Statistische Auswertungen für Anwender ohne SAS Programmierkenntnisse

Jörg Schmidtke
BioMath GmbH
Schnickmannstraße 4
18055 Rostock
joerg.schmidtke@biomath.de

Wenke Mönkemeyer
BioMath GmbH
Schnickmannstraße 4
18055 Rostock
wenke.moenkemeyer@biomath.de

Kerstin Schmidt
BioMath GmbH
Schnickmannstraße 4
18055 Rostock
kerstin.schmidt@biomath.de

Zusammenfassung

Statistische Auswertungen mit SAS verlangen ein hohes Maß einerseits an Programmierkenntnissen und andererseits an fachlicher Kompetenz zur Interpretation der Ergebnisse.
In der Praxis, insbesondere bei immer wiederkehrenden gleichartigen Aufgaben, ist es sinnvoll diese Kompetenzen fachlich aufzuteilen, d.h. die Auswertungsprogramme von SAS-Programmierern so zu entwickeln, dass Fachexperten die Auswertungen einfach steuern können.
An Hand von praktischen Anwendungen bezüglich toxikologischer Studien, landwirtschaftlicher Versuche und Bewertungen von Risiken werden spezifische Lösungen von unterschiedlicher Komplexität dargestellt.

Schlüsselwörter: Toxikologie, Landwirtschaftliche Feldversuche, PIAFStat, SAS Enterprise Guide

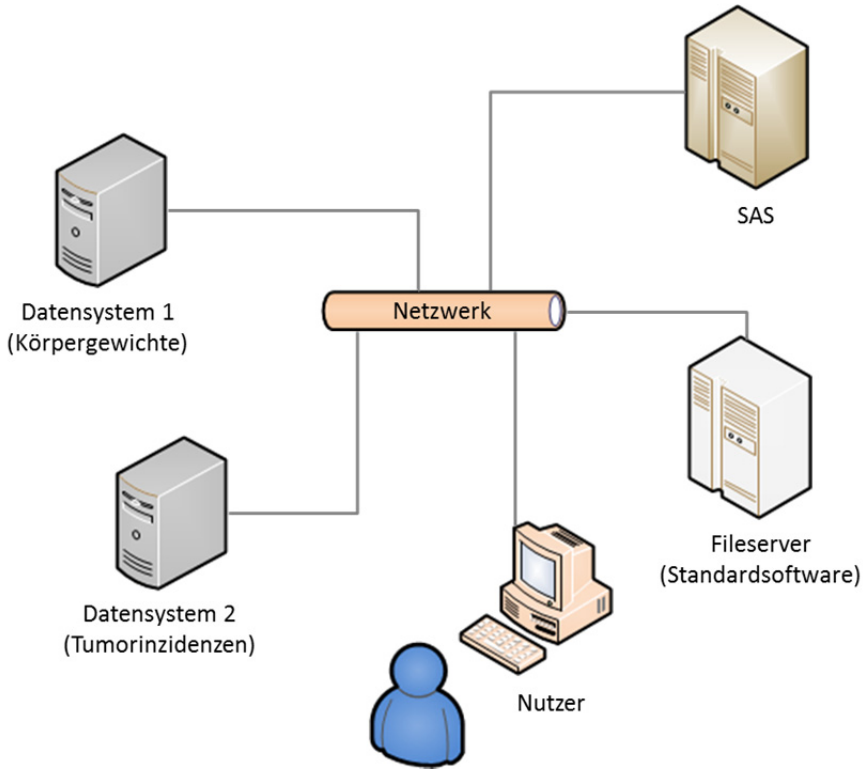
1 Einleitung

Die Durchführung statistischer Auswertungen verlangt von den Anwendern ein umfangreiches Wissen sowohl über die vorliegenden Datenstrukturen als auch über die statistischen Routinen, die in SAS zur Verfügung stehen. Oft sind solche Auswertungen durch Fachwissenschaftler durchzuführen, die nicht über ausgeprägte IT-Kenntnisse sowie Fachwissen zu den statistischen Möglichkeiten von SAS verfügen. Je flexibler die Auswertungsmöglichkeiten gestaltet werden sollen, um so eher sind Lösungen geeignet, die SAS im Hintergrund steuern.
Im Folgenden werden Softwarelösungen vorgestellt, die es den Fachexperten ermöglichen, komplexe statistische Auswertungen einfach durchzuführen.

2 Toxikologische Studien

Für ein toxikologisches Institut der Fraunhofer Gesellschaft sollte eine Lösung entwickelt werden, die den folgenden Bedingungen genügt.

- Die statistischen Standardauswertungen für Lebenszeit, Körpergewicht und Inzidenzen sollten menüorientiert von verschiedenen Nutzern ausgeführt werden können.
- Die notwendigen einzelnen Arbeitsschritte der Analysen sollten in Standardarbeitsanweisungen (SOP[1]) festgeschrieben werden.
- Alle Arbeitsschritte sollten dokumentiert werden (GLP[2]).
- Die Lösung sollte sich in die institutsspezifische IT-Struktur integrieren.



