УДК 681.518.25:004.67

Рынок программного обеспечения для анализа временных рядов

Иващенко А.Б.
Донецкий национальный технический университет
alesya_iva@list.ru

Abstract

Ivashchenko A. "The software market for time series analysis" *In present paper a review of the software products for time series was made. The two groups of software products were highlighted: universal mathematical packages and specialized software. The most promising and popular software solutions for the analysis of time series are considered. The main features and capabilities of the software products are described. Their strengths and weaknesses are identified.*

Keywords: time series, data analysis, forecasting software, mathematical packages.

Введение

Анализ данных – одно из наиболее перспективных направлений современных исследований. Спрос на программные средства для анализа данных постоянно растет. И как следствие, всё стремительнее развивается рынок предложений такого программного обеспечения. На основании обобщенной информации, представленной в [1], можно сделать вывод, что на сегодняшний день для анализа данных уже разработаны десятки программных продуктов. Различные по объему и качеству реализованных методов, области возможного применения, пользовательскому интерфейсу, цене, требованиям к оборудованию и т.п., они отражают многообразие потребностей обработки данных в различных областях человеческой деятельности.

Большинство таких программ предназначены непосредственно для статистической обработки табличных данных и реализуют стандартные методы статистического анализа данных. Сравнительному анализу наиболее популярных программ посвящены обзоры [2, 3]. В них программные продукты рассмотрены, преимущественно, в плане их применимости к статическим данным. Целью настоящей работы является изучение программного обеспечения, позволяющего осуществлять анализ временных рядов.

В результате проведенного анализа был выделен ряд программных продуктов с точки зрения их эффективного использования в части анализа динамических рядов. Условно их можно разделить на две группы: универсальные математические пакеты и специализированные программы.

Универсальные математические пакеты

Компьютерные системы для анализа данных – пакеты статистических программ – считаются наукоемкими программными продуктами, и, пожалуй, наиболее широко применяются в практической и исследовательской работе в самых разнообразных областях [4].

Очень качественное и детальное сравнение универсальных математических пакетов для анализа данных было выполнено Штефаном Штейнхаусом[5]. Его первые отчеты о результатах этих сравнений были опубликованы еще в 1997 году. Последние версии отчетов в этом направлении датируются 2002 и 2008 годами. Штейнхаус исследует следующие программы:

- Gauss, производитель Aptech Systems Inc (www.aptech.com);
- Maple, производитель Waterloo Maple Software Inc (www.maplesoft.com);
- Mathematica, производитель Wolfram Research Inc (www.wolfram.com);
- Matlab, производитель The Mathworks Inc (www.mathworks.com);
- O-Matrix, производитель Harmonic Software (www.omatrix.com);
- OxMetrics (Ox Prof.), производитель Timberlake Consultants Ltd (www.oxmetrics.net);
- Scilab, производитель INRIA (www.scilab.org).

Также в более ранних версиях сравнительного анализа участвовали еще такие пакеты, как Macsyma (производитель Macsyma Inc.), MuPAD (производитель the University of Paderborn) и S-Plus (производитель Insightful Inc).

Сравнительный анализ осуществлялся на основании различных показателей: математическая функциональность (алгебраические выражения, математический анализ, численные методы, вероятностные распределения, описательные статистики и т.д.), функциональность программной среды, графическая функциональность, обработка данных, совместимость с операционными системами, скоростные показатели и т.п.

Интересен тот факт, что «призовые места» стабильно занимала неизменная тройка лидеров: Gauss, Matlab и Mathematica, периодически обгоняя друг друга из года в год и меняясь местами. Это и неудивительно, ведь спустя годы эти пакеты до сих пор очень популярны и продолжают развиваться.

К вышеупомянутым лидерам можно добавить такой популярный продукт, как Statistica (производитель StatSoft). Несмотря на то, что Statistica не была включена в сравнительный анализ Штейнхауса, это программное обеспечение специализируется непосредственно на анализе данных (в том числе и анализе временных рядов) и его использование широко распространено. Для анализа временных рядов данный пакет позволяет оперировать лишь авторегрессионными методами.

Основным недостатком этих пакетов можно считать тот факт, что все эти программные продукты не из дешевых (цены за базовый пакет от 10 000 грн).

