Использование программных средств статистической обработки данных при формировании информационного обеспечения управления

Елизарова Н.Н., канд. техн. наук

Рассматриваются статистические пакеты обработки данных STATISTICA, SPSS, Deductor, а также возможности MS Excel, которые могут быть использованы для задач принятия решений планирования и управления различных объектов.

Ключевые слова: программные средства, пакеты прикладных программ, статистическая обработка данных, информационное обеспечение, описательные методы, дисперсионный анализ, регрессионный анализ, корреляция.

Use of software programs statistical data processing when shaping the dataware of management

The statistical packages of data processing STATISTICA, SPSS, Deductor are considered, and also the opportunities MS Excel, which can be used for tasks of acceptance of the decisions of planning and management of various objects.

Keywords: software programs, packages of the applied programs, statistical data processing, dataware, descriptive methods, analysis of variance, regressional analysis, korrelyaciya.

В современных условиях наблюдается постоянный рост интенсивности информационных потоков и объемов обрабатываемой информации. Это требует непрерывного обновления знаний о состоянии предметной области и перспектив развития. При решении задач планирования и выбора стратегии развития предприятия, фирмы можно выделить ряд блоков задач, в решении которых используется статистическая информация:

– формирование стратегических целевых установок фирмы;

– прогнозирование потребности в материальных, энергетических, трудовых и финансовых ресурсах;

– анализ конкурентов и рынков сбыта;

– анализ спроса и предложений;

– оценка финансовой деятельности предприятия;

и другие.

Особенность решения таких задач заключается не только в обработке большого объема информации, но и в необходимости выявления причинно-следственных связей, построении формализованных моделей для анализа и прогноза. К основным предпосылкам применения современных информационных технологий в области статистической обработки информации можно отнести следующие:

– большое количество объектов статистического наблюдения, многомерность данных;

– необходимость отслеживания динамики массива показателей во времени, формирование на их основе различных сводок;

– низкую оперативность обработки данных;

– высокие материальные и трудовые затраты на сбор и обработку статистической информации;

– территориальную разобщенность исходных данных, необходимость их интеграции и одновременной обработки;

– сложность математических методов анализа данных.

В последнее время получили широкое распространение программные средства или информационные системы, предназначенные для автоматизации работ статистической обработки данных, которые позволяют собирать, хранить и обрабатывать разнородные массивы данных с использованием единой информационной базы. Такие системы на предприятии ориентируются на потребности руководства при выполнении функций управления на основе внутренних и внешних статистических данных. Достоинством таких систем является адаптация информационной базы и функций системы к условиям функционирования предприятия. Однако, в силу сложности реализации математических методов, такие системы, как правило, включают лишь ограниченный набор аналитических методов.

В настоящее время получили распространение статистические пакеты, которые могут быть легко подключены к существующей информационной системе обработки информации на предприятии. В нашей стране наибольшее распространение получили следующие статистические пакеты:

– STATISTICA;

– SPSS;

– Deductor.

Рассмотрим их подробнее.

Пакет прикладных программ STATISTICA /1-3/ – универсальная система анализа данных, разработанная компанией StatSoft, построенная по модульному принципу, каждый модуль выполняет определенный набор функций и может быть использован и автономно. Основные возможности пакета:

– реализует широкий набор математических методов (табл. 1);

– дает возможность представить графическую интерпретацию результатов (в графиках типа 2D, 3D, пиктограммах или в разработанных в собственном дизайне графиках);

– осуществляет поддержку всех стандартов современных офисных приложений (импорт данных из электронных таблиц, в том числе и их MS Excel, экспорт диаграмм в приложения MS Office и др.);

– позволяет расширять возможности пакета за счет встроенного языка программирования Statistica Visual Basic.

Пакет STATISTICA может применяться в разнообразных сферах деятельности:

– в банковской деятельности (для анализа кредитных рисков и прогнозирования финансовых показателей);

– торговой деятельности (для сравнительного анализа поставщиков и прогнозирования потребления товаров и ресурсов);

– маркетинговых исследованиях (для изучения сезонности спроса, классификации товара по потребительским свойствам);

– производственной деятельности (для прогнозирования потребности материальных ресурсов, выявления причинно-следственных связей между технологическими параметрами, анализа надежности и долговечности продукции);

– медицинском обслуживании (для анализа результатов обследования, диагностики);

– социологических исследованиях (для анализа опроса общественного мнения).