Die IT-Struktur stellte sich so dar, dass sowohl die Daten als auch das SAS- System in verschiedenen Umgebungen im Netzwerk zur Verfügung standen. Die Nutzer arbeiteten mit Windowssystemen und hatten über das Netzwerk Zugriff auf die verschiedenen Datensysteme, das SAS-System und einen Fileserver. In den Datensystemen wurden die Körpergewichte und die Tumorinzidenzen der Versuchstiere verwaltet. Auf dem Fileserver stand die Standardsoftware (Textverarbeitung, Tabellenkalkulation...) zur Verfügung.

In der herkömmlichen „manuellen“ Arbeitsweise wurde zunächst eine Auswertungsmethode aus der Lebenszeitanalyse, Gewichtsentwicklung oder Inzidenzanalyse festgelegt und danach die notwendigen Daten aus den Datenbanksystemen exportiert. Mit Hilfe eines Editors erfolgte dann die Anpassung des SAS-Source Codes dieser Auswertungsmethode an die Daten. Nach dem SAS-Run wurden die Ergebnisse in einen Studienbericht als Word-Dokument zusammengestellt (Abb. 1).

[1] SOP - Standard Operating Procedure / Standardarbeitsanweisung
[2] GLP - Good Laboratory Practice / Gute Laborpraxis

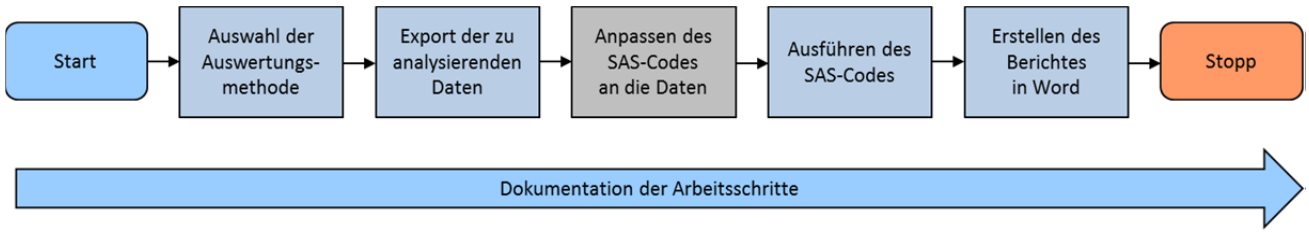


Abbildung 1: Herkömmliche manuelle Arbeitsweise einer statistischen Auswertung

Diese „manuelle“ Bearbeitung der statistischen Auswertungen verlangte von dem Anwender ein sehr hohes Maß an Sorgfalt, insbesondere auch hinsichtlich der Dokumentation der Arbeitsschritte.

Auf Grund der vorliegenden IT-Struktur und des heterogenen Datenbestandes wurde als Lösung ein Windows-Programm geschaffen, das die „manuelle“ Bearbeitung der statistischen Auswertungen automatisierte und gleichzeitig alle Arbeitsschritte protokollierte. Dieses Programm verfügt über Interface für den Zugriff auf die Daten, für die Steuerung des SAS-Systems und für die Nutzung der Textverarbeitung zur Erstellung der Studienberichte (Abb. 2).