На сегодняшний день среди отмеченных программ наиболее эффективным для проведения научных исследований и инженерных расчетов в части анализа временных рядов является Matlab. В сравнении с остальными он достаточно гибкий для программирования: имеет необходимый арсенал готовых встроенных функций и позволяет разрабатывать и внедрять собственные алгоритмы. Однако в Matlab есть ряд проблем, возникающих при работе с большими объемами данных.

Специализированное программное обеспечение

За последние годы рынок программного обеспечения существенно дополнился новыми программными средствами для специализированного анализа временных рядов. Это связано, в первую очередь, с тем, что крупные универсальные пакеты, несмотря на высокую стоимость, содержат лишь традиционные методы анализа данных и не располагают современным инструментарием для анализа и прогнозирования временных рядов.

Красочные страницы веб-сайтов пестрят картинками с изображением графиков, трендов, диаграмм и т.п., а также обещаниями, что именно их программный продукт лучше других анализирует данные и строит прогнозы. Тем не менее, возможности большинства предлагаемых на сегодня программных средств не выходят за рамки заурядного статистического анализа и, как отмечалось ранее, традиционного анализа временных рядов (авторегрессия, выделение тренда, сезонности и т.п.).

Даже при беглом обзоре информации, доступной в просторах Интернета, можно отметить, что появилось множество компаний, специализирующихся на разработке и продаже программных средств и услуг в области бизнес-аналитики. Так, среди платного программного обеспечения для анализа временных рядов и формирования прогнозов на его основе, в основном ориентированного на бизнес-применение, встречаются такие предложения, как, например, CATS, SAS Analytics, Autobox, Forecast Pro, SHAZAM и др. Для программ такого типа характерны высокая стоимость, минимальный набор стандартных методов для анализа и прогнозирования данных, хорошо разработанный модуль визуализации данных (графики, диаграммы и т.п.) и удобный интерфейс.

Альтернативой этим пакетам являются не менее эффективные программные продуты, но распространяющиеся бесплатно, например, Gretl, Zaitun Time Series, ARfit и т.п. Недостатками бесплатных продуктов является слабо разработанный интерфейс, и как следствие, высокие требования к уровню знаний и подготовленности пользователя.

Отдельного внимания заслуживает программное обеспечение, реализующее специфические методы или уникальные методики анализа временных рядов. Такие программы обычно распространяются на платной основе (например, Dataplore, АСТРИД, GMDH Shell, DTREG и др.), хотя все чаще встречаются достойные бесплатные программные продукты (например, TISEAN и Fractan).

В результате анализа информации, представленной в Интернете, были выявлены около двадцати программных продуктов, реализующих перспективные алгоритмы и современные методы анализа временных рядов. Результаты проведенного информационного поиска и анализа информации, имеющейся в Интернете, структурированы в виде таблицы (табл. 1).

Объективно говоря, большая часть представленных платных программных продуктов не предлагает ничего особенного и тем более не стоит тех денег, которые за них просят. Тем не менее, после проведенного анализа можно выделить всего несколько действительно мощных и весомых игроков на рынке программных продуктов для анализа и прогнозирования по временным рядам: SAS, GRETL, GMDH Shell, DTREG, TISEAN. Их трудно сравнивать между собой, поскольку у каждого из них своя специфика, свои достоинства и недостатки. Вкратце опишем их.

Іващенко А.Б.
Донецький національний технічний університет

Таблица 1. – Специализированное программное обеспечение для анализа временных рядов

