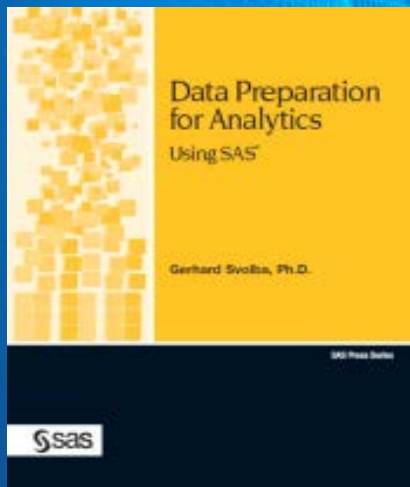


# SIE WOLLEN RECHTZEITIG EIN BILD ÜBER DIE DATENQUALITÄT IHRER ANALYSEDATEN HABEN? – SAS UND JMP HELFEN IHNEN DABEI



**KSFE 2013, ULM, 1. MÄRZ 2013  
DR. GERHARD SVOLBA - SAS AUSTRIA**



# INHALT

Live Demo →

Live Demo →

- Die Idee von „Data Quality for Analytics“
- Datenqualitäts-Profiling von Querschnittsdaten
- Datenqualitäts-Profiling von Längsschnittsdaten
- Entdecken von Auffälligkeiten in den Daten

# DIE IDEE VON „DATA QUALITY FOR ANALYTICS“





“Das gesamte Ökosystem (Entscheidungen, Kriterien, Datenaufbereitungsschritte, fachliche Überlegungen) das zwischen der **FACHLICHEN FRAGESTELLUNG** und dem **FINALEN ANALYTISCHEN MART** liegt.

Fachliche  
Fragestellung

DATA PREPARATION,  
DATA QUALITY

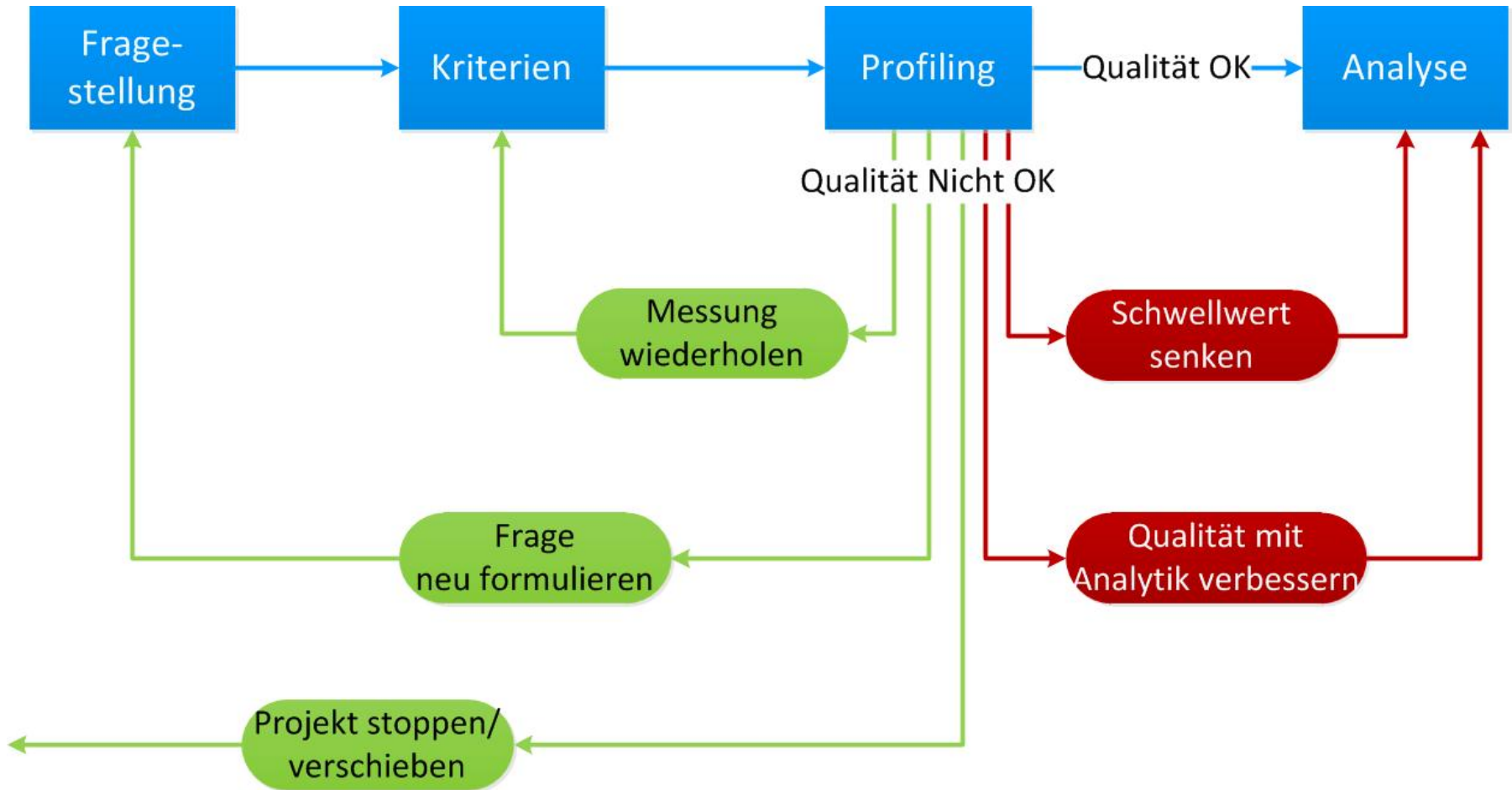
1. Review der fachlichen Fragestellung
2. Definition der Datenqualitätskriterien
3. Erzeugen eines Data Marts aus den Inputdaten
4. Profiling des Datenqualitäts-Satus
5. Entscheidung über den Datenqualitätsstatus
6. Aufbereitung der Daten für die analytische Verwendung

