И.Ю. Сёмочкина, О.В. Прокофьев

ПРИМЕНЕНИЕ ЯЗЫКА R И СРЕДЫ RSTUDIO ДЛЯ МАТЕМАТИЧЕСКОЙ ОБРАБОТКИ ДАННЫХ

Россия, г. Пенза, Пензенский государственный технологический университет,
Пензенский филиал Финансового университета при Правительстве РФ

The article analyzes the possibilities of using the R programming language and the RStudio environment for mathematical data processing. The importance of the central storage and distribution system for CRAN packets is reflected. The requirements to system software, to the level of mastering of professional disciplines by students are considered. The features of the interface, advantages and disadvantages are listed for the user-teacher, engaged in the development of practical teaching methods.

Подготовка магистерской диссертации в области экономических информационных систем нередко связана с необходимость использования интеллектуального анализа данных [1], нечёткой логики [2], эконометрических моделей [3, 4], реализуемых в виде исходных кодов процедур. Коммерческие продукты мирового уровня Statistica, IBM SPSS и подобные требуют существенных материальных затрат даже в рамках льготных программ поддержки высшего образования. Поэтому актуален выбор и использование альтернативных свободно распространяемых программных продуктов, обладающих лицензией GNU General Public License.

Одним из продуктов, отвечающих высоким современным требованиям, является RStudio - свободная среда разработки программного обеспечения с открытым исходным кодом для языка программирования R, который предназначен для статистической обработки данных и работы с графикой [5-9]. В частности, в данной статье задачей исследования являлась оценка степени применимости среды моделирования R для эконометрического моделирования с точки зрения правового подхода, профессионального уровня.

Пакет прикладных программ R обладает универсальной общедоступной лицензией на свободное программное обеспечение GNU General Public License третьей версии (2007 г.). Она предоставляет пользователю права копировать, модифицировать и распространять (в том числе на коммерческой основе) программы, а также гарантировать, что и пользователи всех производных программ получат вышеперечисленные права.

R — это одновременно и свободно распространяемая программная среда с открытым кодом, развиваемая в рамках проекта GNU, и язык программирования для статистической обработки данных и работы с графикой. R можно бесплатно скачать на сайте проекта http://www.r-project.org и применять везде, где нужна работа с данными. Это и сама математическая статистика во всех её приложениях, и первичный анализ данных, и эконометрическое моделирование. R широко используется как статистическое программное обеспечение для анализа данных и фактически стал стандартом для статистических, эконометрическое программ.

Дополнительную популярность R принесло создание центральной системы хранения и распространения пакетов — CRAN (Comprehensive R Archive Network — http://cran.r-project.org), расширяющих возможности базового продукта. На момент написания статьи в системе было размещено 10635 дополнительных пакетов (CRAN Packages). Например, в названиях пакетов понятие «кластер» (Cluster) упоминается 326 раз. Основная мощь R лучше всего проявляется при многомерном статистическом анализе данных, эконометрическом анализе, операций с временными рядами. С помощью R

можно подготовить данные для исследования, которое может быть осуществлено с помощью реализованных в различных функциях статистических методов, а затем вывести полученные результаты для дальнейшего анализа. Сейчас практически во всех западноевропейских и американских университетах изучают и используют R, ежегодно издаются многостраничные учебники и монографии относительно как работы с самим пакетом R, так и его применения при исследовании и обработке данных в прикладных областях.

Особенностью R является интерфейс командной строки, хотя доступны и несколько графических интерфейсов пользователя (коммерческих и бесплатных), например, в системах R Commander, RKWard, RStudio, Weka, Rapid Miner, KNIME[en], а также в средствах интеграции в офисные пакеты.

Существуют версии R for Linux, R for (Mac) OS X, R for Windows. Текущие бинарные версии R запускать на Windows XP или более поздней версии, в том числе на 64-битных версиях. Установка занимает до 150 МБ дискового пространства и не требует специальных навыков.

К перечисленным достоинствам можно добавить:

– обработку массивов данных до несколько сотен тысяч наблюдений;

– наличие встроенной системы помощи и подсказок;

– хорошие графические возможности представления результатов исследований;

– возможность самостоятельного написания необходимых функций;

– наличие свободно распространяемой литературы по R.

Недостатки, отмечаемые рядом пользователей:

– в отличие от большинства коммерческих программ, R имеет не графический интерфейс, а интерфейс командной строки, таким образом, нужно знать необходимые для работы функции и синтаксис языка программирования;

– нет коммерческой поддержки (но есть международная система рассылки сообщений об обновлениях);

– недостаточное количество учебной литературы по R на русском языке.