Кроме этого, пакет STATISTICA является базовым статистическим пакетом в большинстве вузов России, служит для обучения методам статистического анализа.

Пакет прикладных программ SPSS (Statistical Package for Social Science) /4-6/ – статистический пакет, разработанный компанией SPSS Inc, предназначенный для работы в операционной системе MS Windows. Является пакетом обработки и анализа социологических данных. Основные возможности пакета:

– реализует набор математических методов статистической обработки данных (табл. 1);

– осуществляет доступ к территориально распределенным данным и позволяет объединять несколько баз данных;

– формирует нестандартные отчеты, позволяющие оценить данные с разных точек зрения;

– осуществляет настройку интерфейса и процедур работы с данными с помощью встроенного языка сценариев;

– поддерживает связь с большинством форматов данных и обмен данными с другими приложениями MS Windows.

Пакет прикладных программ Deductor /7-8/ – статистический пакет, разработанный фирмой Base Group Labs, состоит из 3-х частей: многомерного хранилища данных Deductor Warehouse, аналитического приложения Deductor Studio и рабочего места конечного пользователя Deductor Viewer.

Deductor Warehouse – многомерное хранилище данных, аккумулирующее всю необходимую для анализа предметной области информацию.

Deductor Studio – программа, реализующая функции импорта, обработки, визуализации и экспорта данных. В Deductor Studio включен полный набор механизмов, позволяющий получить информацию из произвольного источника данных, провести весь цикл обработки, используя **Мастера обработки** (очистку, трансформацию данных, построение моделей), отобразить полученные результаты наиболее удобным образом (OLAP, диаграммы, деревья…) и экспортировать результаты на сторону. Это полностью соответствует концепции извлечения знаний из баз данных.

Deductor Viewer – рабочее место конечного пользователя. Позволяет отделить процесс построения моделей от использования уже готовых моделей. Все сложные операции по подготовке моделей выполняются аналитиками-экспертами при помощи Deductor Studio, а Deductor Viewer обеспечивает пользователям простой способ работы с готовыми результатами.

Реализованные в Deductor обработчики покрывают основную потребность в анализе данных и создании законченных аналитических решений на базе Data Mining.

Кроме описанных трех статистических пакетов, для сравнения рассмотрим пакет MS Excel.

Анализ возможностей различных пакетов (табл. 1) позволил сформулировать их преимущества и недостатки и дать рекомендации по их применению:

1. Хотя пакет MS Excel не является статистическим пакетом, но он входит в MS Office, включает много статистических функций и дает возможность подключить встроенный пакет **Анализа данных** /9-10/. Поэтому следует рассмотреть его возможности для статистического анализа. Для небольших предприятий, когда не требуется проводить кластеризации данных, а лишь необходимо установить некоторые зависимости, дать статистическое описание исследуемым переменным, данный пакет будет экономически выгодным.

2. Пакет STATISTICA является мощным средством статистического анализа, нашедший применение во многих сферах деятельности. Он включает большое количество методов, реализуемых в отдельных модулях, которые могут запускаться автономно. Но для реализации каждого метода не хватает методики их выполнения и толкований полученных результатов. Этот недостаток может затруднить внедрение пакета.

3. Пакет ППП SPSS включает широкий спектр команд и процедур, связанных с описательными методами статистики: описание распределения, анализ связи количественных и качественных переменных, наряду с параметрическими методами сравнения средних, большой набор непараметрических тестов. Такая обработка актуальна в ходе социологических исследованиях. Имеется возможность работать с данными, подготовленными в MS Excel.

4. Пакет Deductor имеет единое хранилище данных (а не отдельные файлы, как ППП STATISTICA), разработанные сценарии, включающие загрузку данных из хранилища или внешнего источника, восстановление пропущенных значений, установления незначимых факторов, построение моделей. В пакете при открытии файла с данными он проверяется на пропущенные данные, идет их восстановление, поэтому результаты дальнейшей обработки могут немного отличаться от других пакетов.

