GENERATION OF SMALL ENTERPRISES' MULTIVARIATE SAMPLE AGGREGATES

N.Ch. Bokun
Statistics Research Institute
Minsk, BELARUS

Abstract

The work gives methodology of small enterprises' multivariate sample aggregates in order to provide more comprehensive statistical evaluation of small institutional units' economic activity. The strategy is found on usage cluster analysis. The article provides results of small enterprises' selection carried out in Minsk City.

Согласно расширительной концепции экономическое производство охватывает весь объем производства товаров, услуг для продажи и собственного потребления. Для наиболее полного измерения деятельности в рамках границ сферы производства СНС, начиная с 1993 г., рекомендуется оценивать размеры теневой экономики, в составе которой наряду со скрытой, нелегальной, неформальной (сектор домашних хозяйств) компонентами, существенное место занимает деятельность, не учтенная вследствие недостатков методологии статистического наблюдения. В первую очередь это относится к малым предприятиям: в последние годы в республике их число превысило 33 тысячи, от каждого из них нецелесообразно требовать предоставления государственной отчетности. Полный охват совокупности сверхмалых и малых предприятий становится экономически неоправданным и практически нереализуемым. В таких условиях единственным достоверным методом оценки представляется выборочное обследование. Но несмотря на явные преимущества выборки: относительно небольшие материальные, временные и стоимостные затраты, оперативность получения статданных, достаточно высокая достоверность, - в ходе проведения выборочных наблюдений возникает ряд проблем: неответы респондентов; наличие в совокупности нетипичных единиц, выборки малого объема, дробление выборок на мелкие группы, неадекватная экстраполяция данных и т.д.

Если две первые проблемы: "неответы респондентов" и "нетипичные единицы совокупности", - могут быть решены в рамках одномерной выборки, то для позитивного разрешения остальных требуется построение оптимальных многомерных выборок. Только многомерный отбор позволяет получить репрезентативные выборки небольшого объема, характеризующие объект наблюдения по большому количеству сильно варьируемых показателей и позволяющие адекватно экстраполировать выборочные данные на генеральную совокупность. Кроме того, при использовании многомерной выборки предприятие рассматривается не как абстрактная единица, описываемая одним или двумя признаками, а как реально существующий объект со своими отличительными свойствами и многообразием связей.

Существующие подходы к построению многомерной выборки основаны либо на расслоении в независимых признаках, предполагающем использование сложной системы методов типизации, оптимизации, комбинаторного анализа и порождающем большое количество конечных слоев слабой заполненности (отбор из типизированных основ выборки, отбор из множественной основы выборки, отбор по решетке), либо на отборе по многомерному нормированному (композитному) показателю. Последний дает возможность использовать методы одномерной выборки, минуя методологические и организационные сложности многомерного отбора, но не позволяет одновременно учесть и числовые, и атрибутивные параметры.

Автором предлагается комбинационный подход к формированию многомерной выборки, в котором сочетаются приемы многомерности, используемые как при расслоении в независимых признаках, так и при расслоении по композитному признаку. Разработана оптимизационная модель многомерной выборки на основе применения кластерного анализа.

Исследуемая совокупность делится с помощью методов кластерного анализа на однородные группы. Внутри каждой полученной группы выделяется основной (ведущий) признак, по которому осуществляется последующий случайный или механический отбор единиц в выборку. Если по ведущему показателю коэффициент вариации превышает 50%, возможно дополнительное расслоение внутри кластера. По каждому признаку считается стандартная ошибка выборки. Если она превышает допустимые границы, то возможно три способа ее снижения: 1) увеличение объема выборки в кластере; 2) дополнительное расслоение предприятий в кластере по ведущему признаку; 3) повторение процесса кластеризации, причем возможно использование того же метода кластеризации, что и ранее, но с увеличением числа шагов, либо использование итеративного метода с заданием числа кластеров $r > l$.

Кластеры включают предприятия, однородные с точки зрения исследуемых k признаков, поэтому предполагается, что экстраполяция выборочных данных на генеральную совокупность с помощью коэффициентов распространения даст репрезентативные результаты:

$$x_{Ej} = \sum_{r=1}^{l} \frac{n_r}{N_r} \cdot x_{brj}, \ (1)$$

$$x_{Ej} = \sum_{r=1}^{l} \sum_{h=1}^{b} \frac{n_{rh}}{N_{rh}} \cdot x_{brj}, \ (2)$$

где x_{Ej} - экстраполированное значение j-го признака;
x_{brj} - выборочное значение $j-$го признака в r-м кластере;
x_{brjh} - выборочное значение j-го признака в h-ой группе r-го кластера;
$r = \overline{1, l}$ - число кластеров, определенных по k признакам;
$h = \overline{1, в}$ - число групп в кластере, выделенных по ведущему признаку;
$j = \overline{1, k}$ - число признаков единицы наблюдения.
При выборе метода кластерного анализа следует учитывать:
1) размер совокупности;
2) возможность задания числа кластеров;

3) наличие непересекающихся кластеров.