Среда разработки программного обеспечения с открытым исходным кодом RStudio доступна в двух версиях: RStudio Desktop, в которой программа выполняется на локальной машине как обычное приложение; и RStudio Server, в которой предоставляется доступ через браузер к RStudio установленной на удаленном Linux-сервере. Дистрибутивы RStudio Desktop доступны для Linux, OS X и Windows. RStudio представляет собой бесплатную интегрированную среду разработки (IDE) для R. Благодаря ряду своих особенностей этот активно развивающийся программный продукт делает работу с R удобной для учебного процесса с подготовленными слушателями. Среда для разработки позволяет использовать не только стандартные процедуры статистической обработки данных, но и процедуры и функции из дополнительных пакетов библиотеки (CRAN Packages, доступ http://cran.r-project.org). На момент написания статьи международным сообществом исследователей было размещено в библиотеке 10635 дополнительных пакетов на языке R, охватывающих всевозможные разделы эконометрического анализа, прогнозирования временных рядов, многомерной статистики, причём каждый из пакетов сопровождается документацией по установленному стандарту.

Среда RStudio изображена на рисунке 1.

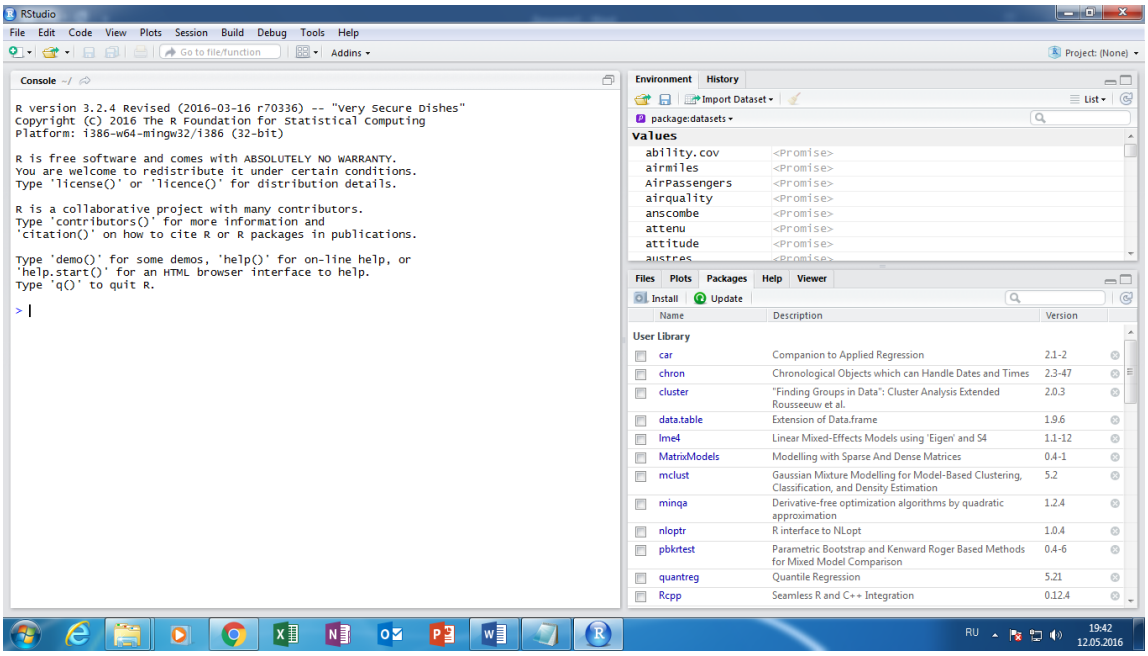

Рис.1. Окна интерфейса пользователя RStudio

Консоль RStudio (Console) предоставляет целый ряд опций, делающих работу с R простой и продуктивной. Освоение этих опций, наряду с возможностями, доступными в панелях Source (Редактор кода) и History (История), может оправдать затраченное на обучение время.

Редактор кода RStudio включает ряд опций для продуктивной работы, в частности подсветку кода, автоматическое завершение кода, одновременное редактирование нескольких файлов, поиск и замену определенных частей кода. Кроме того, в RStudio имеются гибкие возможности по выполнению кода непосредственно из окна редактора. Для многих учащихся это является предпочтительным способом работы с R.

RStudio поддерживает «подсветку» синтаксиса и другие специализированные опции по работе с кодом следующих типов файлов: R-скрипты, документы Sweave, документы TeX.