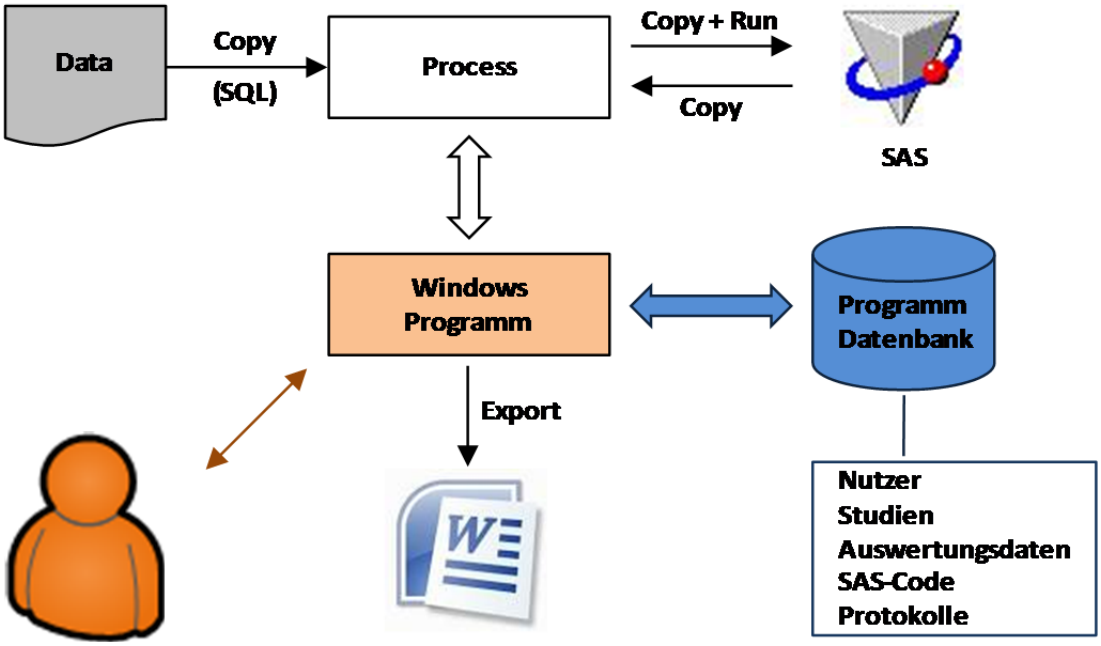


Abbildung 2: Technische Realisierung der Softwarelösung in der vorhandenen IT-Struktur

Die entwickelte Softwarelösung kann als Administrator oder als Nutzer ausgeführt werden. Im Administratormodus werden unter anderem auch die Auswertungsmethoden verwaltet. In diesem Modus kann der SAS-Programmierer den Quellcode für die Auswertungen organisieren, d.h. er kann jedes Analyseproblem (Lebenszeit, Körpergewicht oder Inzidenz) mit verschiedenen Methoden beliebig untersetzen. Der bereits vorhandene Programmquellcode der Auswertungsmethoden wurde angepasst und in die Softwarelösung eingefügt (Tab. 1).

Tabelle 1: Strukturierung der Auswertungsmethoden

Analyseproblem	Auswertungsmethode
Life Time Analysis	Kaplan Meier all animals pairwise
	Kaplan Meier only Tumor bearing animals
	Kaplan Meier all animals against control
Body Weight Analysis	Repeated Measures Analysis of Variance
	Repeated Measures [Proc Mixed]
Incidence Analysis	Numerical Incidence
	Relative Incidence
	Tumor Incidence [Fisher's Exact Test]
	CA Trend Test (non-neoplastic)
	PETO test (two-tailed)
	PETO test (one-tailed, upper)

Im Nutzermodus erfolgen das Datenmanagement und die statistische Auswertung ohne „Einblick“ in den SAS-Quellcode und ohne expliziten Aufruf der SAS-Umgebung. Folgende Arbeitsschritte werden dabei im Programm ausgeführt:

- Laden (automatisches Speichern und Synchronisieren) der Auswertungsdaten
- Wahl der statistischen Auswertungsmethode über einen Selektionsbaum, der durch SAS-Experten bereitgestellt wurde
- eventuelle notwendige Einschränkung der Datenmenge für die Auswertung
- SAS-Run und Import der Ergebnisse in die Textverarbeitung

Das grundsätzliche Prinzip der automatischen Erzeugung eines lauffähigen SAS-Programms ist in Abb.3 dargestellt.

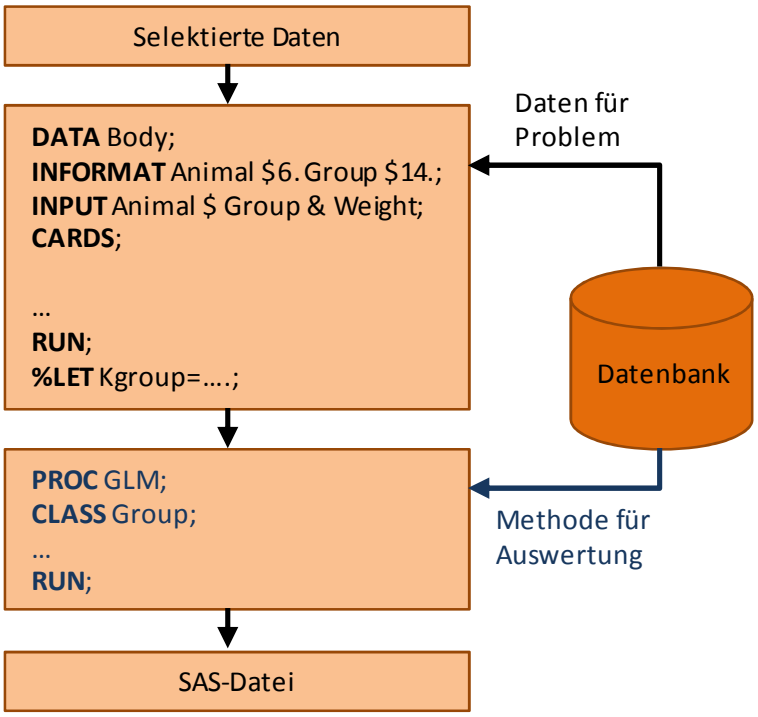


Abbildung 3: Prinzip der Generierung eines SAS-Programms

Auf Grund der vom Nutzer durchgeführten Datenselektion erzeugt das Programm zunächst einen Datenschritt. Für die gewählte Auswertungsmethode ist in der Datenbank durch den Administrator der SAS-Source Code hinterlegt worden. Dieser Code wird entsprechend gelinkt. Der erzeugte SAS-Code (einschließlich Daten) wird über System-Prozesse abgearbeitet und die Ergebnisse in die Textverarbeitung exportiert. Alle Arbeitsschritte werden protokolliert und in der Datenbank gespeichert.

