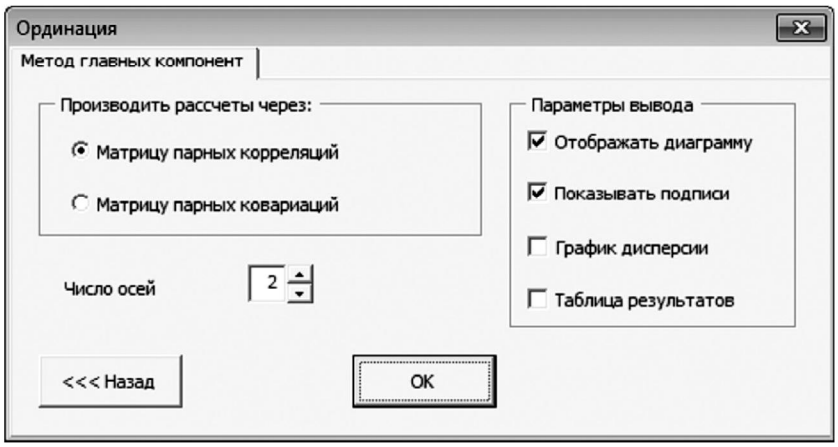


Рис. 4. Диалоговое окно настроек для PCA анализа (метод главных компонент).

В окне настроек PCA анализа пользователь задает тип анализируемой матрицы (корреляция или ковариация), число осей ординации и параметры отображения результатов (рис. 4).

Второй метод – неметрического шкалирования – относится к принципиально иным методам снижения размерностей. Он основан на минимизации некоторой функции стресса, сравнивающей между собой попарные расстояния в исходном пространстве признаков изучаемых объектов с евклидовыми расстояниями в новом (уменьшенном) пространстве. Чем меньше значение стресса, там лучше точки в новом пространстве отображают взаимное расположение анализируемых объектов в исходном пространстве. Стресс, равный нулю, обозначает полную тождественность сравниваемых матриц. Соответственно, задача ординации сводится к подбору таких координат точек в новом пространстве, чтобы величина стресса между модельной и эмпирической матрицами была минимальной (Kruskal, 1964; Clarke, 1993).

В окне настроек (рис. 5) пользователь выбирает тип входной матрицы. В случае, когда выбрана многомерная выборка, на вход подается стандартная таблица с данными, на основе которой строится матрица различий. При этом для построения используется один из коэффициентов сходства/различия, выбираемый в соответствующем разделе. На сегодняшний день в надстройке реализованы евклидово и манхеттоновское расстояния, количественные коэффициенты сходства Жаккара и Съеренсена-Чекановского, коэффициент корреляции Пирсона.

Кроме многомерной выборки пользователь при выборе типа входной матрицы может указать матрицу сходств (на вход подается квадратная симметричная матрица, диагональные элементы которой равны 1, все недиагональные элементы варьируются от 0 до 1) или же матрицу расстояний (диагональные элементы равны 0, недиагональные – варьируют от 0 до 1).

Результатом работы ординационных алгоритмов является диаграмма взаимного расположения объектов в новом (уменьшенном) пространстве (рис. 6) и табличное представление результатов (рис. 7). В таблице содержатся обобщенные статистики, описывающие качество проведенной ординации и полученные координаты точек в новом пространстве.

Другим распространенным методом анализа данных, применяемым в экологии, является кла-

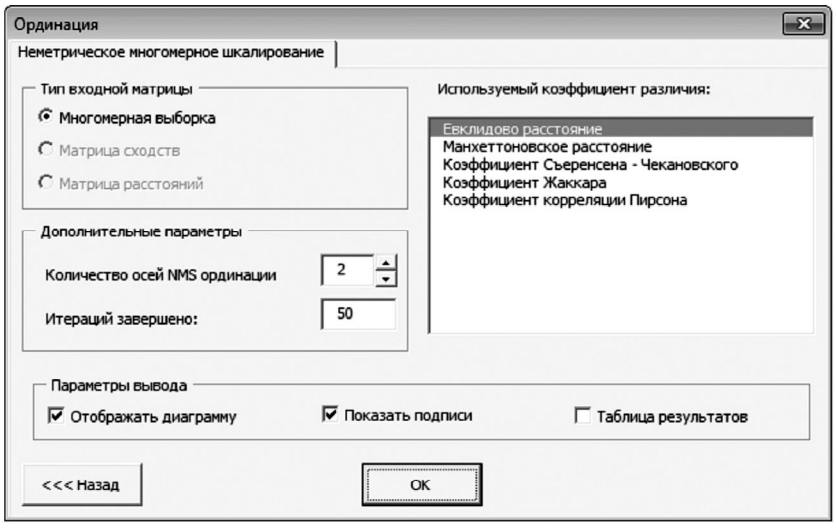


Рис. 5. Диалоговое окно настроек NMS анализа (многомерное неметрическое шкалирование).