№ п/п	Программный продукт	Компания-производитель, веб-страница	Страна	Цена / Демо-версия / Требования к уровню знаний пользователя	Возможности продукта
1	CATS (Centre for the Analysis of Time Series)	London School of Economics http://www2.lse.ac.uk/CATS/home.aspx	Англия	Платно, цены не указаны /Демо НЕТ/ Не известно	Опционно, оговариваются с заказчиком
2	SAS® Analytics (SAS® Forecasting and Econometrics software	SAS (Statistical Analysis System) http://www.sas.com/technologies/analytics/index.html	США	6000 $ / Демо НЕТ / НЕТ	Опционно, оговариваются с заказчиком
3	Autobox	Automatic forecasting systems, http://www.autobox.com/cms/index.php/home	США	8000-57 000 $ / Демо ЕСТЬ / НЕТ	Модели АРПСС (методика Бокса-Дженкинса) с доп. опциями (в зависимости от стоимости)
4	Аналитическая система для прогнозирования временных рядов	KAI Development, http://kaidev.ru/	Россия	Платно, цены не указаны / Демо НЕТ / НЕТ	Нейронные сети; вейвлеты; сглаживающие сплайны; автокорреляция и др.
5	Forecast Pro	Business Forecast Systems, http://www.forecastpro.com/	США	1000-22000 $ / Демо ЕСТЬ / НЕТ	Экспоненциальное сглаживание, модели Бокса-Дженкинса, подгонка кривых и т.п. (в зависимости от стоимости)
6	GRETL (GNU Regression, Econometrics and Time-series Library)	Allin Cottrell, Wake Forest University http://gretl.sourceforge.net/	США	Распространяется бесплатно // ДА	Статистический анализ, авторегрессионный анализ, фильтр Калмана и т.п.
7	Zaitun Time Series	http://zaitunsoftware.com/home	Индонезия	Распространяется бесплатно // НЕТ	Выделение тренда, скользящее среднее, экспоненциальное сглаживание, линейная регрессия, нейронные сети
8	TISEAN (Time Series Analysis)	R. Hegger, H. Kantz, Th.Schreiber http://www.mpipks-dresden.mpg.de/~tisean/	Германия	Распространяется бесплатно // ДА	Методы, основанные на теории хаоса и нелинейных динамических систем
9	ARfit (Multivariate Autoregressive Model Fitting)	Tapio Schneider, California Institute of Technology http://www.gps.caltech.edu/~tapio/arfit/	США	Распространяется бесплатно // НЕТ	Множественные авторегрессионые модели
10	Dataplore (Analysis of signals and time series data)	Ixellence GmbH http://www.ixellence.com/	Германия	1000-3000€ / Демо – НЕТ / ДА	Статистический анализ, регрессионный анализ, обработка сигналов, методы нелинейной динамики
11	RATS (Regression Analysis of Time Series)	Estima http://www.estima.com/ratsmain.shtml	США	300-650 $ / Демо НЕТ / ДА	Экспоненциальное сглаживание, авторегрессионный анализ, фильтр Калмана, спектральный анализ и т.п. (в зависимости от стоимости)
12	SHAZAM	www.econometrics.com/	Канада и Англия	250-2200$ / Демо ЕСТЬ / ДА	Авторегрессионный анализ, анализ главных компонент, факторный анализ и др. (в зависимости от стоимости)
13	АСТРИД (Автоматизированная СТРуктурная ИДентификация)	DITIM www.mgua.irtc.org.ua/ukr/index.php?page=astrid	Украина	Платно, цены не указаны/ Демо НЕТ / Не известно	Методы группового учета аргументов (МГУА и его модификации)
14	DataX (Inteligent Data Mining Software Suite)	Zaptron System www.zaptron.com/datax/index.htm	США	Платно, цены не указаны / Демо НЕТ / Не известно	Регрессионный анализ, МГУА, алгоритмы с нечеткой логикой, выделение тренда и циклической компоненты
15	GMDH Shell	Geos Research Group www.gmdhshell.com/	США	1000-5000 $ / Демо ЕСТЬ / ДА	Методы группового учета аргументов
16	CDA (Chaos Data Analyzer)	J. C. Sprott http://sprott.physics.wisc.edu/cda.htm	США	150 $ / Демо НЕТ / ДА	Некоторые методы теории хаоса (вычисляются максимальный показатель Ляпунова, корреляционная размерность, автокорреляционная функция); работает только в среде DOS
17	Fractan (FRACTal ANalysis)	http://www.iki.rssi.ru/magbase/RESULT/APPENDIX/fractan.boom.ru/soft.htm	Россия	Распространяется бесплатно // ДА	Некоторые методы теории хаоса (вычисляются корреляционная размерность, энтропия, показатель Херста, средняя взаимная информация)
18	DTREG (Decision Trees and REGressions)	Phillip H Sherrod http://www.dtreg.com/	США	1000-5000 $ / Демо ЕСТЬ/ ДА	Деревья решений, регрессионный анализ, нейронные сети, МГУА, генетические алгоритмы и пр. (в зависимости от стоимости)

Возможности некоторых перспективных программных решений в части анализа временных рядов

Пакет приложений SAS. SAS Systems – это очень известное мировое имя среди разработчиков программного обеспечения IT-решений, а также услуг для бизнеса. Стоимость их программных продуктов очень высока.

