Auswertung von heterogenen Daten mit SAS-Routinen unter GLP-Bedingungen

Jörg Schmidtke, BioMath GmbH, J.-Jungius Str.9 18059 Rostock

Aufgabenstellung

Die Durchführung statistischer Auswertungen von heterogen gespeicherten Daten ist in der Praxis oft sehr aufwendig und bedarf von den Anwendern ein umfangreiches Wissen sowohl über die vorliegenden Datenstrukturen als auch über die verfügbaren statistischen Routinen die innerhalb von SAS angeboten werden. Oft sind solche Auswertungen durch wissenschaftliche Mitarbeiter durchzuführen, die nicht über Expertenwissen hinsichtlich der Daten und der statistischen Möglichkeiten von SAS verfügen. Durch ein geeignetes Softarewerkzeug können dem Praktiker solche Auswertungen sehr vereinfacht werden. Dieses Werkzeug dient als Interface sowohl zu den heterogenen Daten als auch zu den statistischen SAS-Routinen. Weiterhin verfügt dieses Werkzeug über Expertenwissen bezüglich statistischer Auswertungen.
Im folgenden wird ein Werkzeug vorgestellt, daß von BioMath für das Institut für Toxikologie und Aerosolforschung der Fraunhofer Gesellschaft Hannover entwickelt wurde. Die zugrundeliegenden Strukturen wurden den vorliegenden institutspezifischen Systemeigenschaften angepaßt. Andere Systemvorrausetzungen wären ebenfalls denkbar.

System und Netzwerk

Das zu entwickelnde Werkzeug sollte folgenden Bedingungen genügen:

- **Statistische Standardauswertungen** sollen menüorientiert von verschiedenen Nutzern ausgeführt werden können.
- Die notwendigen einzelnen Arbeitsschritte zu den Analysen müssen in **SAA's** (Standardarbeits-anweisungen) festgeschrieben werden.
- Alle Arbeitsschritte und Benutzereingaben sollen **dokumentiert** werden (**GLP**).

Vor der eigentlichen Umsetzung dieser Aufgabe war es notwendig die Systemarchitektur und die bis dahin „manuelle" Durchführung statistischer Auswertungen zu analysieren. Der Datenbestand wurde in verschiedenen Systemen unter Open-VMS geführt. Weiterhin stand das SAS System ebenfalls unter VMS den Anwendern zur Verfügung. Der Zugang zu diesen einzelnen Systemen erfolgte über Terminalsitzungen von den einzelnen Arbeitsplätzen aus. Über die Netzwerkarchitektur wurden alle Sicherheitsmechanismen zur Verfügung gestellt. Der eigentliche Endanwender arbeitete mit einem Windows NT-System (siehe Abb.1).

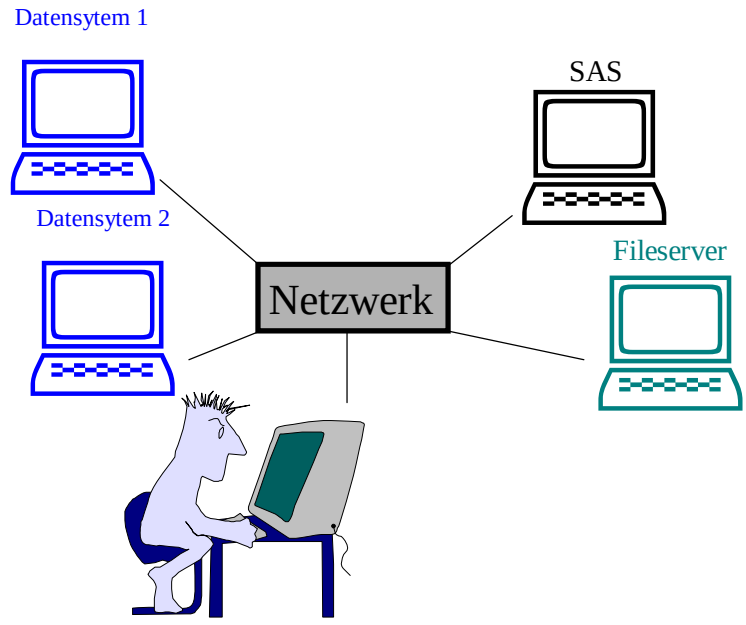

Abb.1

Netzwerksystem
Der Endanwender steuert ein Windows NT-System. Über das Netzwerk hat er Zugriff auf die verschiedenen Datensysteme, das SAS-System und einen Fileserver. Dieser Fileserver verwaltet die Standardsoftware (Textverarbeitung Tabellenkalkulation...)