Finaler analytischer  
Mart

	Customer ID	Date of Birth	Age (years)	Gender	Marital St
1	1000002	26DEC1958	44	Male	Married
2	1000005	25JUN1947	56	Male	Single
3	1000006	10DEC1945	57	Female	Married
4	1000007	02JUN1934	69	Male	Married
5	1000008	15DEC1957	45	Male	Single
6	1000009	11MAR1959	44	Male	Single
7	1000014	23AUG1952	51	Male	Single
8	1000015	12MAY1959	44	Male	Single
9	1000016	11FEB1967	36	Male	Married

	CUSTOMER	TIME	PRODUCT
1	0	0	herring
2	0	1	comet_b
3	0	2	olives
4	0	3	ham
5	0	4	turkey
6	0	5	bourbon
7	0	6	ice_cream
8	1	0	baguette
9	1	1	soda
10	1	2	herring
11	1	3	cracker
12	1	4	herring
13	1	5	olives
14	1	6	comet_b
15	2	0	avocado
16	2	1	cracker
17	2	2	anchovy
18	2	3	herring
19	2	4	ham
20	2	5	turkey
21	2	6	sardines

## DER ANALYSEPROZESS UND DIE ENTSCHEIDUNG ÜBER DIE DATENQUALITÄT



# ENTDECKUNG UND BEHANDLUNG VON FEHLENDEN WERTEN IN QUERSCHNITTSDATEN



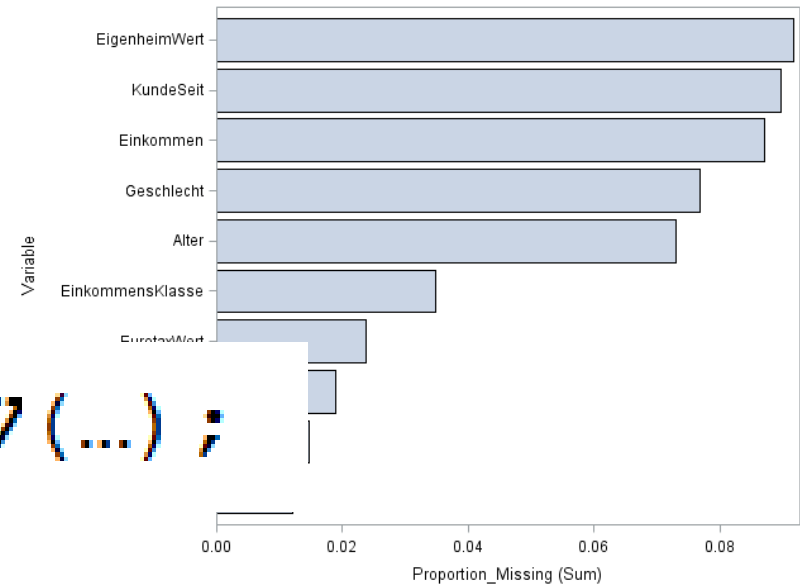


# FEHLENDE WERTE IN QUERSCHNITTS- DATEN

## Univariate Analyse

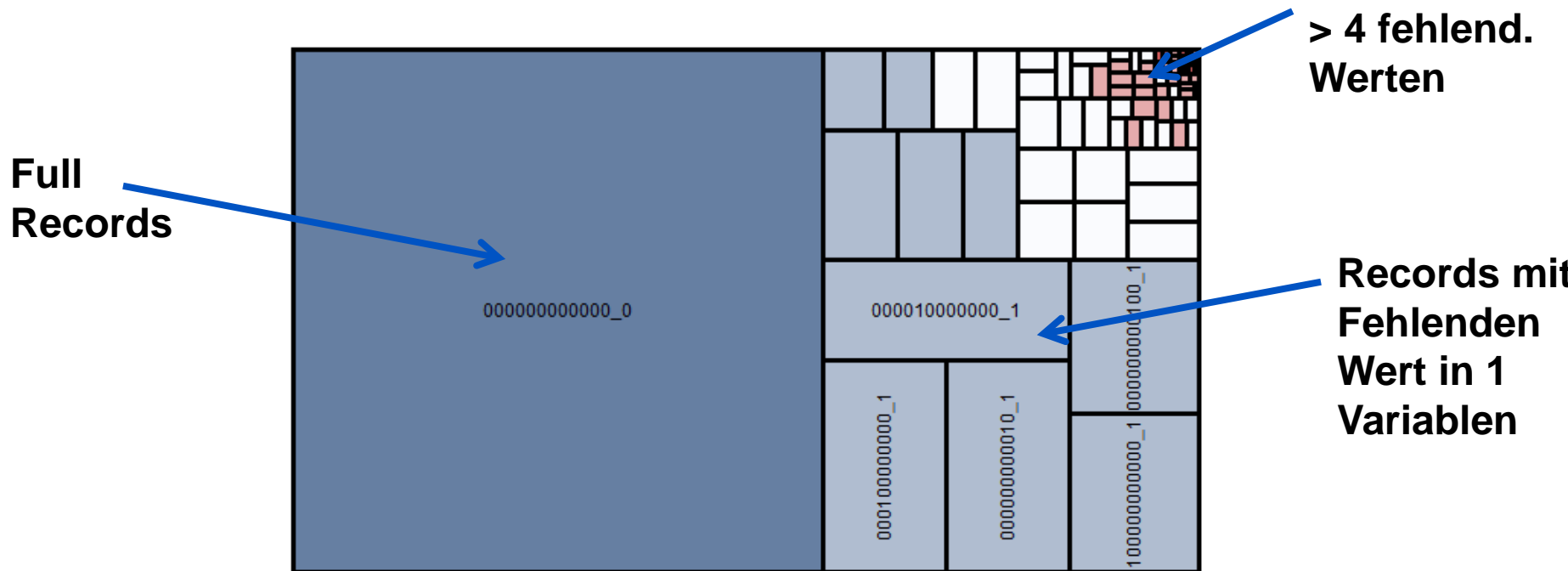
Obs	Variable	NumberMissing	Proportion_Missing	N
1	Alter	753	0.07	10303
2	EigenheimWert	945	0.09	10303
3	Einkommen	898	0.09	10303
4	EinkommensKlasse	359	0.03	10303
5	EurotaxKlasse	196	0.02	10303
6	EurotaxWert	211	0.02	10303
7	Fuehrerscheinenzug	1	0.00	10303
8	Geschlecht	0	0.00	10303
9	KundeSeit	0	0.00	10303
10	VermahnungsPunkte	124	0.01	10303