Таблица 1. **Сопоставление возможностей статистических пакетов**

Функции и методы	Пакеты прикладных программ			
	MS Excel	**STATISTICA**	**SPSS**	**Deductor**
Описательные методы статистического анализа:				
1) вычисления математических ожиданий, дисперсий изучаемых величин и др.	встроенные функции Excel	модуль **Описательной статистики**	команда **Descriptives**	при выполнении функции **Линейная регрессия**
2) проверка гипотез о равенстве математических ожиданий	функции пакета **Анализа данных**	модуль **Описательной статистики**	широкий спектр команд **One sample T-test**, **Independent sample T-test** и др. непараметрические методы	–
3) построение гистограмм	функции пакета **Анализа данных**	модуль **Описательной статистики**	команды **FREQUENCIES STATISTICS, HISTOGRAM**	–
Построение модели временного ряда и прогнозирование с учетом сезонных колебаний и периодических трендов	требуется самостоятельно создавать шаблон на листе Excel	модуль **Временные ряды и прогнозирование** с поквартальной и месячной десонализацией	–	–
Построение многомерной линейной регрессионной модели	встроенная функция **ЛИНЕЙН** и функция пакета **Анализ данных РЕГРЕССИЯ**	модуль **Множественная регрессия**	линейная регрессия в процедуре **REGRESSION**	функция **Линейная регрессия**
Построение нелинейной регрессионной модели	встроенные функции позволяют построить полиномиальную и экспоненциальную модели	модуль **Множественная регрессия** дает большой выбор нелинейных моделей	логистическая регрессия в процедуре **REGRESSION**	–
Корреляционный анализ	встроенные функции Excel **КОРРЕЛ**, **КОВАР**, функции пакета **Анализа данных**	модули **Описательной статистики, Непараметрический анализ.**	процедуры связи количественных переменных **CORRELATIONS** и неколичественных переменных **CROSSTABS**	функция **Корреляционный анализ**
Одномерный и двухмерный дисперсионный анализ	функции пакета **Анализ данных**	модуль **Дисперсионный анализ**	процедура **ANOVA**	–
Кластерный анализ	–	модуль **Кластерный анализ**	процедуры **CLUSTER, QUICK CLUSTER** или команда **k-means**.	функции **Дерево решений и Карта Кохонена**
Факторный анализ	–	модуль **Факторный анализ**	процедура **FACTOR**	функция **Факторный анализ**
Дискриминантный анализ	–	модуль **Дискриминантный функциональный анализ**	–	–
Многомерное шкалирование	–	модуль **Многомерное шкалирование**	процедура **Multidimentional scaling**	–
Возможности графического отображения результатов	встроенные функции **Мастер диаграмм**	графики типа 2M, 3M, пиктограммы	графики, дендрограммы в процедуре **PLOT DEND-ROGRAM**	диаграммы, гистограммы, **OLAP** – многомерное представление данных в виде кросс-таблиц и кросс-диаграмм
Возможности импорта данных	из других приложений MS Office	из других приложений MS Office, в том числе из MS Excel	из других приложений MS Office, в том числе из MS Excel	из других приложений MS Office программой **Deductor Studio**
Возможности экспорта данных	таблицы и диаграммы в другие приложения MS Office	таблицы и диаграммы в другие приложения MS Office	таблицы и диаграммы в другие приложения MS Office	таблицы и диаграммы в другие приложения MS Office программой **Deductor Studio**
Возможности интеллектуализации данных	–	дополнительный модуль **Нейронные сети**	–	методы **Мастера обработки: Нейросеть**
Очистка и трансформация данных	–	модуль **Временные ряды и прогнозирование**	–	широкий спектр, в том числе: сглаживание (скользящее окно), очистка от шумов (фильтрация), группировка

Пример вычисления коэффициентов корреляции в статистических пакетах

1. Корреляционный анализ в MS Excel (рис. 1, 2).

Таблица 2. **Вычисление парных коэффициентов корреляции с применением встроенной функции Excel КОРРЕЛ для выходной переменной *Доход***

Основные средства	Нематериальные активы	Запасы	Отложенные налоговые активы	Расходы	Уплаченные налоги	Максимальная дополнительная прибыль от проверки
Столбец 1	*Столбец 2*	*Столбец 3*	*Столбец 4*	*Столбец 5*	*Столбец 6*	*Столбец 7*
-0,106	-0,100	-0,106	0,272	-0,130	0,058	0,012

Таблица 3. **Вычисление корреляционной матрицы с применением функции пакета Анализа данных в Excel**