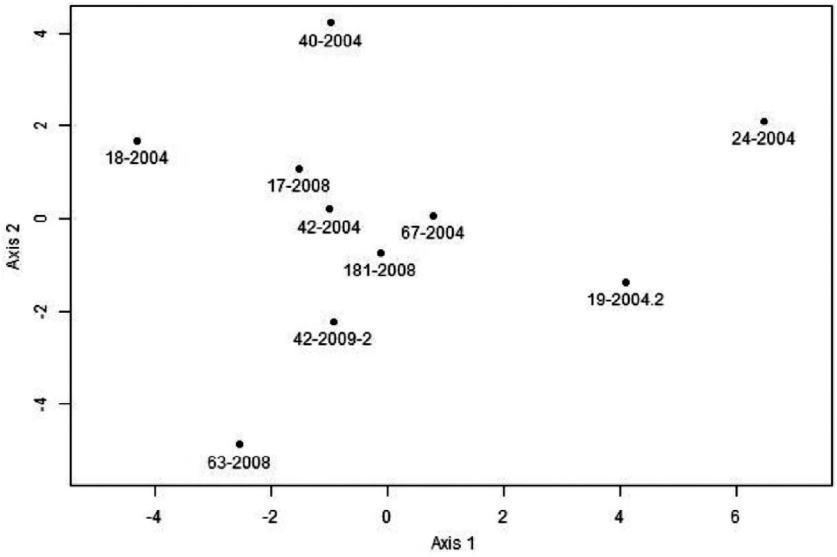


Рис. 6. Результирующая диаграмма NMS ординации.

	A	B	C
1	Результаты анализа главных компонент		
2	Используется матрица парных корреляций		
3			
4		PCA 1	PCA 2
5	Значения собственных векоров (дисперсия по осям)	4.624689	2.699556
6	Доля объясненной дисперсии от общей суммы (в процентах)	27.20406	15.87974
7			
8	42-2004	0.80932	-0.22458
9	181-2008	1.562143	1.153166
10	67-2004	0.564014	-0.48835
11	18-2004	2.264484	-1.90679
12	19-2004.2	-1.1467	-0.33939
13	63-2008	-1.1776	3.729382
14	17-2008	1.885857	-1.01031
15	42-2009-2	-0.2976	1.662671
16	24-2004	-5.50271	-1.8424
17	40-2004	1.038798	-0.73341
18			
19	Корреляция дополнительных факторов с осями ординации		
20	Фактор 1	-0.48685	-0.19734
21	Фактор 2	-0.30056	-0.62811

Рис. 7. Табличное представление результатов PCA ординации.

стерный анализ (Песенко, 1982; Шитиков, 2003; Пузачено, 2004). В целом, под кластерным анализом (кластеризацией) понимают задачу разбиения всей совокупности рассматриваемых объектов на отдельные группы (классы) со сходными характеристиками и определение взаимных отношений между ними (Ward, 1963; Уиллиамс, 1986; Jongman, 1987; Ким, 1989). Алгоритмы кластерного анализа можно разделить на два основных типа: иерархические и не иерархические. В модуле ExcelToR реализованы наиболее распространенные алгоритмы первого типа, поскольку иерархический кластерный анализ позволяет графически (в виде дендрограмм) отобразить полученные результаты. Кроме того, он показывает не только разбиение объектов на группы, но и их взаимное расположение (иерархию).

Основой иерархического кластерного анализа является определение расстояния между объектами посредством различных мер сходства/различия и способ их группировки (агломерации). В надстройке ExcelToR реализованы наиболее часто используемые в экологии меры сходства/различия: Жаккара, Съеренсена-Чекановского, евклидово и манхеттоновское расстояние, коэффициент корреляции Пирсона и ранговой корреляции Кэндела (Андреев, 1980; Песенко, 1982; Шитиков, 2003). Реализованы следующие методы группировки: метод среднего расстояния (UPGMA – unweighted pair-group method using arithmetic averages, average-linkage clustering); максимального расстояния (Complete-linkage clustering); метод Ward-a (Ward's method).

