О ПРОГРАММНОМ ОБЕСПЕЧЕНИИ АНАЛИЗА ТАБЛИЦ ЭМПИРИЧЕСКИХ ДАННЫХ

К.А. Рыбаков, Е.В. Черепанов

Московский авиационный институт (государственный технический университет)
Институт экономики и комплексных проблем связи

Введение. В статье приводится краткое описание программного обеспечения «Анализ таблиц эмпирических данных», позволяющего решать ряд актуальных задач обработки информации: выявление недостоверных данных, восстановление пропущенных значений в таблицах, прогнозирование и планирование временных рядов, классификация данных и анализ уровня.

1. Проверка достоверности данных. В эконометрических, социально- и технико-экономических исследованиях, как правило, базой для анализа служат таблицы эмпирических данных (таблицы «объект–свойство»). Они часто оказываются неполными (содержат пропуски значений показателей для некоторых наблюдений) и обладают заметной недостоверностью (часть данных неточна – случайные ошибки, ложные сведения). В этой связи проблема выявления недостающей и недостоверной информации в таблицах эмпирических данных может считаться неотъемлемой частью первичной статистической обработки практически во всех прикладных статистических работах [1].

В разработанном программном обеспечении использовались новые алгоритмы анализа эмпирических таблиц с целью выявления недостающей и ложной информации. По сути, применяемая методика опирается на то, что, во-первых, числовые показатели, как правило, коррелированны, и, во-вторых, наблюдения в таблице обладают мерами подобия, что также поддается формализации [2–4].

Отечественный и зарубежный опыт свидетельствует о том, что решение задачи выявления недостающей и недостоверной информации в таблицах данных, основанное на прямом использовании числовой информации таблицы, может давать грубые и неточные результаты. Связано это с тем, что классические оценки обладают весьма низкой робастностью и в такой ситуации требуется использование более тонких алгоритмов анализа данных. В программе используются различные подходы, в частности, предварительный анализ и отбраковка сомнительных данных («выбросов»), переход к ранговым переменным, использование робастных оценок и др. Отметим, что применение ранговых статистик является одним из наиболее радикальных методов непараметрической статистики, дающих стабильность и относительно высокую эффективность. Кроме того, могут использоваться полиграммные методы оценивания, для которых в ряде социально-экономических исследований установлена высокая стабильность и хорошая точность [5].

В качестве примера проведен анализ достоверности некоторых показателей социально-экономического положения субъектов Российской Федерации, входящих в Северо-западный федеральный округ (данные 2008 г., рис. 1). В исходной таблице среднемесячная начисленная заработная плата в республике Коми введена ошибочно (исходное значение – 20638.3, введенное – 206383, восстановленное – 20322.9, относительная погрешность оценивания – 1.53%), указанный объем иностранных инвестиций в Вологодской области признан недостоверным. Данные по строительству жилых домов в Псковской области и числу безработных в Ленинградской области были помечены как отсутствующие и в результате довольно точно восстановлены: для объема

строительства исходное значение – 0.214, восстановленное – 0.262, относительная погрешность оценивания – 22.43%; для относительного числа безработных – 0.034, 0.032, 5.88% соответственно.