Die herkömmliche „manuelle“ Arbeitsweise einer statistischen Auswertung ist in Abb.2 dargestellt:

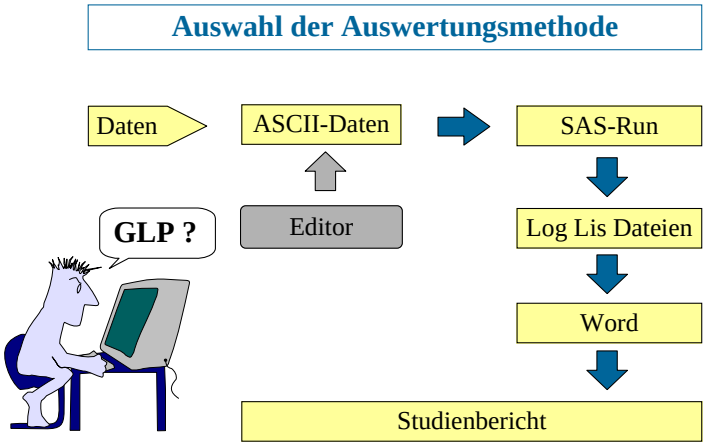

Abb. 2

Vor der eigentlichen statistischen Auswertung wurde zunächst die Auswertungsmethode festgelegt, danach die entsprechenden Daten aus den Datenbanksystemen exportiert. Mit Hilfe eines Editors erfolgte dann der Link des SAS-Source Codes und der Daten. Nach dem SAS-Run wurden die statistischen Ergebnisse zu einem entsprechenden Studienbericht in einem Word-Dokument zusammengestellt.

Diese „manuelle“ Bearbeitung statistischer Auswertungen verlangt von dem Anwender ein sehr hohes Maß an Konzentration und Sorgfalt. Weiterhin ist auch für die Kommunikation zwischen den unterschiedlichen Systemen ein erheblicher Zeitaufwand notwendig.

Die Hauptanwendung

Auf Grund der Systemvoraussetzungen bietet sich ein dynamisches NT-Programm an, daß die „manuelle“ Bearbeitung statistischer Auswertungen sehr vereinfachen kann. Die Abb.3 beschreibt die Übersicht der Programmarchitektur.

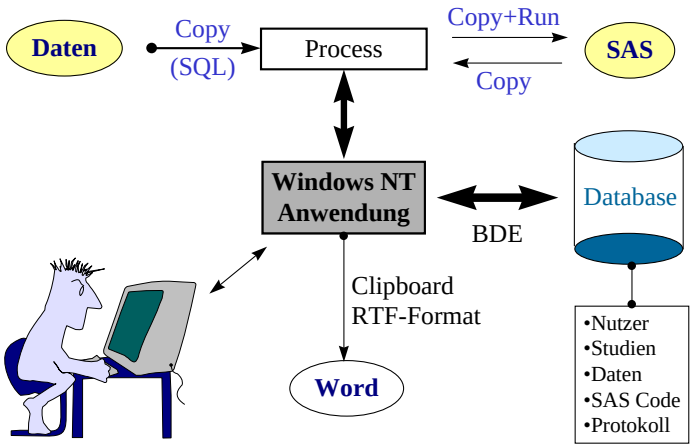

Abb.3

Die Hauptanwendung ist ein Windows NT Programm und bettet sich durch seine Standardbedienelemente optimal in das Windowssystem ein. Dieses System verfügt einerseits über ein Interface zum Zugriff auf die Daten und andererseits zum Zugriff auf das SAS-System. Dieses Interface kann an die gegebenen Systemeigenschaften entsprechend angepaßt werden. Auf der anderen Seite bildet die Datenbank den Kern für die Steuerung der Interface hinsichtlich der Datenbestände, des SAS-Systems und der Textverarbeitung.