В диалоговом окне (рис. 8) пользователь задает тип исходной матрицы. Это может быть как матрица объектов (в этом случае необходимо выбрать метод расчета сходства/расстояния между анализируемыми объектами), так и рассчитанные заранее любым удобным способом матрицы сходства или расстояния, аналогичные тем, что используются в NMS ординации. Пользователь задает метод группировки данных и дополнительные настройки, влияющие на внешний вид результирующей дендрограммы (рис. 9).

Наряду с построением ординационных диаграмм или дендрограмм, в надстройке ExcelToR существует возможность наложения дополнительных экологических факторов. Наложение факторов проходит либо в виде раскраски объектов в соответствии с выбранными группами (используется при кластерном анализе и ординации), либо в виде векторов, отражающих корреляцию между изучаемыми экологическими факторами и осями ординации (используется только для ординации) (рис. 10).

Построение теплокарт (рис. 11 и 12) скорее является способом визуализации данных, а не специализированным методом их анализа. Его суть состоит в том, что строится прямоугольная сетка, в клетках которой цветом отображаются числовые значения. Минимальным значениям соответствует один цвет, максимальным – другой. Промежуточные значения окрашиваются по цветам градиента.

В окне настроек (рис. 11) пользователь задает область с подписями по осям X и Y, область данных, которая содержит числовые значения, отображаемые цветами, выбирает цветовую схему и число градаций цвета на градиенте.

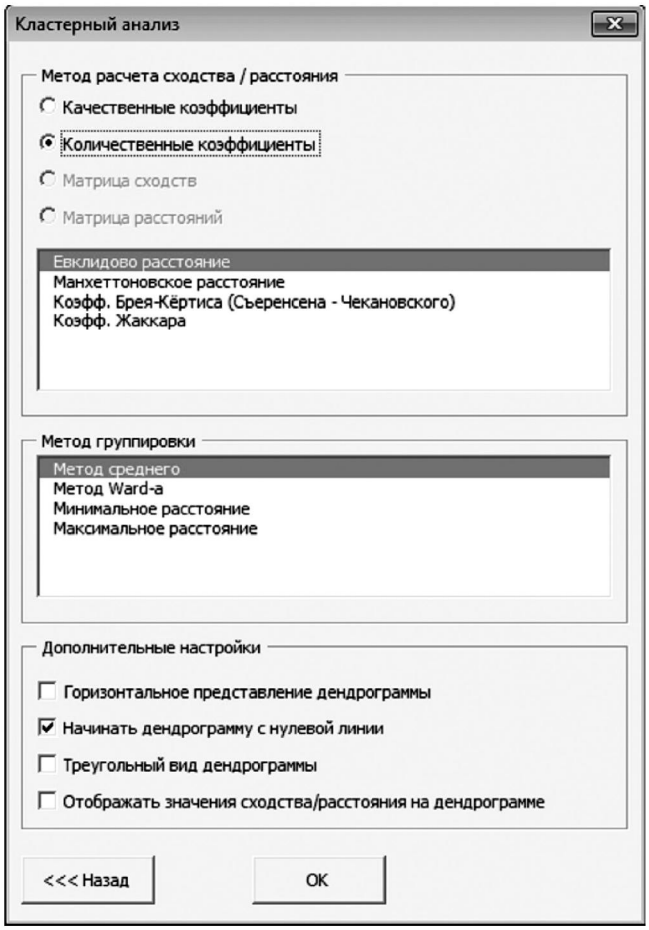


Рис. 8. Диалоговое окно настроек кластерного анализа.