**% COUNT\_MV (...);**



- Welche Variablen in meinen Daten leiden unter der „Fehlende-Werte Krankheit“?
- Betrachtung aber nur aus einer „Spalten-Perspektive“
- Nicht: Wie viele „Full-Records“ habe ich?
- Nicht: Gibt es ein Muster in der Struktur der fehlenden Daten?

- Zusammenhängen der „Fehlender Wert“-Indikatoren für jede Variable zu einer Kette. Z.B: 00100100



```
%MV_Profiling(data=EM.KFZ_STORNO_DQ,  
vars= Alter AutoTyp AutoVerwendung EigenheimWert Einkommen  
EinkommensKlasse  
EurotaxKlasse EurotaxWert Fuehrerscheinenzug Geschlecht KundeSeit  
Vermahnungspunkte );
```

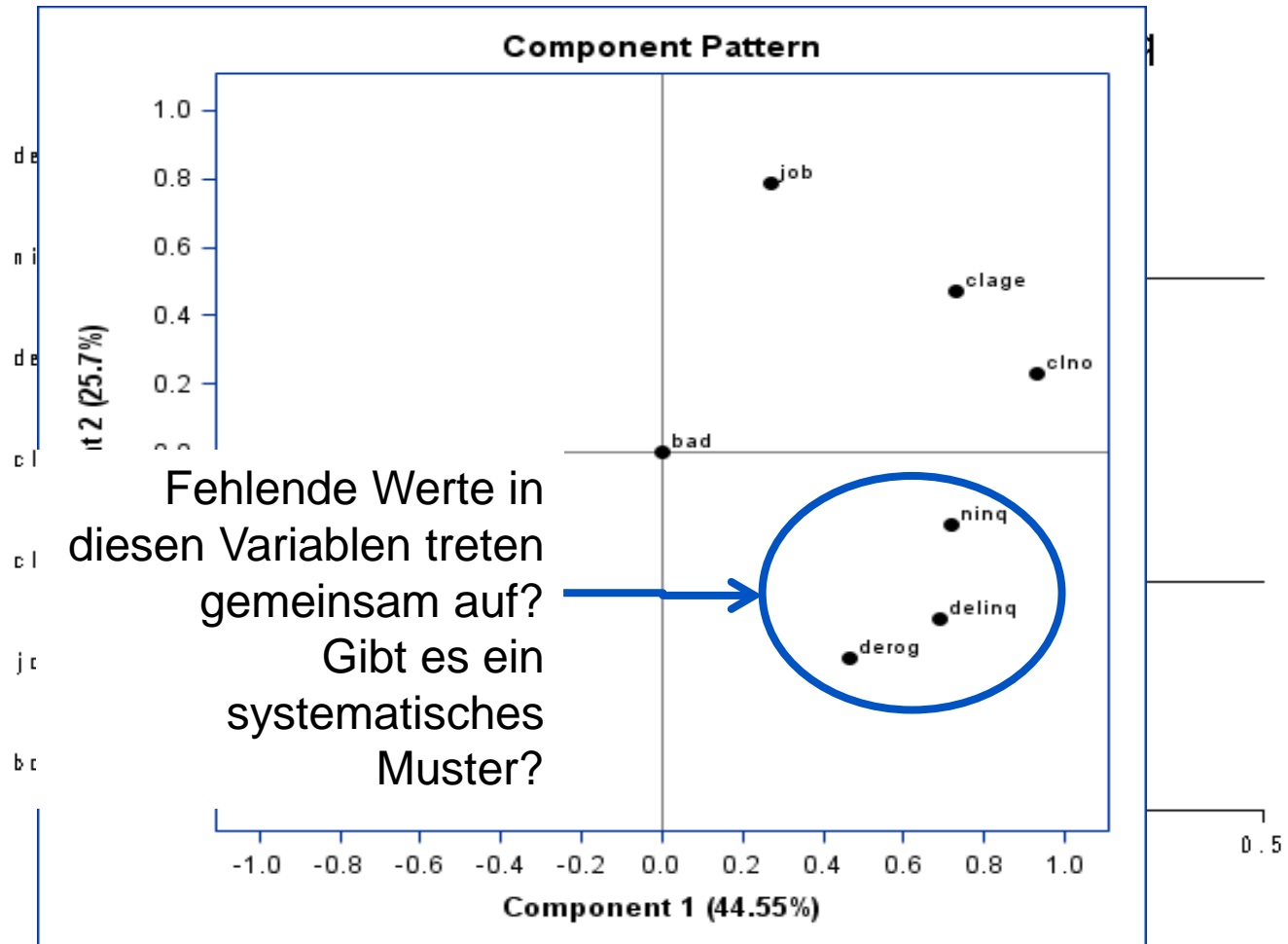


FEHLENDE  
WERTE  
IN QUERSCHNITTS-  
DATEN

Multivariate Analyse zeigt Muster in den fehlenden Werten

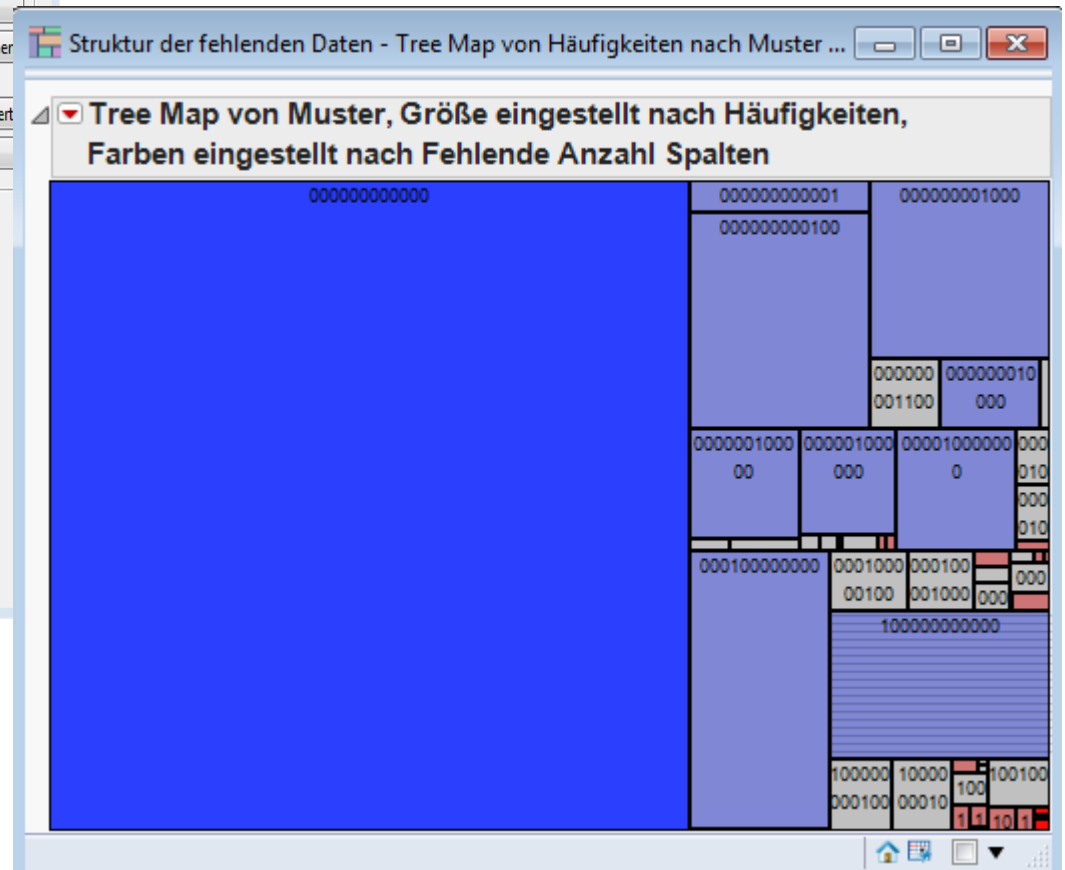
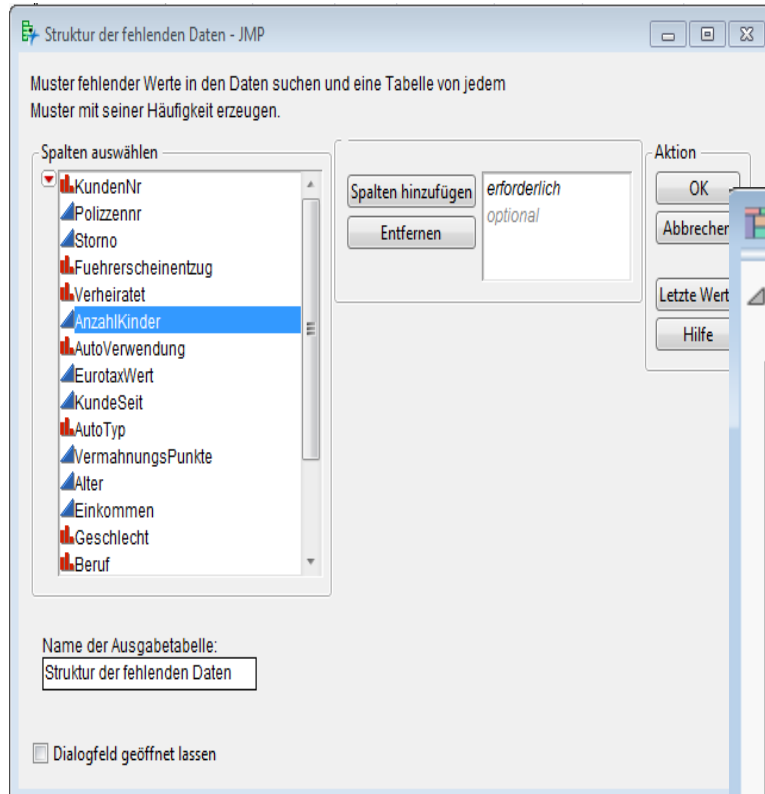
Variablen  
Clustering