Eine wichtige Schnittstelle des NT-Systems ist die Bereitstellung und Kontrolle von Processen. Diese Processe werden der Hauptanwendung über eine Konfigurationsdatei bekannt gemacht. Durch Steuerung dieser Processe kann dann auf den Datenbestand in sehr dynamischer Weise zugegriffen werden. Weiterhin wird über diese Architektur der SAS-Run durchgeführt. Alle diese Steuerungen erfolgen auf Nutzerebene unabhängig von Terminalsitzungen.
In der zur Verfügung stehende Datenbank (auf einem Fileserver) werden sowohl die Nutzerrechte als auch das Expertenwissen hinsichtlich der statistischen Auswertung durch einen Systemadministrator verwaltet.
Nach der Konfiguration des Systems kann die statistische Auswertung durch den Nutzer wie folgt durchgeführt werden:

- Laden (automatisches speichern und synchronisieren) der Auswertungsdaten
- Wahl der statistischen Auswertungsmethode über einen Selektionsbaum, der durch Experten bereitgestellt wurde
- Einschränkung der Datenmenge bezüglich der Auswertung
- SAS-Run und Import der Ergebnisse in einen Textverarbeitung

Das grundsätzliche Prinzip der automatischen Erzeugung eines lauffähigen SAS-Programms durch die Hauptanwendung ist in Abb.4 dargestellt.

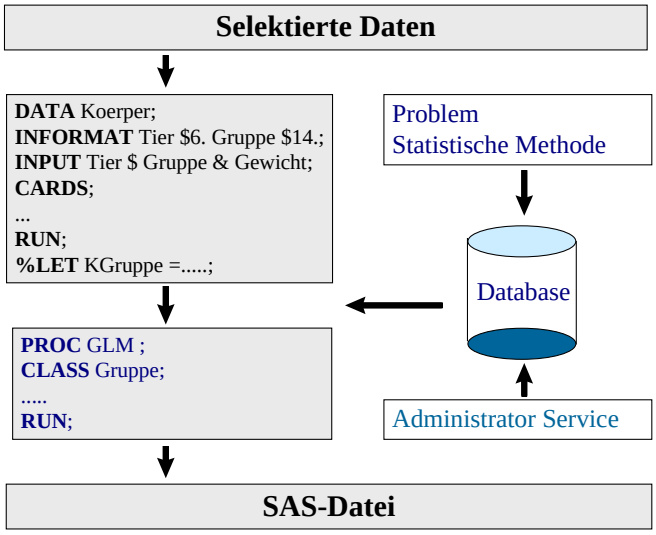

Abb.4

Auf Grund der vom Nutzer durchgeführten Datenselektion erzeugt die Hauptanwendung zunächst einen Data Step. Über einen Entscheidungsbaum wählt der Nutzer dann eine statistische Methode aus. Für diese Methode ist in der Datenbank, durch den Administrator, SAS-Source Code hinterlegt worden. Dieser Code wird entsprechend gelinkt.
Der so erzeugte SAS-Code (einschließlich Daten) wird über NT-Processe abgearbeitet und die Ergebnisse in die Textverarbeitung importiert. Alle Arbeitsschritte werden systematisch protokolliert und in der Datenbank gespeichert.

Zusammenfassung

Das vorgestellte System wurde mit Delphi 32bit entwickelt und setzt intern auf ein 32bit Windowsbetriebssystem auf, einschließlich Datenbank und Datenbanktreiber. Auf Grund der offenen Struktur des Systems ist eine Anpassung an gegebene Netzwerkarchitekturen relativ problemlos möglich. Die Schnittstelle zu den statistischen SAS-Routinen wird durch die Administratorsoftware bereitgestellt und kann den Erfordernissen der Auswertungen beliebig zugeschnitten werden. Das komplette System zeichnet sich durch folgendes aus:

- Synchronisation von **heterogenen** Daten
- **Experten** beschränken das System auf die relevanten statistischen Auswertungsmethoden
- Nutzer kann über einen **Selektionsbaum** die statistischen Methoden auswählen
- **Zentralisierung** aller Transaktionen
- **minimale** Systemanforderungen (Win95/NT + Netzwerk) mit sehr gutem Laufzeitverhalten