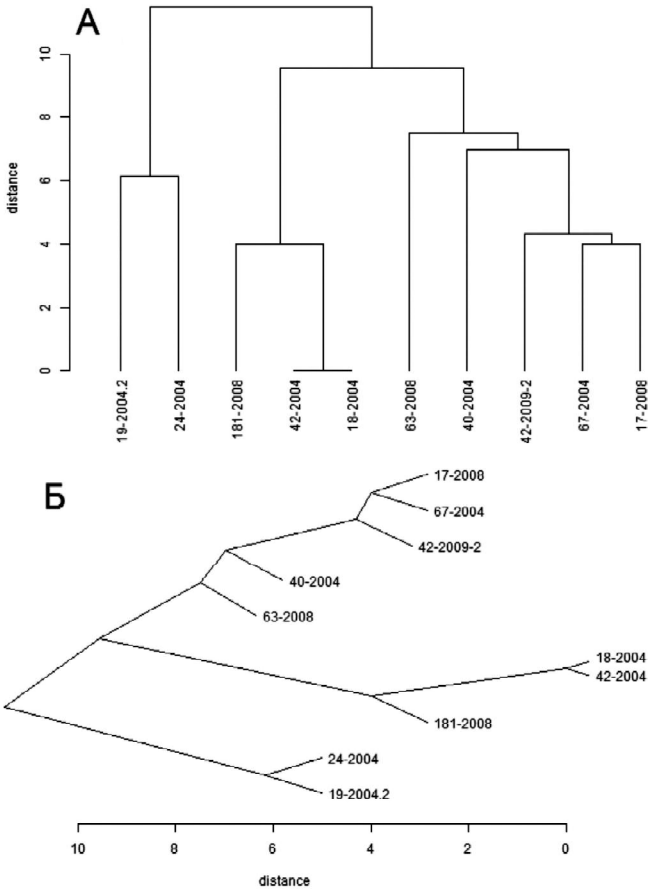


Рис. 9. Различие внешнего вида результирующих дендрограмм для одного и того же набора данных с полностью идентичными способами анализа. А – вертикальное отображение, представление в виде прямоугольников, элементы начинаются с 0. Б – горизонтальное отображение, представление в виде треугольников, подвешенные элементы.

В качестве примера приведем результат использования метода для анализа температурных данных по европейской части России (рис. 12). На теплокарте приведены отклонения температур определенных периодов от средних значений. Области желтого, оранжевого и красного цвета на теплокарте означают, что текущие показатели выше (теплее) усредненных значений, а синий и голубой – температурные показатели ниже (холоднее) среднего значения. Такая форма представления данных позволяет визуально определить области с повышенными и пониженными температурами и выявить не только внутригодовую, но и межгодовую температурную динамику (Novakovskiy, 2014).

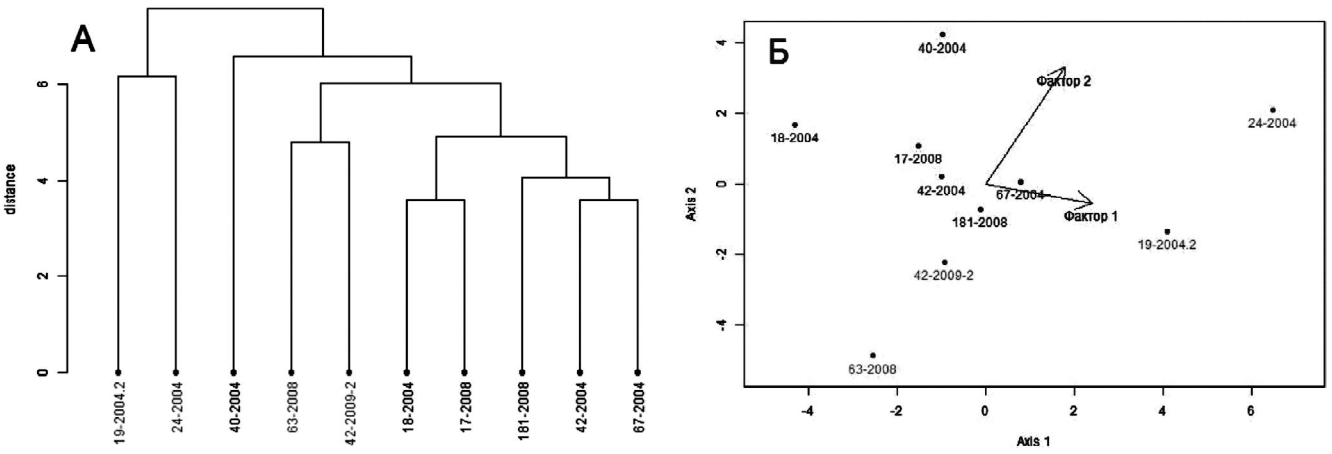


Рис. 10. Пример наложения дополнительных экологических факторов на результаты обработки данных. А – наложение различной окраски элементов для кластерного анализа; Б – наложение дополнительных факторов в виде раскраски и корреляционных векторов на результирующую ординационную диаграмму.