Die kurz beschriebene Softwarelösung hat eine offene Struktur, die an gegebene Netzwerkarchitekturen einfach angepasst werden kann. Die statistischen SAS-Programme können im Administratormodus den Erfordernissen der Auswertungen beliebig zugeschnitten werden. Das komplette System zeichnet sich insbesondere durch die Synchronisation von heterogenen Daten, die Beschränkung auf die relevanten statistischen Auswertungsmethoden durch Experten und die Protokollierung der Arbeitsschritte aus.

3 Landwirtschaftliche Versuche

Im landwirtschaftlichen Feldversuchswesen der Bund-Länder-Gemeinschaft sollten die statistischen Auswertungen standardisiert werden. Die Versuchsdaten wurden bei jedem Mitglied der Gemeinschaft in einer eigenen PIAF[3]-Anwendung verwaltet und die statistischen Auswertungen erfolgten „manuell“ mit SAS. Es wurde nach einer Lösung gesucht die den folgenden Anforderungen genügt:

- Die statistischen Analysen sollten weiterhin mit SAS erfolgen.
- Die Auswertungsverfahren sollten nicht in die bestehende PIAF-Anwendung integriert werden.
- Statistische Verfahren sollten in der Gemeinschaft untereinander austauschbar sein.
- Der Anwender der statistischen Verfahren sollte die Möglichkeit einer Anpassung, ohne mit dem Quellcode in Berührung zu kommen, haben.

Als Lösung wurde das eigenständige Programm PIAFStat entwickelt, das die PIAF-Anwendung und SAS über eine statistische Verfahrensbibliothek verbindet. Die einzelnen Verfahren der Bibliothek sind SAS-Programme, die um syntaktische Elemente erweitert wurden. Durch die Einführung dieser Elemente können die Verfahren allgemein entwickelt werden. Erst durch die Zuordnung der Daten zu den Datenvariablen der syntaktischen Elemente wird ein lauffähiges SAS-Programm generiert. PIAFStat steuert danach das SAS-System und stellt die Ergebnisse für die weitere Bearbeitung zur Verfügung (Abb. 4).

[3] PIAF - Planung, Information und Auswertung von Feldversuchen

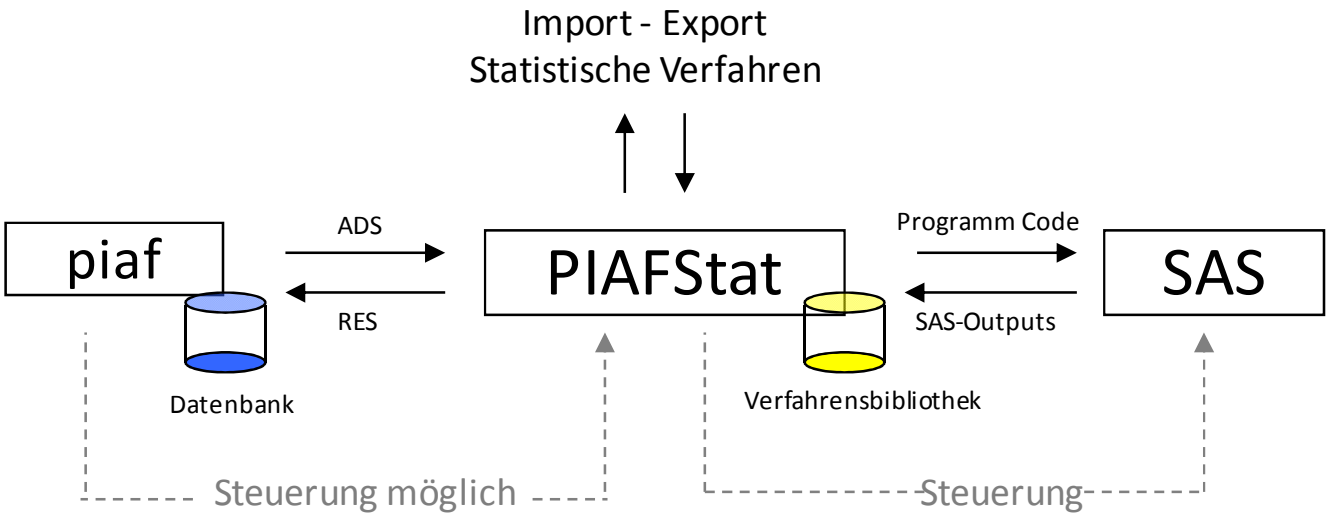


Abbildung 4: PIAFStat - Steuerung und Datenaustausch

Der SAS-Quellcode jedes Verfahrens wird in die Abschnitte Deklaration und Programm unterteilt. Im Abschnitt Deklaration werden alle benötigten Platzhalter für die

- Klassifizierungs-Merkmale
- Analyse-Merkmale
- Optionale Abschnitte und fest definierte Platzhalter

einschließlich deren Eigenschaften definiert. Im Abschnitt Programm wird das eigentliche SAS-Programm hinterlegt. Die in das SAS-Programm eingefügten Elemente sind definierte Platzhalter, die Statement-konform (z.B. für Variable) platziert werden.