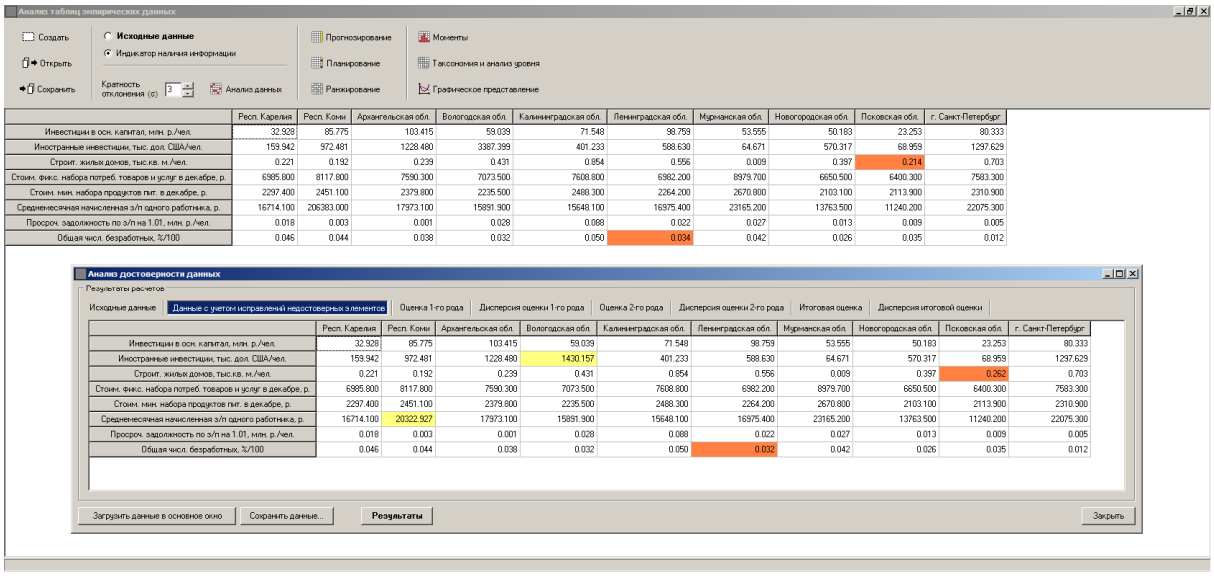

Рис. 1. Апробация алгоритмов проверки достоверности данных

2. Прогнозирование и планирование. Во многих случаях для обоснования динамики изменения показателей деятельности предприятия имеется очень короткий ряд статистических данных (5–8 наблюдений). В программе реализована методика, ориентированная на прогнозирование очень коротких ретроспективных рядов (менее 10 наблюдений). В таких задачах традиционные методы прогнозирования (на основе регрессионных и спектральных методов) не могут дать какой-либо надежный результат. В этой связи перспективным является непараметрическое прогнозирование последовательностей показателей, не предусматривающее сглаживание данных какой-либо кривой. Разработанный подход основан на представлении (аналитически неизвестного) тренда взвешенной суммой ретроспективных наблюдений. Идея метода впервые была предложена в работе [6] и развита в [7]. Другой задачей, которая может быть решена с помощью разработанного программного обеспечения, является задача планирования, отличающаяся от прогнозирования тем, что часть показателей должна иметь заданные значения, т. е. динамика этих показателей либо априори такова, задана экспертом, либо нужно обеспечить эту динамику.

На рис. 2 представлен пример расчетов: прогнозирование курсов доллара США, евро, австралийского доллара, фунта стерлингов и китайского юаня при запланированных курсе японской иены и цене на нефть (данные – недельные котировки за несколько месяцев 2009 г.; относительная погрешность прогноза не превосходит 2.63%).

3. Классификация и анализ уровня. Важными составляющими программного обеспечения «Анализ таблиц эмпирических данных» являются возможности классификации и анализа уровня. Вообще, любая классификация является снижением размерности пространства наблюдений. Ценность процедур такого рода связана с тем, что при экспертном анализе объектов изучаемой области гораздо удобнее работать не с большим массивом исходных наблюдений, а с небольшим числом некоторых классов («обобщенных наблюдений») [8].

В социально-экономических и социологических исследованиях для процедур классификации наблюдений часто используется термин «типологизация» [9]. Заметим, что термины «таксономия», «кластеризация», «типологизация» разными авторами применяются в различных смыслах классификации, причем общепринятой трактовки различий в использовании указанных терминов нет. Впрочем, кластеризация обычно подразумевает некоторую процедуру разбиения множества наблюдений на классы, число которых заранее (до проведения формальной процедуры систематизации) не задано. В этом смысле реализованный алгоритм относится к кластерному анализу [10]. Отметим, что в результате автоматизированной классификации эмпирических данных крайне редко можно получить в каком-то смысле «фундаментальную» систематизацию объектов изучаемой области. В этой связи содержательная трактовка полученных результатов часто носит нетривиальный характер.