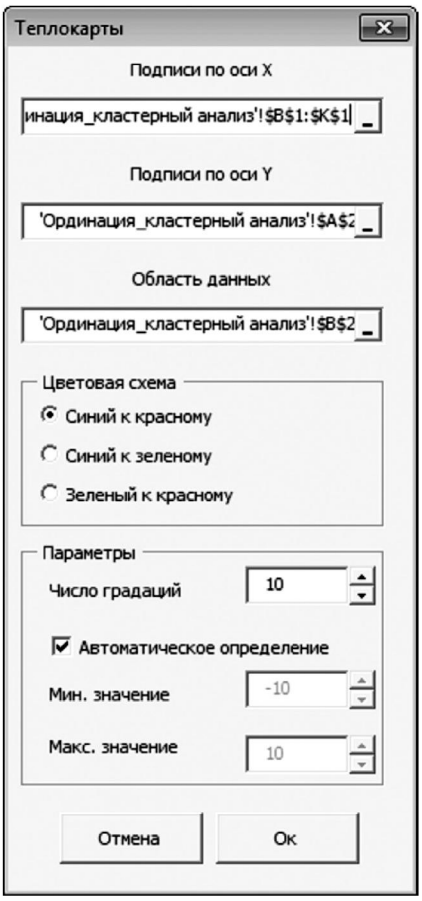


Рис. 11. Диалоговое окно настроек для построения теплокарт.

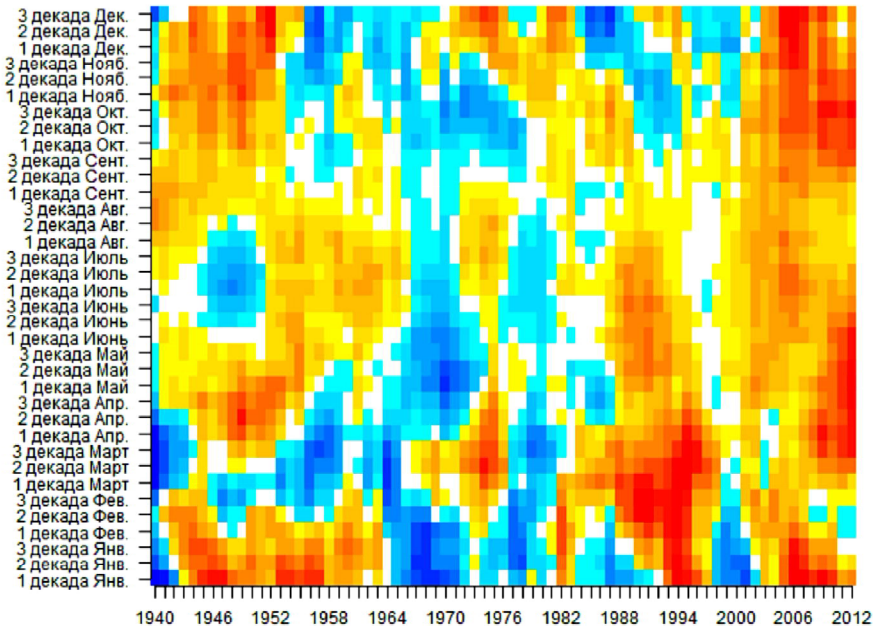


Рис. 12. Построенная теплокарта отклонений среднедекадных температур от многолетних средний значений для европейской части России.

Таким образом, созданная нами программно-аналитическая надстройка ExcelToR расширяет возможности Microsoft Excel по статистической обработке экологических данных. Универсальность предложенных алгоритмов позволяет использовать модуль для решения задач и в других областях науки, не связанных с экологией и биологией. Надстройка проста в использовании и не требует специальной подготовки данных.

Отметим, что надстройка ExcelToR является продолжением нашей работы по автоматизации обработки экологических данных. Ранее для этих целей нами был разработан оригинальный программный модуль «GRAPHS» (Новаковский, 2004, 2006). Программа ExcelToR – это новый программный продукт, основанный на иных принципах. Вычислительным ядром надстройки является статистический пакет R, что позволяет использовать все его возможности в области расчетов и визуализации данных. Кроме того, разработанная система состоит из взаимно независимых модулей и построена таким образом, чтобы быстро модернизировать существующие алгоритмы и добавлять новые, реализованные в пакете R.