Principal  
Components



# FEHLENDE WERTE IN QUERSCHNITTS- DATEN

## „Struktur der fehlenden Daten“ Task in JMP



# FEHLENDE WERTE IN QUERSCHNITTS- DATEN

## ERSETZEN VON FEHLENDEN WERTEN MIT DEM SAS ENTERPRISE MINER UND SAS STAT

### SAS Enterprise Miner



Method

Mean

Median

Distribution

Tree

Tree Surrogate

Midrange

Mid-Minimum

Tukey's Biweight

Huber

Andrew's Wave

Default constant

### SAS STAT – PROC MI

```
proc mi data = CreditData  
        out = CreditData_MI;  
    var clage mortdue value;  
run;
```

```
proc logistic data = CreditData_MI;  
    model bad(event='1') =  
        clage mortdue value / covb ;  
    ods output  
        ParameterEstimates = Estimates  
        Covb = CovMatrix;  
    by _Imputation_;  
run;
```

```
proc mianalyze parms=Estimates  
               covb=CovMatrix;  
    modeleffects intercept  
        clage  
        mortdue  
        value;  
run;
```

# ENTDECKUNG UND BEHANDLUNG VON FEHLENDEN WERTEN IN LÄNGSSCHNITTSDATEN

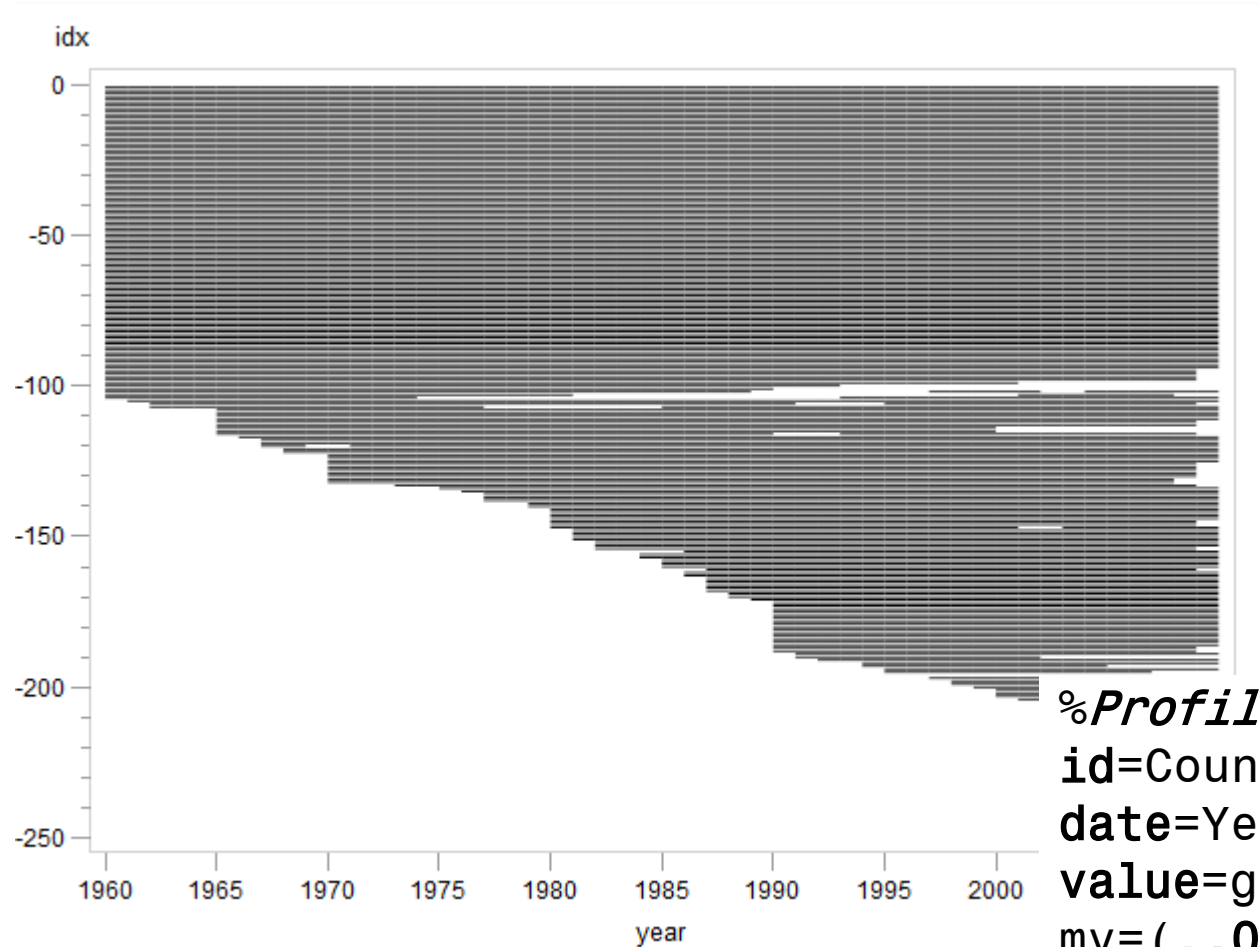


# Profiling der Struktur der fehlenden Werte und der 0-Werte in Zeitreihendaten

## Fehlender Wert

# FEHLENDE WERTE IN LÄNGSSCHNITTS- DATEN

Profiling der Struktur der fehlenden Werte und der  
0-Werte in Zeitreihendaten




```
%Profile_TS_MV(data=tmp.gdp2,  
id=Country_name,  
date=Year,  
value=gdp,  
mv=(.,0),  
w=1,  
NMAX_TS=300);
```




# FEHLENDE WERTE IN LÄNGSSCHNITTS- DATEN

## ERSETZEN VON FEHLENDEN WERTEN MIT PROC EXPAND (SAS ETS)

	DATE	AIR	air_mv
1	JAN49	112	112
2	FEB49	118	118
3	MAR49	132	132
4	APR49	129	129
5	MAY49	121	.
6	JUN49	135	135
7	JUL49	148	.
8	AUG49	148	148
9	SEP49	136	136
10	OCT49	119	119
11	NOV49	104	.
12	DEC49	118	118
13	JAN50	115	115
14	FEB50	126	126
15	MAR50	141	141



```
proc expand data = air_missing  
            out = air_expand;  
    convert air_mv=air_expand;  
    id date;  
run;
```



	date	air	air_mv	air_expand
1	JAN49	112	112	112
2	FEB49	118	118	118
3	MAR49	132	132	132
4	APR49	129	129	129
5	MAY49	121	.	128.29783049
6	JUN49	135	135	135
7	JUL49	148	.	144.73734152
8	AUG49	148	148	148
9	SEP49	136	136	136
10	OCT49	119	119	119
11	NOV49	104	.	116.19900978
12	DEC49	118	118	118
13	JAN50	115	115	115
14	FEB50	126	126	126
15	MAR50	141	141	141
16	APR50	135	135	135
17	MAY50	125	125	125