	Основные средства	Нематериальные активы	Запасы	Отложенные налоговые активы	Расходы	Уплаченные налоги	Максимальная дополнительная прибыль от проверки	*Доход*
	Столбец 1	*Столбец 2*	*Столбец 3*	*Столбец 4*	*Столбец 5*	*Столбец 6*	*Столбец 7*	*Столбец 8*
Столбец 1	1,000							
Столбец 2	0,994	1,000						
Столбец 3	1,000	0,994	1,000					
Столбец 4	-0,095	-0,094	-0,095	1,000				
Столбец 5	0,084	0,066	0,084	-0,002	1,000			
Столбец 6	-0,451	-0,462	-0,451	-0,115	-0,141	1,000		
Столбец 7	0,507	0,508	0,507	0,079	0,096	-0,798	1,000	
Столбец 8	-0,106	-0,100	-0,106	0,272	-0,130	0,058	0,012	1,000

2. Вычисление парных коэффициентов корреляции для выходной переменной *Доход* (табл. 2, 3).

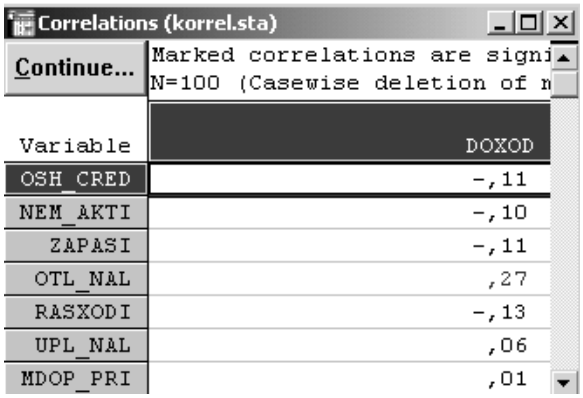

Рис. 1. STATISTICA

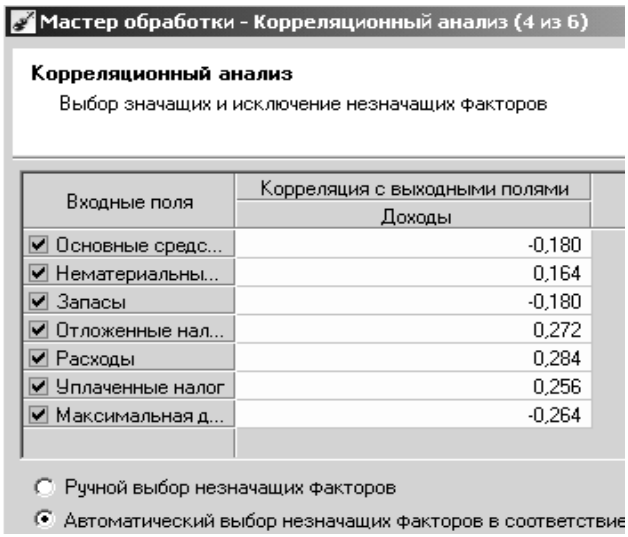

Рис. 2. Deductor

3. Вычисление корреляционной матрицы (рис. 3, 4)