Надстройка является свободно распространяемой. Скачать установочную версию ExelToR и подробную инструкцию по использованию можно по адресу http://ib.komisc.ru/exceltor.

ЛИТЕРАТУРА

Андерсон, Т. Введение в многомерный статистический анализ / Т. Андерсон. – М. : Физматгиз, 1963. – 500 с.

Андреев, В. Л. Классификационные построения в экологии и систематике / В. Л. Андреев. – М. : Наука, 1980. – 142 с.

Андрукович, П. Ф. Применение метода главных компонент в практических исследованиях / П. Ф. Андрукович. – М. : Изд-во московского университета, 1973. – 124 с.

Бигон, М. К. Экология. Особи, популяции и сообщества: В 2-х т. Т. 1. Пер. с англ. / М. Бигон, Дж. Харпер, К. Таунсенд. – М. : Мир, 1989. – 667 с.

Василевич, В. И. Статистические методы в геоботанике / В. И. Василевич. – Л. : Наука, 1969. – 232 с.

Зиновьев, А. Ю. Визуализация многомерных данных / А. Ю. Зиновьев. – Красноярск : Изд-во Красноярского государственного технического университета, 2000. – 180 с.

Ким, Дж. О. Факторный, дискриминантный и кластерный анализ / Дж. О. Ким; отв. ред. И. С. Енюков. – М. : Финансы и статистика, 1989. – 215 с.

Лебедева, Н. В. Биологическое разнообразие и методы его оценки / Н. В. Лебедева, Д. А. Криволуцкий // География и мониторинг биоразнообразия. – М. : НУМЦ, 2002. С. 8-76.

Мастицкий, С. Э. Статистический анализ и визуализация данных с помощью R [Электронный ресурс] / С. Э. Мастицкий, В. К. Шитиков. – Хайдельберг-Лондон-Тольятти, 2014. – Режим доступа: http://r-analytics.blogspot.ru/2014/12/r.html.

Нешатаев, Ю. Н. Выборочно-статистический методы выделения растительных ассоциаций / Ю. Н. Нешатаев // Методы выделения растительных ассоциаций. – Л. : Наука, 1971. – С. 181-205.

Новаковский, А. Б. Возможности и принципы работы программного модуля «GRAPHS» / А. Б. Новаковский. – Сыктывкар, 2004. – 31 с. – (Сер. Автоматизация науч. исследований / Коми НЦ УрО РАН; Вып. 27).

Новаковский, А. Б. Обзор современных программных средств, используемых для анализа гео-

ботанических данных / А. Б. Новаковский // Растительность России. – 2006. – № 9. – С. 86-96.

Одум, Ю. Экология: В 2-х т. Т. 1. Пер. с англ. / Ю. Юдум. – М. : Мир, 1986. – 328 с.

Песенко, Ю. А. Принципы и методы количественного анализа в фаунистических исследованиях / Ю. А. Песенко. – М. : Наука, 1982. – 287 с.

Пузаченко, Ю. Г. Математические методы в экологических и географических исследованиях / Ю. Г. Пузаченко. – М. : Изд-во Академия, 2004. – 416 с.

Уиллиамс, У. Т. Методы иерархической классификации / У. Т. Уиллиамс, Дж. Н. Ланс // Статистические методы для ЭВМ / Под ред. М. Б. Малютов. – М. : Наука, 1986. – С. 269-301.

Шитиков, В. К. Количественная гидроэкология: методы системной идентификации / В. К. Шитиков, Г. С. Розенберг, Т. Д. Зинченко. – Тольятти : ИЭВБ РАН, 2003. – 463 с.

Экологическая оценка кормовых угодий по растительному покрову / Л. Г. Раменский, И. А. Цаценкин, О. Н. Чижиков, Н. А. Антипин. – М. : Сельхозгиз, 1956. – 472 с.

Borcard, D. Numerical Ecology with R / D. Borcard, F. Gillet, P. Legendre. – N.Y. : Springer, 2011. – 319 p.