Im folgenden Beispiel werden im Deklarationsabschnitt die Platzhalter K1 und K2 für die Klassifikations-Merkmale Sorte und Intensität definiert. Im Programmabschnitt kann dann [K1] und [K2] Statement-konform genutzt werden.

```
DEKLARATION
K1(L:Faktor Sorte, I:Faktor Sorte ,M:1-1,F1)
K2(L:Faktor Intensität, I:nur bei zweifaktoriellen Versuchen)
...

PROGRAMM
...
Data sortiment; Set sortiment; Keep s [K1] Name;
Run;
Data sortiment; Set sortiment; pgnr=[K1]; Diagramm=[K1];BB='B';
Run;
...
```

Die Inhalte der Platzhalter [K1] und [K2] werden nach Aufruf des Verfahrens vom Nutzer zugeordnet. In Abb. 5 hat der Nutzer die Datenvariable F1 (Sorte) dem Platzhalter [K1] (Faktor Sorte) zugeordnet.

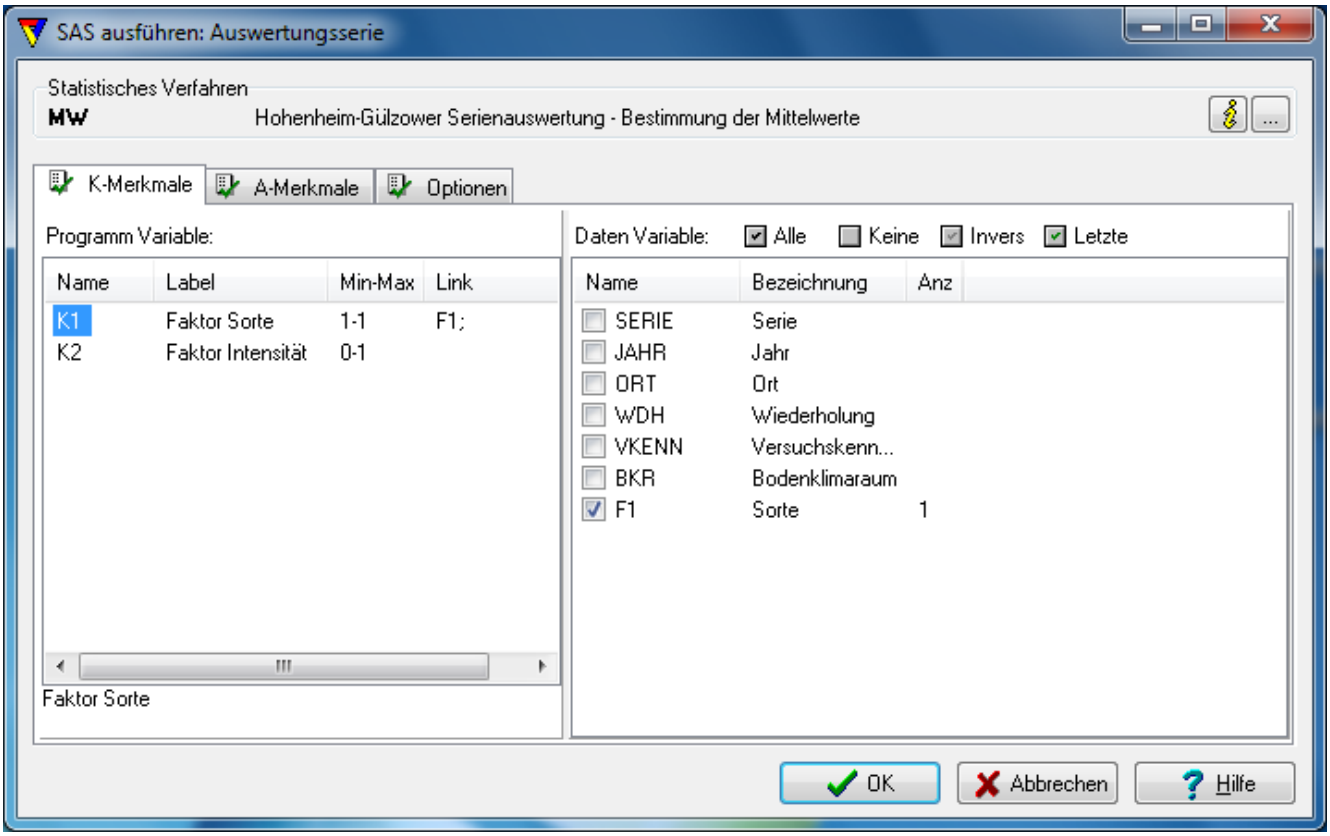


Abbildung 5: Nutzerdialog für die Zuordnung der Datenvariablen zu den Programmvariablen

PIAFStat ist eine komplexe IT-Lösung für die Auswertung von landwirtschaftlichen Versuchen. Die Auswertungsverfahren sind SAS-Programme mit zusätzlichen syntaktischen Elementen, über die eine umfangreiche Nutzersteuerung initiiert werden kann.