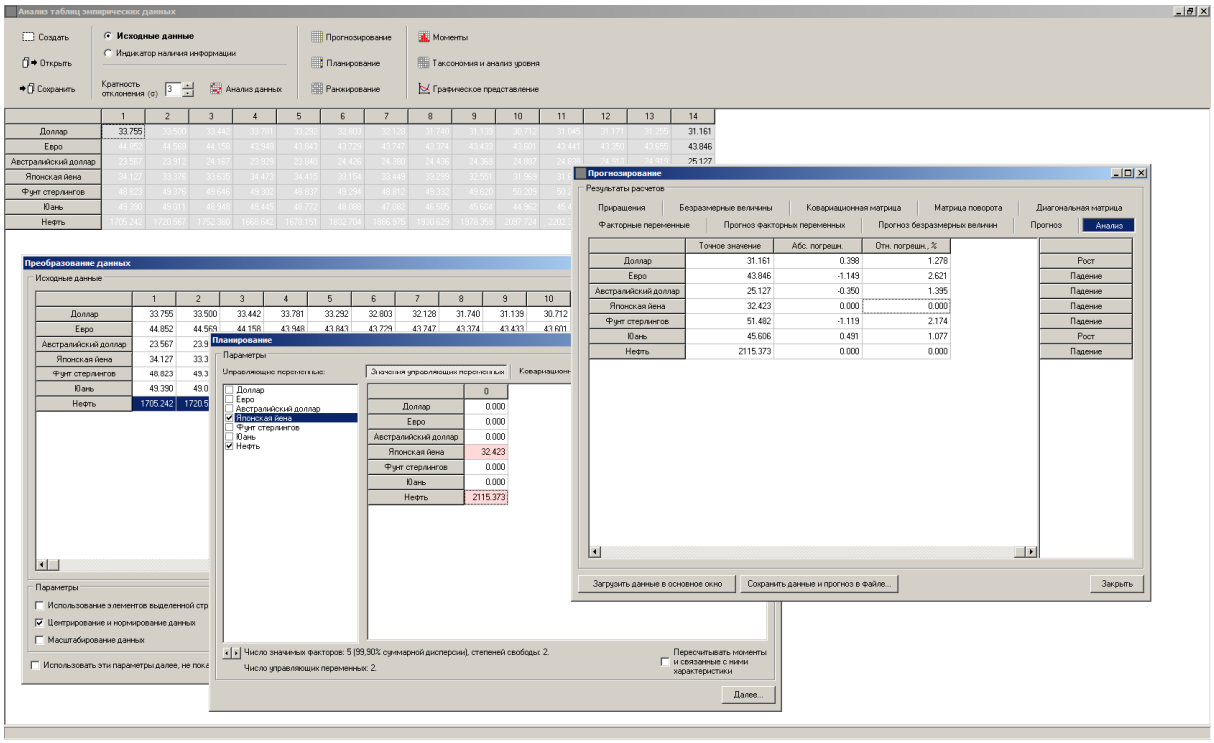

Рис. 2. Апробация алгоритмов прогнозирования и планирования

Идея реализованного метода возникла в связи с решением задач сопоставительного анализа технико-экономических объектов [11]. В начале метод применялся при решении ряда задач технико-экономических исследований [12, 13], а затем был развит для решения задач социальной классификации (с использованием слабых шкал), прикладной социологии и маркетинга потребительских рынков [14, 15].

Наряду с традиционными методами классификации в приложениях большую роль играют задачи систематизации объектов по их уровню. Корректное сопоставление объектов требует использования многокритериального выбора, а применение каких-либо методов квалиметрии очень ненадежно в силу чрезвычайно высокой чувствительности оценочной функции к вариациям параметров модели. Здесь использован подход к анализу уровня объектов, впервые предложенный в [16] и развитый в [11, 13].

Для апробации проанализированы данные о характеристиках 88 самолетов по одиннадцати показателям, результаты приведены на рис. 3. Классификация зависит от числа значимых факторов. В данном случае выбрано четыре фактора; предварительно в исходной таблице были восстановлены пропущенные значения (27 пустых ячеек).