# FEHLENDE WERTE IN LÄNGSSCHNITTS- DATEN

## AUFFINDEN VON FEHLENDEN RECORDS MIT PROC TIMESERIES (SAS ETS)

PNR	date	air
56	2004-02-01	
56	2004-03-01	
56	2004-04-01	
56	2004-05-01	
56	2004-06-01	
56	2004-07-01	
56	2004-08-01	48
56	2004-09-01	36
56	2004-10-01	66
56	2004-11-01	15
56	2004-12-01	33
58	2005-06-01	39
58	2005-07-01	63
58	2005-08-01	84
58	2005-09-01	18
58	2005-12-01	69
58	2006-03-01	0
58	2006-07-01	90
58	2006-10-01	57
58	2007-01-01	48

```
proc timeseries data = air_missing
                out  = TIMEID_INSERTED;
  id date interval = MONTH setmiss=0;
  var air;
run;
```

← Record  
vorhanden,  
Wert fehlt

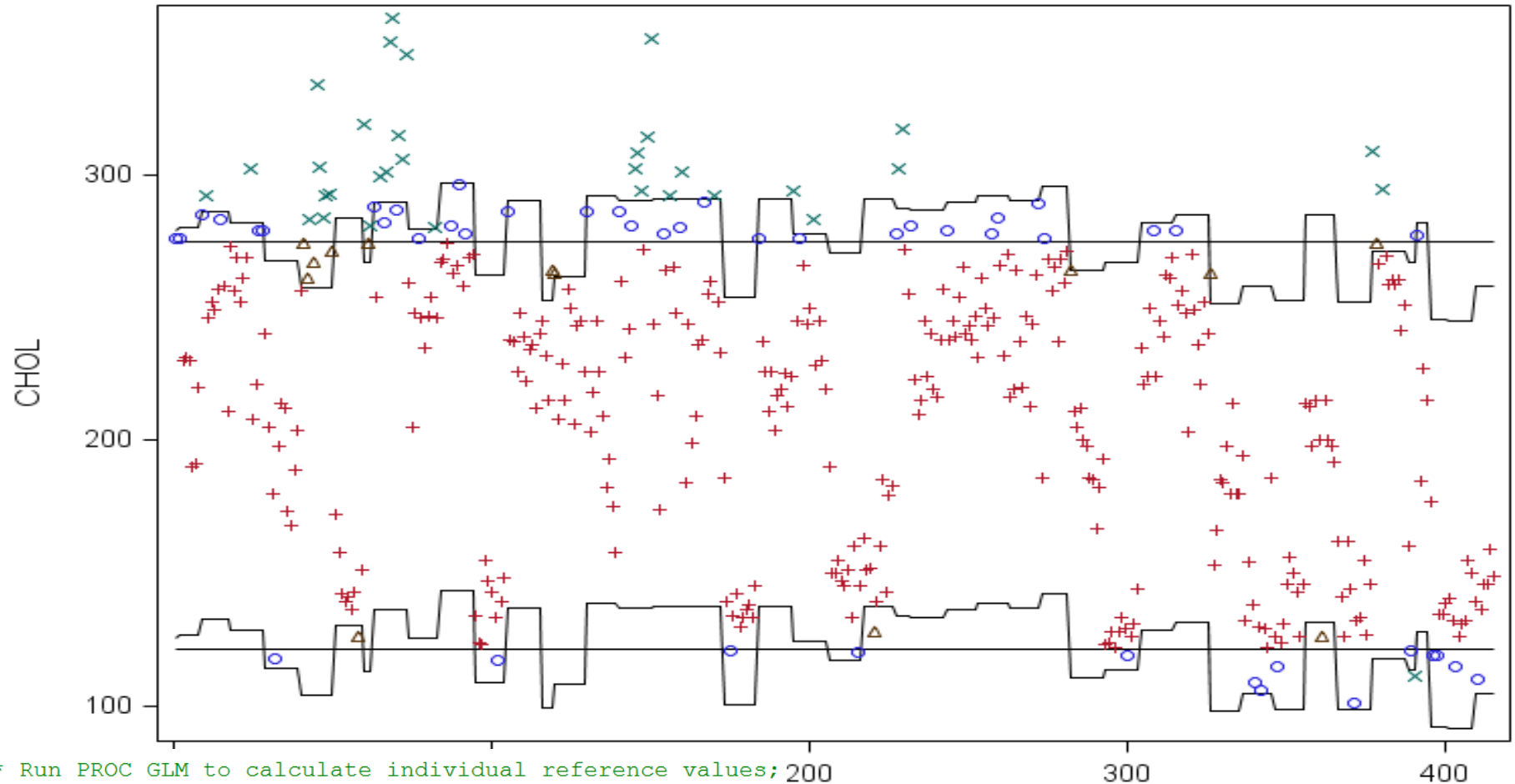
← ?