К преимуществам этого пакета можно отнести то, что он предназначен для пользователя, не имеющего специальной подготовки в области статистики и не программирующего на входном языке. В нем предоставляется не полный, но достаточный набор возможностей анализа данных: описательная статистика, расчеты обобщающих показателей, прогноз временных рядов, анализ «что-если» и прочее.

Особый интерес представляет приложение SAS Analyst. Это приложение, также ориентированное на пользователя без специальной статистической подготовки, позволяет быстро осуществить статистический анализ данных, табличное и графическое представление результатов. Также в SAS System разработано средство для интеллектуального анализа данных (SAS Data Mining Solution), дающее пользователю возможность осуществить весь цикл работы с исходными данными, имеющими большие объемы и невыясненную статистическую структуру.

Среди основных недостатков – отсутствие бесплатных демо-версий, высокая стоимость программных решений и ежегодная плата за лицензионный продукт.

Программный продукт GMDH Shell. Это программный инструмент для интеллектуального анализа данных и прогнозирования на основе алгоритма метода группового учета аргументов (МГУА, Group Method of Data Handling – GMDH), автором которого является украинский ученый Ивахненко А.Г. Сложность и специфика реализованных в продукте алгоритмов делают данный продукт в некотором смысле уникальным, а потому и ценным программным обеспечением для анализа данных.

Полезные функциональные возможности пакета GMDH Shell обеспечиваются модулями, описанными ниже:

- модуль предобработки включает в себя такие полезные функции, как, например, замена пропущенных значений на: ноль, среднее, центральное, наиболее встречаемое или интерполяция соседних значений; задание экзаменационной выборки; длина обучающего окна, интервал прогноза и др.;
- модуль «решателя» содержит, кроме всего прочего, такие полезные функции, как опция «дополнительные переменные» – формирование расширенного пространства переменных в полиномиальном базисе (формируются все возможные слагаемые полинома Колмогорова-Габора с учетом задаваемых ограничений); возможность опционного перемешивания наблюдений для множественной кросс-валидации; выбор алгоритма обучения: комбинаторный алгоритм с возможностью ограничения максимальной сложности моделей (COMBI) и многорядный итерационный алгоритм (улучшенный MIA);
- модуль пост-обработки включает функции усреднения прогнозов N лучших моделей, визуализации результатов, оценки качества прогнозирующих моделей (60 различных видов ошибки).

Недостаток – высокая стоимость продукта, программа требует высокой подготовленности пользователя и знания принципов методики.

Свободное программное обеспечение GRETL. GRETL (GNU Regression, Econometrics and Time-series Library – библиотека для регрессий, эконометрики и временных рядов) – прикладной программный пакет для эконометрического моделирования, часть проекта GNU.

Важным является то, что GRETL – это свободное бесплатное программное обеспечение с отрытым исходным кодом.

Наличие простого интуитивного интерфейса на различных языках мира, в том числе и на русском, делает этот пакет особенно привлекательным для пользователей.

Основные возможности:

- оценка параметров с помощью метода наименьших квадратов (OLS), метода максимального правдоподобия (ML), обобщённого метода моментов (GMM) и др.;
- выделение сезонности при помощи встраиваемых пакетов X-12-ARIMA и TRAMO/SEATS (Time series Regression with ARIMA noise, Missing values and Outliers / Signal Extraction in ARIMA Time Series);
- создание моделей временных рядов (авторегрессия скользящего среднего (ARMA), авторегрессия интегрированного скользящего среднего (ARIMA), обобщённая авторегрессия условной гетероскедастичности (GARCH), векторная авторегрессия (VAR), векторная модель коррекции ошибок (VECM) и др.);
- построение моделей с ограниченными зависимыми переменными: логит (logit), пробит (probit), тобит (tobit), интервальная регрессия и др.;
- скриптовый язык сценариев с поддержкой циклов для реализации метода Монте-Карло и итерационных процедур оценки.

Принципиальное преимущество: благодаря встроенному языку имеется возможность писать собственные программы.