4 Bewertungen von Risiken

Für die Zulassung von gentechnisch veränderten Pflanzen muss eine Sicherheitsbewertung erfolgen. Dazu wurde ein Sicherheits-Prüfsystem, das technisch als Decision Support System (DSS) realisiert wurde, entwickelt (Abb. 6). Das System besteht aus den Kernbestandteilen:

- Datenkomponente
- Methodenkomponente und
- Dialogkomponente

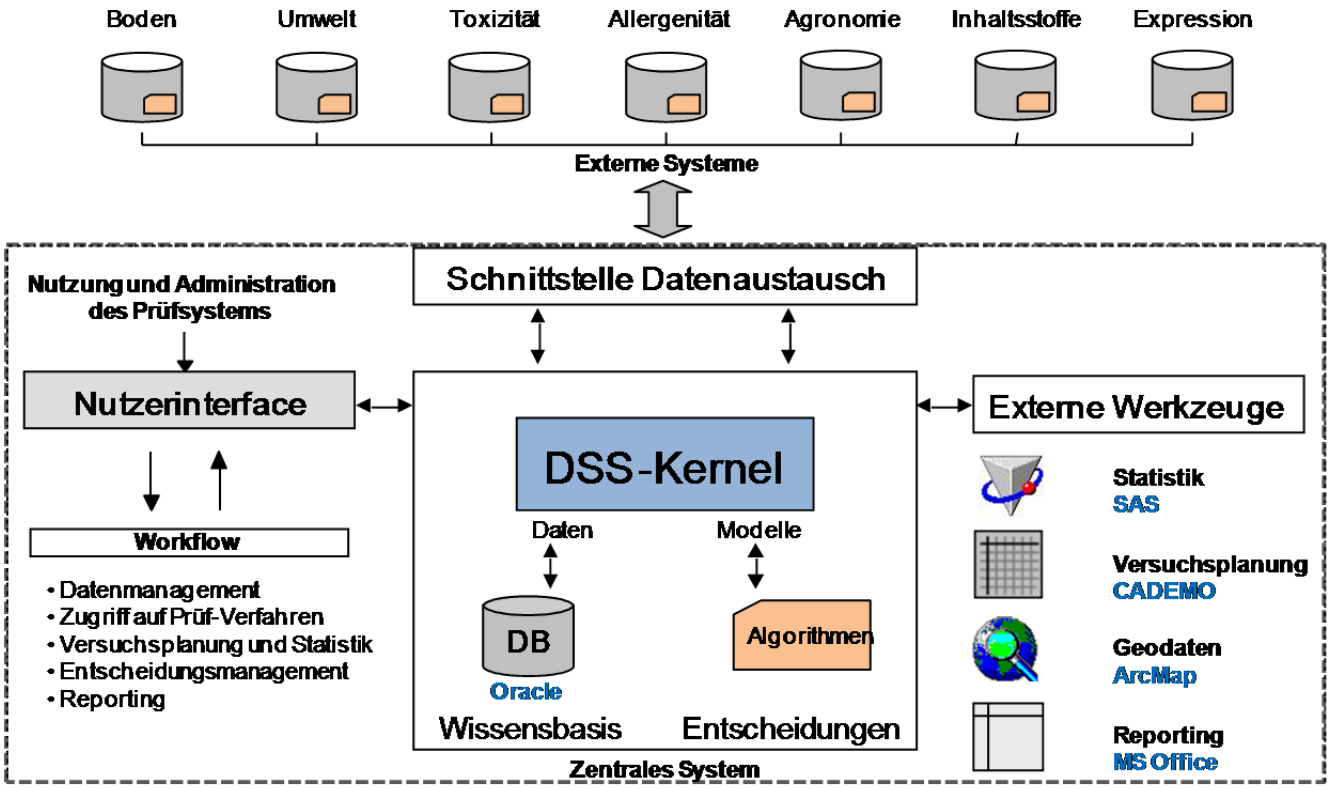


Abbildung 6: DSS mit SAS als externes Werkzeug zur statistischen Analyse

Ein wichtiger Teil der Methodenkomponente sind die statistischen Analysen für die Risikobestimmung. Die statistischen Verfahren sind strukturiert nach der Herkunft der Indikatormerkmale (Boden, Umwelt, Inhaltsstoffe, …) und in der zentralen Datenbank als SAS-Programme gespeichert.