Fehlende Records  
Kontinuität nicht  
gegeben

# ENTDECKEN VON AUFFÄLLIGKEITEN IN DEN DATEN



# AUFFINDEN VON AUFFÄLLIGKEITEN MIT INDIVIDUELLEN VALIDIERUNGSLIMITS



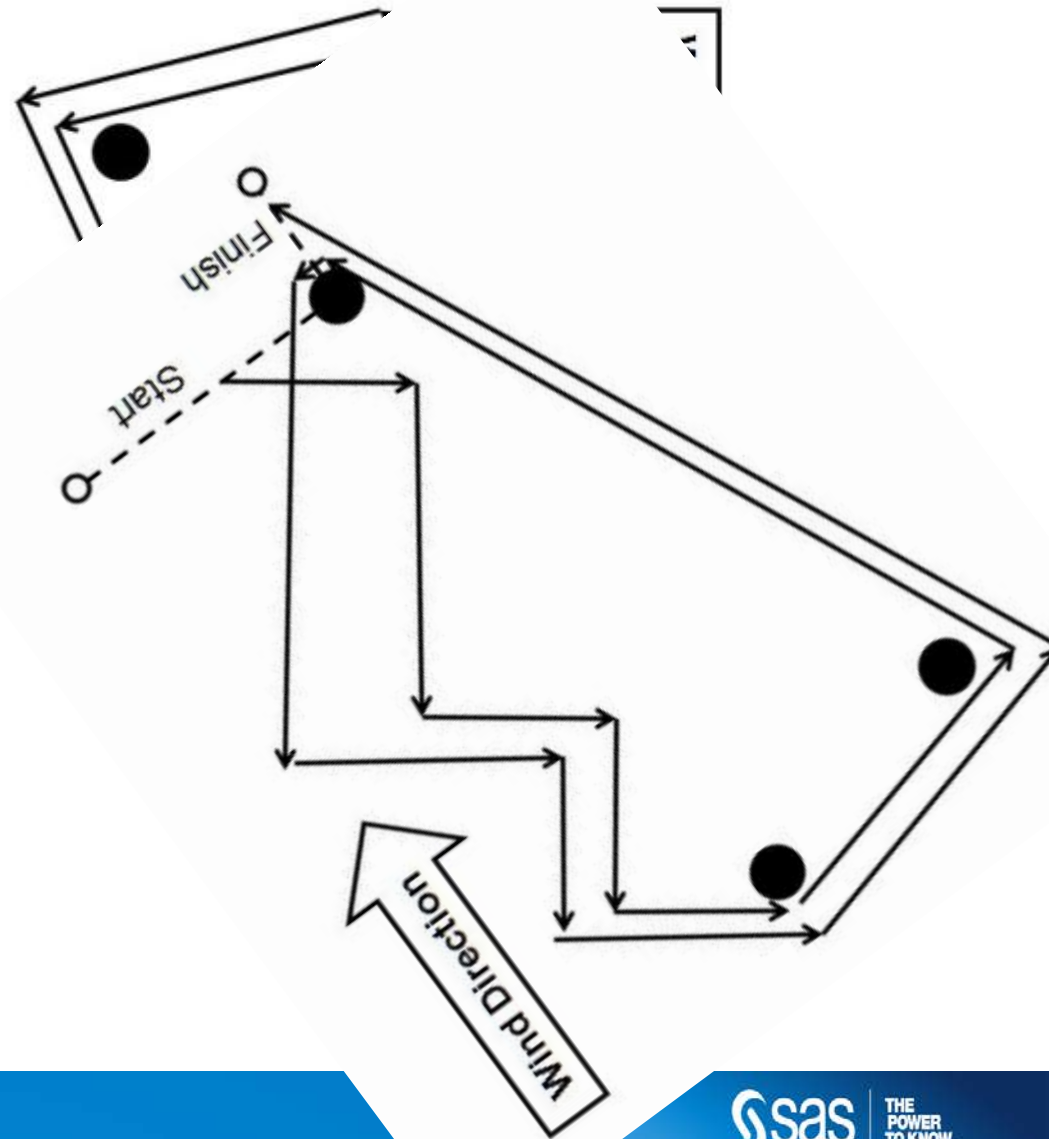
\*\*\* Run PROC GLM to calculate individual reference values;

```
proc glm data=labor_chol_data;
class sex centernr stage age_grp weight_grp ;
id visitdate;
model chol = age_grp sex weight_grp centernr stage;
output out=pred_chol p=reference r=residual
      stdi=stdi stdr=stdr stdp=stdp;
```