GRETL интенсивно развивается благодаря бесплатности продукта и возможности пополнения библиотеки собственным кодом. В перспективе может

Іващенко А.Б.
Донецький національний технічний університет

составить достойную конкуренцию дорогостоящему программному обеспечению (типа Matlab, Statistica и т.д.).

Недостатки: незрелость программного продукта; требуется предварительное изучение команд языка.

Программное приложение DTREG. DTREG является мощным средством статистического анализа данных. Базовая (минимальная) комплектация пакета предназначена в основном для классификации и кластеризации данных и располагает простейшими регрессионными моделями и методами для построения деревьев решений. Также присутствует дискриминантный, корреляционный, факторный анализ, анализ главных компонент. Более дорогостоящие версии пакета дополнены возможностью использования генетических алгоритмов, исследования временных рядов с помощью авторегрессионных моделей, построения различных искусственных нейронных сетей (модель нейронной сети на основе множественного персептрона, сети с радиальными базисными функциями, полиномиальные сети на основе МГУА).

Ограниченная демонстрационная версия DTREG предназначена для ознакомления лишь с базовой комплектацией пакета и не позволяет оценить возможности дополнительных модулей программы. Главным недостатком DTREG является высокая стоимость продукта.

Программный проект TISEAN. Пакет программного обеспечения TISEAN – это библиотека отдельных алгоритмов, реализующих методы, основанные на парадигме детерминированного хаоса. Включает алгоритмы для представления и визуализации данных, понижения шума, прогнозирования, оценки показателя Ляпунова и корреляционной размерности. Каждый алгоритм реализован в виде подпрограммы, хранящейся в отдельном файле и вызываемой с помощью командной строки. Для всех алгоритмов предусмотрено использование опционных «ключей» для настройки параметров команды.

Данный пакет фактически является единственным продуктом, наиболее полно реализующим подходы теории нелинейных динамических систем. Содержит мощные инструменты для анализа временных рядов. И к тому же находится в свободном пользовании.

Недостаток: требуется хорошая теоретическая подготовка пользователя; отсутствие интерфейса; работа с командной строкой.

Заключение

На первый взгляд, рынок программных решений для анализа временных рядов не выглядит несостоятельным, но при более детальном его анализе становится понятно, что доступного и качественного программного обеспечения для анализа временных рядов однозначно недостаточно, особенно касательно программ, реализующих интеллектуальные методы или методы нелинейной динамики.

Перспективные качественные бесплатные программные продукты требуют от пользователя профессиональных навыков и высокой квалификации, предполагают наличие статистического образования и знания английского языка, поскольку требуют внимательного изучения документации на английском языке.

Крупные профессиональные приложения требуют очень больших финансовых затрат, так как немаловажное значение имеет цена пакета. Профессиональные статистические или математические пакеты (SAS, Statistica, Matlab и т.д.) обычно стоят от 1 до 10 тыс. долларов.

Среди универсальных математических программных средств лидерами являются пакеты GAUSS, Matlab и Mathematica. Среди бесплатных программ в качестве перспективных можно назвать пакеты GRETL и TISEAN.

Литература

1. Comparison of statistical packages // The free encyclopedia «Wikipedia» – http://en.wikipedia.org/wiki/Comparison_of_statistical_packages (21.11.2012).
2. Brendan O'Connor. Comparison of data analysis packages: R, Matlab, SciPy, Excel, SAS, SPSS, Stata – http://brenocon.com/blog/2009/02/comparison-of-data-analysis-packages-r-matlab-scipy-excel-sas-spss-stata/ (21.11.2012).
3. И.А. Чучуева. Сравнение программных продуктов для анализа данных: R, Matlab, SciPy, MS Excel, SAS, SPSS, Stata – http://www.mbureau.ru/blog/sravnenie-programmnyh-produktov-dlya-analiza-dannyh-r-matlab-scipy-ms-excel-sas-spss-stata (21.11.2012).
4. Ю.Н. Тюрин, А.А. Макаров. Статистический анализ данных на компьютере / Под ред. В.Э. Фигурнова – М.: ИНФРА, 1998. – 528 с.
5. S. Steinhaus. Comparison of mathematical programs for data analysis – http://www.scientificweb.de/ncrunch/ (21.11.2012).

Іващенко А.Б.
Донецький національний технічний університет