RStudio поддерживает выполнение кода непосредственно из окна Редактора (выполняемые команды посылаются в Консоль, где появляется также результат их выполнения).

RStudio включает ряд опций, обеспечивающих быструю навигацию по R-коду. Во время работы RStudio создает базу данных всех команд, которые пользователь вводит в Консоль. Имеется возможность просмотра этой базы данных при помощи панели History (История).

Если все файлы, имеющие отношение к определенному проекту, хранятся в одной папке, имеет смысл сделать ее исходной для работы. RStudio автоматически будет делать рабочей папкой ту, в которой хранится открываемый файл. RStudio позволяет организовать работу в соответствующие контексту проекты так, что каждый проект будет иметь свою собственную рабочую директорию, рабочее пространство, историю и скрипты. Проекты RStudio ассоциированы с рабочими директориями R. Следовательно, проект можно создать: в новой директории; в существующей директории, где уже хранятся скрипты с R-кодом и данные; путем копирования файлов, хранящихся в одной из онлайн-систем контроля версий. Для создания нового проекта служит команда New Project (Новый Проект), доступная из закладки Projects главного меню и из панели инструментов (в дальнем правом углу рабочего окна программы). Имеется несколько опций для настройки поведения каждого конкретного проекта в RStudio. Эти опции доступны по команде Project Options из раздела Project главного меню программы.

Краткое перечисление этих возможностей позволяет сделать вывод о доступности применения среды RStudio в учебном процессе и научном исследовании для магистрантов, имеющих базовую подготовку в области программирования и общей теории статистики. Таким образом, язык R, система программирования и отладки R-Studio, библиотека CRANE по научному уровню и диапазону возможностей составляют конкуренцию коммерческим продуктам, могут быть использованы для укрепления междисциплинарных связей, в написании выпускной квалификационной работы. В то же время выявлено требование по отличному уровню освоения базового курса статистики на этапе бакалавриата и по наличию навыков программирования, включая применение объектно-ориентированной технологии в написании исходных кодов. Ограничения приводят к выводу об индивидуальном подходе к применению этого программного обеспечения, с учётом возможностей и потребностей учащегося магистратуры. Тем не менее, в случае оправданного использования, появляется возможность использовать потенциал научного роста обучаемого, введения готовых или самостоятельно написанных исходных кодов в выпускную квалификационную работу, ознакомления магистранта с мировым уровнем достижений в области интеллектуального анализа данных.

1. Прокофьев О.В., Семочкина И.Ю. Применение интеллектуального анализа данных в системах поддержки принятия решений для управления профессиональным образованием. Современные информационные технологии. - 2012. - № 16. - С. 140-143.
2. Прокофьев О.В., Семочкина И.Ю. Моделирование управления образовательным процессом на основе методологии нечёткой логики. XXI век: итоги прошлого и проблемы настоящего плюс. - 2011. - № 3 (03). - С. 155-161.
3. Прокофьев О.В., Савочкин А.Е. Алгоритмическая модификация теста Чоу для автоматизированной проверки гипотезы о структурной стабильности тренда. XXI век: итоги прошлого и проблемы настоящего плюс. - 2014. - № 3 (19). - С. 183-188.
4. Прокофьев О.В. Формы проявления структурной нестабильности временного ряда в современных статистических данных. В сборнике: Проблемы экономики, организации и управления в России и мире Материалы V международной научно-практической конференции. Ответственный редактор Уварина Н.В. - 2014. - С. 257-258.
5. How_to_work_with_IDE_RStudio. Электронная книга, адрес доступа: http://rstudio.org.
6. Мастицкий С.Э., Шитиков В.К. (2014) Статистический анализ и визуализация данных с помощью R. – Электронная книга, адрес доступа:
http://r-analytics.blogspot.com.
7. Статистический анализ данных в системе R. Учебное пособие /А.Г.
Буховец, П.В. Москалев, В.П. Богатова, Т.Я. Бирючинская; Под ред. проф. Буховца А.Г. — Воронеж: ВГАУ, 2010. — 124 с.
8. Теория пространственных точечных процессов в задачах экологии и природопользования (с применением пакета R): учебное пособие / сост.: А.А. Савельев, С.С. Мухарамова, Н.А. Чижикова, А.Г. Пилюгин. – Казань: Изд-во Казан. ун-та, 2014. – 146 с.
9. Зарядов И.С. Введение в статистический пакет R: типы переменных, структуры данных, чтение и запись информации, графика. Учебно-методическое пособие. – М.: Издательство Российского университета дружбы народов. - 2010. – 207 с.