Предполагается использовать сочетание агломеративного иерархического метода и итеративного метода k-средних, с помощью которых на практике решается до 70-80% задач кластерного анализа. Первый из методов не допускает пересечений, размытых кластеров и может использоваться для получения исходного разбиения совокупности на группы, если пользователь затрудняется задать количество кластеров; второй удобен для обработки больших статистических массивов.

При использовании агломеративного кластерного анализа для нормирования данных и определения метрического расстояния между объектами рекомендуется применять традиционные формулы:

$p = (x - \overline{x})/\sigma$,(3)

$d_{ij}^E = \sqrt{\sum_{k=1}^{m}(x_{ik} - x_{jk})^2}$, (4)

где p - нормированное значение признака;

x - индивидуальное значение признака;

$\overline{x}, \sigma$ - соответственно среднее значение и среднее квадратическое отклонение признака;

d_{ij}^E - евклидово расстояние между i-ым и j-ым объектами;

$k = \overline{1, m}$ - число признаков;

x_{ik}, x_{jk} - значение k-ой переменной соответственно у i-го и j-го объектов.

Мера сходства для объединения двух кластеров определяется методом “дальнего соседа”, т.е. степень сходства оценивается по степени сходства между наиболее отдаленными (несхожими) объектами кластеров. Метод “дальнего соседа” позволяет выделить четко различающиеся кластеры.

При использовании метода k-средних в качестве базовых данных выступают исходные значения переменных. После задания r случайно отобранных объектов, которые будут служить эталонами, или центрами, кластеров из оставшихся *(N - r)* объектов извлекается точка X_i с координатами $(x_{i1}, x_{i2,\ldots}x_{im})$ и проверяется, к какому из эталонов она находится ближе всего. Для этого применяется евклидово расстояние $\left(d_{ik} = \sqrt{\sum_{k=1}^{m}(x_{ik} - x_{jk})^2}\right)$. Проверяемый объект присоединяется к тому центру, которому соответствует min_{dik}. Эталон заменяется новым, пересчитанным с учетом присоединенной точки, его вес увеличивается на единицу. На следующем шаге выбирается новая точка и для нее процедуры повторяются. Через *(N - r)* шагов все точки (объекты) совокупности окажутся отнесенными к одному из rкластеров. С целью достижения устойчивости разбиения процесс кластеризации повторяется. Если новое разбиение совпадает с предыдущим, то кластеризация заканчивается.

В соответствии с предложенным алгоритмом в НИИ статистики разработана программа "Многомерная выборка", которая предназначена для проведения выборочных исследований по k признакам на основе конкретной базы данных,

сформированных на основе статотчетности по ф. № 1-МП, используя методы одномерной и многомерной выборки. Апробация программного обеспечения проведена на основе информационного массива МП г. Минска, отчитавшихся в 2005 г. по ф.№1-МП (годовая). Анализируемые признаки: объем производства продукции, работ, услуг; списочная численность работников в среднем за период; фонд заработной платы работников списочного и несписочного состава.

Поиск оптимальной выборочной совокупности осуществлялся в несколько этапов:

1) после создания генеральной совокупности методом кластерного анализа происходило ее разбиение на однородные группы (по трем признакам), т.е. кластеризация;

2) внутри кластеров производилось расслоение совокупности на группы по ведущему признаку (объем производства);

3) по каждому кластеру путем перебора различных методов выборочных обследований были построены оптимальные выборочные совокупности, причем критериями результативности выступали ошибки суммарных значений трех исследуемых признаков: "объем производства" ($x1$), "списочная численность работников" ($x2$), "фонд заработной платы" ($x3$).

Оценка итогов пробного обследования показала, что наиболее оптимальным представляется многомерная выборка с простым расслоением. Для больших объемов генеральной совокупности (свыше 400-500 единиц) целесообразно использовать одномерную выборку, которая позволяет получить допустимые ошибки выборки по всем трем показателям: так, в оптовой торговле по г. Минску ($N = 4159$) в результате построения 25-% одномерной выборки с простым расслоением ошибки выборки оказались менее 1%. Сводные экстраполированные значения исследуемых признаков в целом по г. Минску в 2005 г. составили: по $x1$ – 4720178 млн. руб., по $x2$ – 137085,6 млн. руб., по $x3$ – 612757,8 млн. руб., что лишь на 0,2; 0, 51 и 0,33% отличается от величин, полученных по данным сплошной отчетности. Общая доля отбора - 26,8%. Полученные результаты свидетельствуют о высокой репрезентативности выборочной совокупности.

Разработанное методологическое и программное обеспечение выборочных обследований малых предприятий предусматривает возможность проведения многомерного отбора, получения выборочной совокупности, репрезентативной по группе признаков и достаточно адекватно отражающей деятельность мелких хозяйственных единиц.