Stichprobenziehung mit dem SAS System

Burkhard Remppis

Universität Heidelberg
SAS Fellowship

Für eine Repräsentativitätsanalyse wurden aus einer ungarischen Volkszählungsdatei mit 10,4 Mio. Datensätzen mit Hilfe der SAS-Technologie je 30 Simulationsstichproben nach dem Auswahlplan des ungarischen Mikrozensus und analog zum Auswahlplan des deutschen Mikrozensus gezogen, und zwar jeweils mit einem 2%-Auswahlsatz. Der deutsche Mikrozensus basiert auf einer systematischen Stichprobe, der ungarische Mikrozensus auf einer dreistufigen, zweifach geschichteten Stichprobe, die in der Simulation auf zwei der drei Stufen nachvollzogen wurde. Die Stichprobenziehungen wurden auf einer Windows NT 4.0 Workstation (SP6a) mit Intel Pentium III Prozessor (400 MHz, 256 MB-RAM) am Geographischen Institut Heidelberg durchgeführt.

Zur Umgehung relativ komplexer eigener Programmierungen in SAS Base wurde auf die Stichprobenfunktionen der SEMMA-Technologie des SAS ENTERPRISE MINER zurückgegriffen. Hierbei lassen sich über in einem Projektfenster als Flussdiagramm angeordnete, menügesteuerte *Sample Nodes* sehr einfach Stichproben realisieren.
Gänzlich problemlos war die Ziehung der systematischen Stichproben: Im *Sample Node* wird hierzu die Stichprobenmethode ”Nth” (Ziehung jedes n-ten Datensatzes) sowie der Auswahlsatz (im vorliegenden Fall 2%) ausgewählt. Zur Sicherstellung der Zufallsauswahl wurde dem *Sample Node* ein *SAS Code* vorweggestellt, der mittels der Zuteilung von Zufallszahlen und der PROC SORT-Funktion eine zufällige Mischung der Datensätze gewährleistete. Alle 30 systematischen Stichproben konnten parallelgeschaltet innerhalb eines EM-Projektes durchgeführt werden. Die Berechnungszeit betrug ca. 3,5 h.

Bei der Durchführung der komplexen Stichprobensimulation des ungarischen Mikrozensus ergaben sich zwei Probleme: Zum einen die Ziehung gewichteter Stichproben, zum anderen eine Programminstabilität bei zu großen Projektdiagrammen im EM-Projektfenster.

Auf der ersten Auswahlstufe mussten ungarische Gemeinden mit Wahrscheinlichkeiten proportional zur Einwohnergröße gezogen werden. Der ENTERPRISE MINER unterstützt jedoch keine Stichproben mit gewichteten Wahrscheinlichkeiten. Die Ziehung der Auswahlgemeinden musste deshalb über eine Programmierung in SAS BASE auf Grundlage der RANUNI-Funktion erfolgen. Dabei wurde in Schleifenprogrammierungen die zweifache Stichprobenschichtung nach Regierungsbezirken und Gemeindegrößenklassen berücksichtigt. Die den Auswahlgemeinden zugehörigen Datensätze wurden anschließend zu neuen Grundgesamtheiten für die Stichprobenziehung zweiter Stufe zusammengefasst.

Die Stichprobenziehung zweiter Stufe wurde im ENTERPRISE MINER in Form von 129 Teilstichproben realisiert: eine Teilstichprobe je Schichtungszelle aus der Kombination von Regierungsbezirk und Gemeindegrößenklasse. Über *Data Partition Nodes* wurden jeweils nur die Datensätze der entsprechenden Schichtungszelle an die *Sample Nodes* weitergegeben. In den *Sample Nodes* wurden dann die 2%-Stichproben mittels der Einstellung ”stratified“ (geschichtet) gleichmäßig aus den Auswahlgemeinden der jeweiligen Schichtungszelle gezogen. Die 129 Teilstichproben wurden zum Schluss mittels eines *SAS Code* zur Gesamtstichprobe zusammengefasst.

Einschließlich *Data Input Node* und *SAS Code Node* bestand somit jede Gesamtstichprobe zweiter Stufe aus 260 Einzelknotenpunkten. Die Berechnungszeit betrug pro Gesamtstichprobe ca. 3 h. Aus Zeiteffizienzgründen wurden mehrere Gesamtstichproben in einem EM-Projektdiagramm parallelgeschaltet. Dementsprechend multiplizierte sich die Anzahl der

Einzelknotenpunkte pro Projektdiagramm. Ab fünf parallelgeschalteten Gesamtstichproben trat dabei Programminstabilität durch Überlastung des Arbeitsspeichers auf. Der ENTERPRISE MINER legt für jeden Projektknotenpunkt Arbeits- und Metadateien an. Die Programminstabilität war vermutlich auf die außerordentlich große Anzahl dieser EM-Dateien in den Projektdiagrammen zurückzuführen. Bei einer maximalen Anzahl von vier Gesamtstichproben je Projektdiagramm konnten dagegen die Stichproben zweiter Stufe ohne Probleme mit dem ENTERPRISE MINER realisiert werden.