Im Ablauf der Sicherheitsprüfung wird das gesamte Spektrum der statistischen Verfahren abgearbeitet. Der Nutzer/Fachexperte selektiert jeweils die dazu notwendigen Daten, wählt das statistische Verfahren, führt es aus und speichert die Ergebnisse in das DSS zurück. Diese Ergebnisse gehen dann in weitere Algorithmen des Entscheidungsbaumes der Sicherheitsprüfung ein.

Die technische Realisierung der Abarbeitung der statistischen Verfahren ist dem Prinzip der toxikologischen Studien (Abschnitt 1) sehr ähnlich. D.h. abhängig von der Datenherkunft wird ein Datenschritt erzeugt und durch die gewählte Auswertungsmethode zu einem lauffähigen SAS-Programm ergänzt.

Das DSS stellt dem Fachexperten die Möglichkeit zur Verfügung komplexe Versuchs-Datenstrukturen mit SAS auszuwerten. Der SAS-Programmierer kann über eine Schnittstelle die statistischen Verfahren zur Risikoanalyse in das System integrieren.

5 Weitere Möglichkeiten

In einer Teilaufgabe bei der Entwicklung von SAS-Programmierrichtlinien für eine europäische Behörde sollte ein Makro bereitgestellt werden, das standardisierte Programmstrukturen generiert und diese in Dateien zur weiteren Bearbeitung speichert. Dieses Makro sollte sowohl von SAS-Programmierern als auch von Enterprise Guide Nutzern verwendet werden.

Die Argumente des entwickelten Makros CreateProgramTemplateFiles sind Zeichenketten unterschiedlicher Verwendung, numerische Werte mit festgelegten Wertebereichen und Flags zur Steuerung des Programmablaufes (Tab. 2).

Tabelle 2: Auszug aus der Argumentenliste des Makros CreateProgramTemplateFiles

Argument	Type	Description
project_short_name	Input	Project short name
author	Input	Name of the author (optional)
number_of_parts	Input	Number of parts (0 to 10)
is_part_external	Input	Flag for external parts
		TRUE all parts are in external files
		FASLE all parts in the main program
path_of_project	Input	Project path
returncode	Output	Flag of the macro result
		0 - successful
		1 - invalid argument project_short_name
		2 - argument number_of_parts out of range
		3 - project path does not exists

Die SAS-Programmierer verwenden dieses Makro wie gewohnt in der Entwicklungsumgebung, die Enterprise Guide-Nutzer erwarten einen Dialog zur Eingabe der Argumente. Dafür kann im Enterprise Guide der Prompt Manager genutzt werden. Mit dessen Mechanismus ist es möglich, SAS-Programme auf der Ebene dieser Nutzer auszuführen.

Das Makro wurde zunächst in ein SAS-Programm gekapselt in dem globale Variablen für die Argumente des Makros definiert wurden.

```
/*******************   BEGIN OF PROGRAM  ************************/

/* include MacroCreateProgramTemplateFiles.sas */
%INCLUDE "...\MacroCreateProgramTemplateFiles.sas";

/* define all variable for use at marco arguments */

%let result       = ;     /* return code */
```

```
%let prjshortname = ;     /* project short name */
%let author       = ;     /* name of the author */
%let nparts       = 1;    /* number of parts (0 to 10) */
%let partexternal = TRUE; /* flag for external parts (TRUE,FALSE) */
%let pathproject  = ;     /* project path */

/* call the marco */
%CreateProgramTemplateFiles(&prjshortname,...,result);

/* output the return code of the macro */
%put Result = &result;

/*********************   END OF PROGRAM   ************************/
```

Mit dem Prompt Manager wurden dann die globalen Variablen festgelegt, die über eine Dialogbox vom Nutzer abgefragt werden. Jede Variable erhielt dabei den notwendigen Prompt Typ.

Tabelle 3:

Variable	Prompt Type	Description
prjshortname	Text	Project short name
author	Text	Name of the author
nparts	Numeric (integer in range)	Number of parts (0 to 10)
partexternal	Text (fixed text to selected)	Flag for external parts (TRUE,FALSE)
pathproject	Directory	Project path

Der Prompt Typ für die Variable nparts wurde so angepasst, dass der Nutzer nur noch zwischen den Integer-Werten 0 bis 10 auswählen kann. Analoges gilt für das Flag bezüglich der Auswahl der Zeichenketten TRUE und FALSE.

Wird das Programm im Enterprise Guide ausgeführt, so kann der Nutzer in einer Dialogbox, die auf Grund der Festlegungen durch den Prompt Manager generiert wird, die Programmparameter eingeben (Abb. 7). Danach wird das Programm weiter ausgeführt.