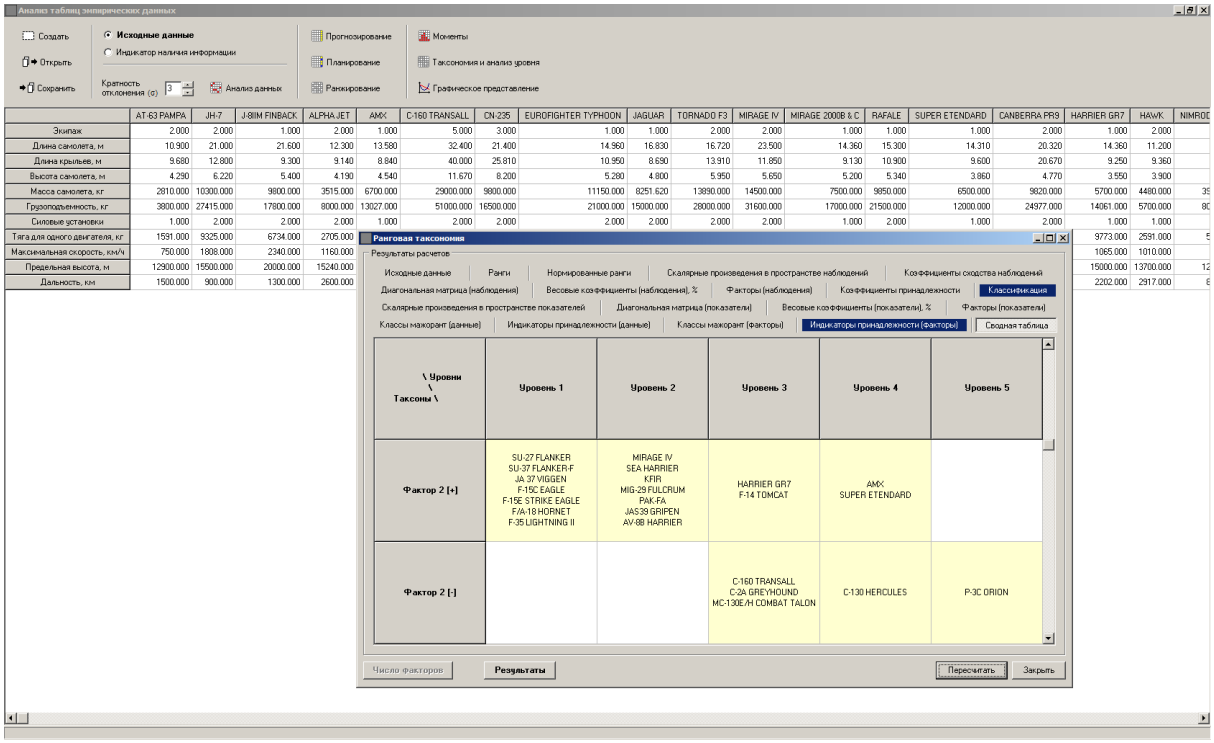

Рис. 3. Апробация алгоритмов классификации данных и анализа уровня (фрагмент таблицы, Фактор 2 характеризует максимальную скорость и предельную высоту).

Заключение. В работе описано программное обеспечение «Анализ таблиц эмпирических данных», перечислены его основные возможности. Приведены иллюстрирующие примеры применения алгоритмов проверки достоверности данных, прогнозирования и планирования, классификации данных и анализа уровня.

Библиографический список

1. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. – Новосибирск: Изд-во Института математики СО РАН, 1999.
2. Сюнтюренко О.В., Черепанов Е.В. Информатика: анализ данных и эконометрия // Средства связи. – 1986. № 4. – С. 36–44.
3. Черепанов Е.В. Анализ полноты и достоверности информации в таблицах эмпирических данных // Анализ социально-экономических и политических процессов и систем. Вып. 4: Сб. науч. работ. – М.: АМИ, 2007. – С. 147–153.
4. Рыбаков К.А., Черепанов Е.В. Анализ данных в эмпирических таблицах с использованием порядковых статистик // Информатика, социология, экономика, менеджмент. Вып. 7, ч. 2: Межвуз. сб. науч. тр. – М.: АМИ, 2010. – С. 60–65.

5. Тарасенко Ф.П., Черепанов Е.В. Полиграммные оценки линейных функционалов // Математическая статистика и ее приложения. Вып. X. – Томск: Изд-во ТГУ, 1986. – С. 204–211.
6. Черепанов Е.В. Статистическое прогнозирование экономической динамики в терминах конечных разностей // IV Сибирская научно-практическая конференция по надежности научно-технических прогнозов: Тез. докл. – Новосибирск: ВСНТО, 1987. – С. 182–185.
7. Жеруль А.О., Черепанов Е.В. Экономическое прогнозирование на основе непараметрического экстраполирования коротких временных рядов // Анализ социально-экономических и политических процессов и систем. Вып. 3. Математические вопросы социально-экономических исследований: Сб. науч. работ. – М.: АМИ, 2006. – С. 28–35.
8. Айвазян С.А. и др. Прикладная статистика. Классификация и снижение размерности. – М.: Финансы и статистика, 1989.
9. Татарова Г.Г. Основы типологического анализа в социологических исследованиях. – М.: Высшее образование и наука, 2007.
10. Дюран Б., Одел П. Кластерный анализ. – М.: Статистика, 1977.
11. Черепанов Е.В. и др. О комплексе статистических методов и моделей для анализа информационных данных // Научно-технический прогресс и информация. – 1980. № 4. – С. 55–61.
12. Черепанов Е.В. и др. Обработка фактографической технико-экономической информации статистическими методами // Вопросы радиоэлектроники. Сер. АСУПР. – 1985. № 1. – С. 60–66.
13. Черепанов Е.В., Щиренко Е.Г. Применение методов многомерного анализа данных при проведении технико-экономических исследований // Техника средств связи. Сер. Техника, экономика, управление. Вып. 3. – 1985. – С. 25–30.
14. Азаров С.В., Зотова Е.А., Черепанов Е.В. Кластеризация многомерных наблюдений на основе компонентного анализа статистик бинарного отношения на множествах // Математические методы и компьютерные технологии в маркетинговых и социальных исследованиях: Сб. науч. работ. – М.: АМИ, 2004. – С. 79–86.
15. Черепанов Е.В. Понятие типологического пространства в задачах многомерной систематизации // Информатика, социология, экономика, менеджмент. Вып. 7, ч. 2: Межвуз. сб. науч. тр. – М.: АМИ, 2010. – С. 107–115.
16. Абросов В.И., Черепанов Е.В. О выявлении современного уровня техники рассматриваемой области // Бюллетень межотраслевой информационной службы. – М.: ВНИИ межотраслевой информации, 1979, № 1. – С. 41–48.