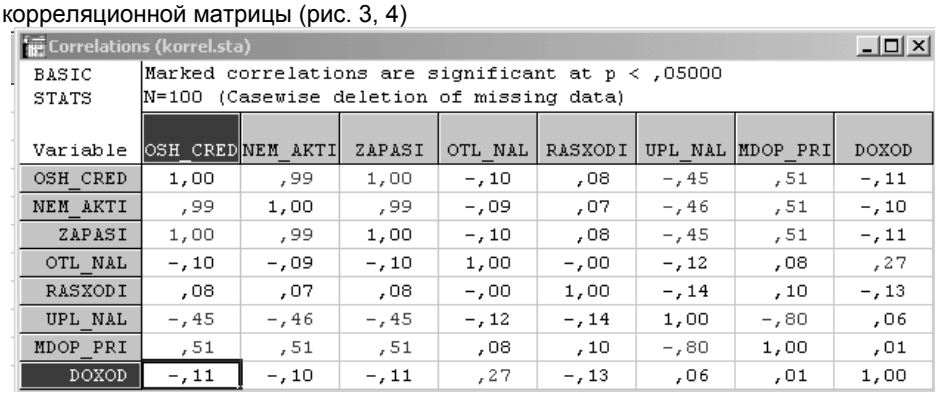

Рис. 3. ППП STATISTICA

Correlations

		Основные средства	Нематериальные активы	Запасы	Отложенные налоговые активы	Расходы	Уплаченные налоги	Максимальная дополнительная прибыль от проверки	Доходы
Основные_средства	Pearson Correlation	1	,994**	1,000**	-,095	,084	-,451**	,507**	-,106
	Sig. (2-tailed)		,000	,000	,346	,404	,000	,000	,295
	N	100	100	100	100	100	100	100	100
Нематериальные_активы	Pearson Correlation	,994**	1	,994**	-,094	,066	-,462**	,508**	-,100
	Sig. (2-tailed)	,000		,000	,354	,516	,000	,000	,324
	N	100	100	100	100	100	100	100	100
Запасы	Pearson Correlation	1,000**	,994**	1	-,095	,084	-,451**	,507**	-,106
	Sig. (2-tailed)	,000	,000		,346	,404	,000	,000	,295
	N	100	100	100	100	100	100	100	100
Отложенные_налоговые_активы	Pearson Correlation	-,095	-,094	-,095	1	-,002	-,115	,079	,272**
	Sig. (2-tailed)	,346	,354	,346		,985	,253	,435	,006
	N	100	100	100	100	100	100	100	100
Расходы	Pearson Correlation	,084	,066	,084	-,002	1	-,141	,096	-,130
	Sig. (2-tailed)	,404	,516	,404	,985		,162	,343	,198
	N	100	100	100	100	100	100	100	100
Уплаченные_налог	Pearson Correlation	-,451**	-,462**	-,451**	-,115	-,141	1	-,798**	,058
	Sig. (2-tailed)	,000	,000	,000	,253	,162		,000	,565
	N	100	100	100	100	100	100	100	100
Максимальная_дополнительная_прибыль_от_проверки	Pearson Correlation	,507**	,508**	,507**	,079	,096	-,798**	1	,012
	Sig. (2-tailed)	,000	,000	,000	,435	,343	,000		,905
	N	100	100	100	100	100	100	100	100
Доходы	Pearson Correlation	-,106	-,100	-,106	,272**	-,130	,058	,012	1
	Sig. (2-tailed)	,295	,324	,295	,006	,198	,565	,905	
	N	100	100	100	100	100	100	100	100

**. Correlation is significant at the 0.01 level (2-tailed).

Рис. 4. ППП SPSS

Список литературы

1. **Боровиков В.П., Ивченко Г.И.** Прогнозирование в системе STATISTICA в среде Windows. Основы теории и интенсивная практика на компьютере: Учеб. пособие. 2-е изд., перераб. и доп. – М.: Финансы и статистика, 2006.
2. **StatSoft, Inc.** (2001). Электронный учебник по статистике. Москва, StatSoft. WEB: http://www.statsoft.ru/home/textbook/default.htm.
3. **Белов А.А., Баллод Б.А., Елизарова Н.Н.** Теория вероятностей и математическая статистика: Учебник / ГОУВПО «Ивановский государственный энергетический университет имени В.И. Ленина». – Иваново, 2006.
4. **Божко В.П., Хорошилов А.В.** Информационные технологии в статистике. – М.: Финстатинформ, 2002.
5. **Румянцева Е.Л.** Информационные технологии: Учеб. пособие / Под ред. прф. Л.Г. Гагариной. – М.: ИД «ФОРУМ»: ИНФРА-М, 2007.
6. **SPSS для Windows**. Руководство пользователя SPSS. Кн. 1. – М.: Статистические системы и сервис, 1998
7. **Интеллектуальные** модели анализа экономической информации: Электронный курс лекций. – BaseGroup Labs, 2005.
8. **Аналитическая** платформа Deductor 4. Руководство пользователя. – BaseGroup Labs, 2005.
9. **Минько А.А.** Принятие решений с помощью Excel. Просто как дважды два. – М.: Эксмо, 2006.
10. **Поиск** оптимальных решений средствами Excel 7.0. – СПб.: BHV – Санкт-Петербург, 1997.

Елизарова Надежда Николаевна,
Ивановский государственный энергетический университет,
кандидат технических наук, доцент кафедры информационных технологий,
e-mail: elisarova@it.ispu.ru