Bray, J. R. An ordination of upland forest communities of southern Wisconsin / J. R. Bray, J. T. Curtis // Ecological monographs. – 1957. – Vol. 27. – P. 325-349.

Clarke, K. R. Non-parametric multivariate analyses of changes in community structure / K. R. Clarke // Austral. J. Ecol. – 1993. – Vol. 18. – P. 117-143.

Cox, T. F. Multidimensional scaling (2nd edition) / T. F. Cox, M. A. Cox. – Chapman and Hall, 2001. – 294 p.

Dunteman, G. H. Principal Component Analysis / G. H. Dunteman. – Iowa : SAGE Publications, 1989. – 96 p.

Hammer, O. Past: paleontological statistics software package for education and data analysis / O. Hammer, D. A. T. Harper, P. D. Ryan // Palaeontologia Electronica. – 2001. –Vol. 4(1). – 9 p.

Hill, M. O. DECORANA and TWINSPAN for ordination and classification of multivariate species data: a new edition, together with supporting programs, in FORTRAN 77 / M. O. Hill. – Huntingdon : Institute of Terrestrial Ecology, 1994. – 58 p.

Hill, M. O. DECORANA – a FORTRAN program for detrended correspondence analysis and reciprocal averaging / Hill M. O. – N.Y. : Cornell University, Ithaca, 1979. – 31 p.

Jongman, R. H. G. Data analysis in community and landscape ecology / R. H. G. Jongman, C. J. F. Ter Braak, O. F. R. Van Tongeren. – Wageningen, 1987. – 299 p.

Kabacoff, R. I. R in Action. Data analysis and graphics with R. 2nd ed. / R. I. Kabacoff. – N.Y. : Manning, 2015. – 608 p.

Kruskal, J. B. Nonmetric multidimensional scaling: A numerical method / J. B. Kruskal // Psychometrica. – 1964. – Vol. 29(2). – P. 115-130.

Legendre, P. Numerical Ecology. 2nd ed. / P. Legendre, L. Legendre. – Amsterdam, 1998. – 853 p.

Leps, J. Multivariate Analysis of Ecological Data / J. Leps, P. Smilauer. – Ceska Budejovice, 1999. – 110 p.

McCune, B. Analysis of ecological communities / B. McCune, J. B. Grace, D. L. – Oregon : Urban MjM Software Design, 2002. – 285 p.

Novakovskiy, A. Hydrometeorological database (HMDB) for practical research in ecology / A. Novakovskiy, V. Elsakov // Data Science Journal. – 2014. – Vol. 13. – P. 57-63.

Seefeld, K. Statistics Using R with Biological Examples / K. Seefeld, E. Linder. – Durham : University of New Hampshire, 2007. – 325 p.

Shepard, R. N. Analysis of proximities: Multidimensional scaling with an unknown distance function I & II / R. N. Shepard // Psychometrica. – 1962. – Vol. 27. – P. 125-140, 219-246.

Ter Braak, C. J. F. Canonical community ordination. Part I: Basic theory and linear methods / C. J. F. Ter Braak // Ecoscience. – 1994. – Vol. 1. – P. 127-140.

Tichy, L. JUICE, software for vegetation classification / L. Tichy // Journal of Vegetation Science. – 2002. – Vol. 13. – P. 451-453.

Ward, J. H. Hierarchical grouping to optimize an objective function / J. H. Ward // Journal of the American statistical association. – 1963. – Vol. 58, № 301. – P. 236-244.

INTERACTION BETWEEN EXCEL AND STATISTICAL PACKAGE R FOR ECOLOGICAL DATA ANALYSIS

A.B. Novakovskiy

Institute of Biology of Komi Scientific Centre of the Ural Branch of the Russian Academy of Sciences, Syktyvkar

Abstract. The add-on ExcelToR which design for joint use Microsoft Excel and statistical package R is considered in this paper. Microsoft Excel is used for input, storage and preparing data for analysis. Statistical package R is a processing «core» for analysis. Special separate modules which standardize the data transferring from/to Excel to R were developed. This approach allows implementing/modifying necessary algorithms for statistical analysis or data visualization fast and easy.

Now following algorithms of statistical analysis have been designed in ExcelToR: clustering (Single-linkage clustering и Complete-linkage clustering, Ward's method, UPGMA), ordination (PCA, NMS), heat maps.

Key words: statistical analysis, Excel, R, clustering, ordination, heat maps