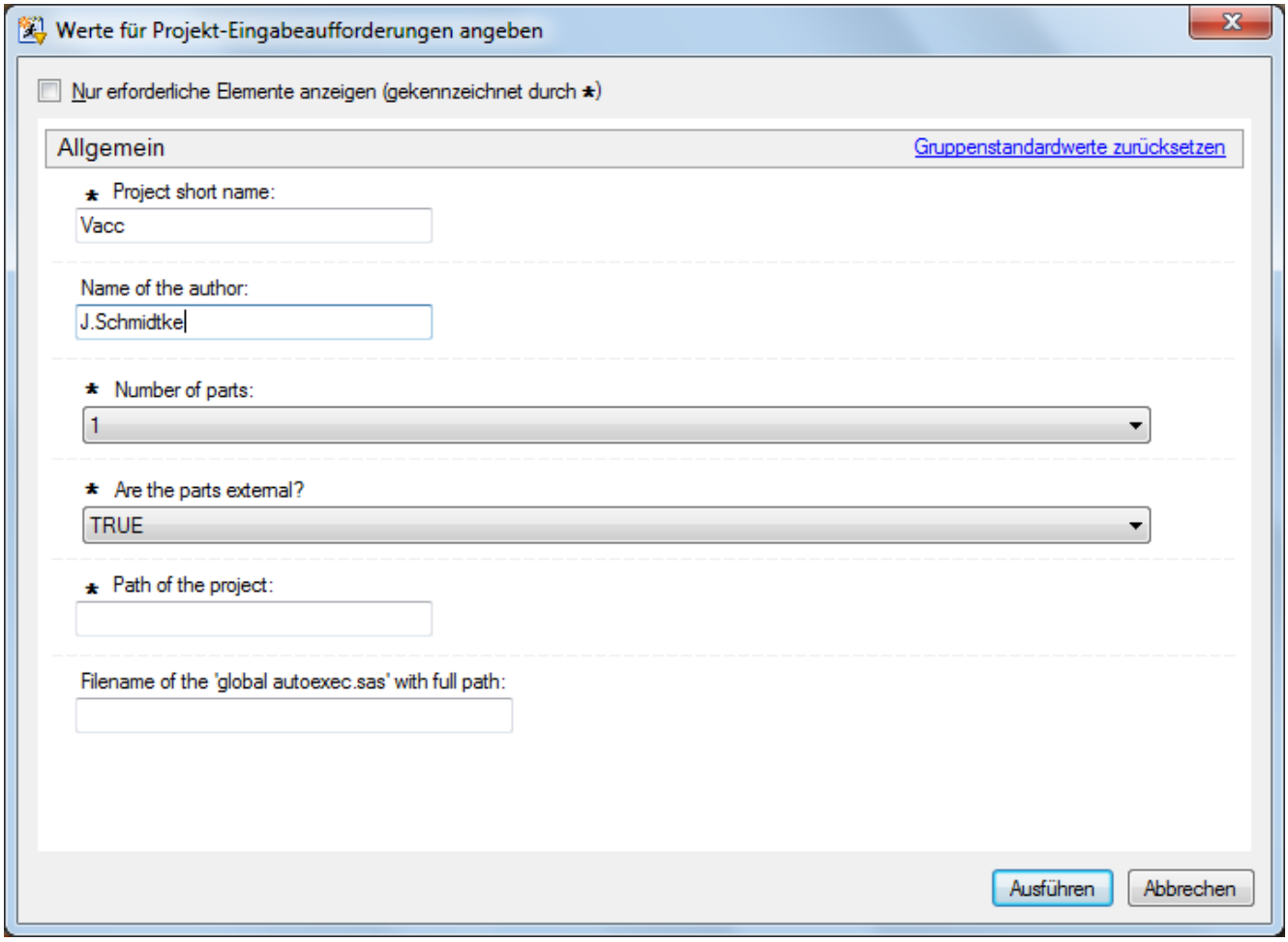


Abbildung 7: Dialogbox zur Eingabe der Programmparameter

Der Prompt Manager ist geeignet um SAS-Programme dialogorientiert dem Enterprise Guide-Nutzer zur Verfügung zu stellen. Eine übersichtliche Beschreibung der Funktionalitäten ist in [1] gegeben.

6 Fazit

Sollen die fachlichen Kompetenzen für statistische Auswertungen in SAS auf Experten aufgeteilt werden, so sind zunächst die vorhandenen Organisations- und IT-Strukturen zu analysieren und dann auf Grundlage der spezifischen Anforderungen eine Lösung zu entwickeln.
Liegen heterogene Datenstrukturen vor oder sollen dynamische Anpassungen des SAS-Codes durch den Endanwender möglich sein, dann könnte eine individuelle Softwarelösung geeignet sein, die das Daten- und Auswertungsmanagement nutzerfreundlich kapselt und SAS im Hintergrund steuert.
Wird bereits der Enterprise Guide für die Auswertung verwendet, so sind die Funktionalitäten mit zusätzlichen SAS-Programmen erweiterbar. Mit Hilfe des Prompt Managers können solche Programme dialoggesteuert ausgeführt werden.

Literatur

[1] Angela Hall. Creating Reusable Programs by Using SAS Enterprise Guide Prompt Manager, Paper 309-2011, SAS Global Forum 2011
http://support.sas.com/resources/papers/proceedings11/309-2011.pdf