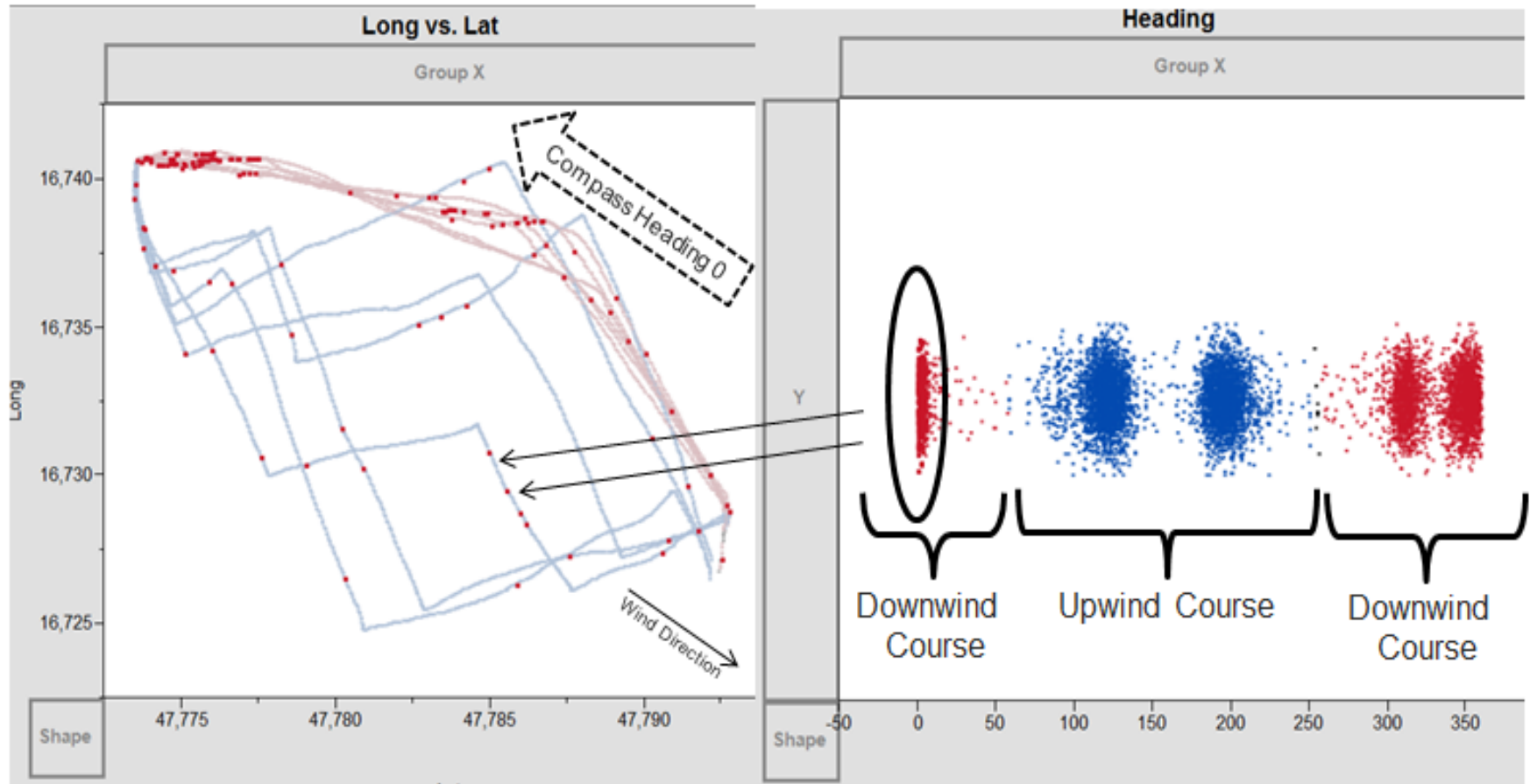
```
run;
quit;
```

ID  
0 + 0 x 11 Δ 1

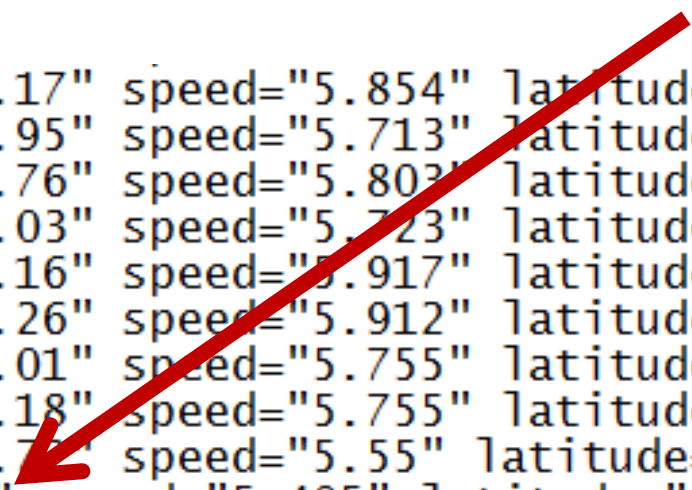
## SKIZZE EINER REGATTABAHN MIT 3 BOJEN.



## VISUELLE AUFDECKUNG VON FEHLERN IN DEN DATEN





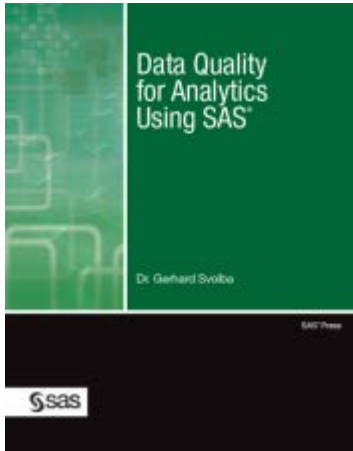


The image displays a list of data rows, each containing a timestamp, a heading value, a speed value, and a latitude value. A red arrow points to the row where the heading value is missing, indicated by a single quote character.

Timestamp	heading	speed	latitude
'2009-05-21T14:04:32+02:00"	heading="202.17"	speed="5.854"	latitude="47.
'2009-05-21T14:04:34+02:00"	heading="200.95"	speed="5.713"	latitude="47.
'2009-05-21T14:04:36+02:00"	heading="200.76"	speed="5.803"	latitude="47.
'2009-05-21T14:04:38+02:00"	heading="200.03"	speed="5.723"	latitude="47.
'2009-05-21T14:04:40+02:00"	heading="199.16"	speed="5.917"	latitude="47.
'2009-05-21T14:04:42+02:00"	heading="197.26"	speed="5.912"	latitude="47.
'2009-05-21T14:04:44+02:00"	heading="200.01"	speed="5.755"	latitude="47.
'2009-05-21T14:04:46+02:00"	heading="200.18"	speed="5.755"	latitude="47.
'2009-05-21T14:04:48+02:00"	heading="205.7"	speed="5.55"	latitude="47.7
'2009-05-21T14:04:50+02:00"	heading="198"	speed="5.405"	latitude="47.785
'2009-05-21T14:04:52+02:00"	heading="205.26"	speed="5.619"	latitude="47.
'2009-05-21T14:04:54+02:00"	heading="195.28"	speed="5.598"	latitude="47.
'2009-05-21T14:04:56+02:00"	heading="198.07"	speed="5.558"	latitude="47.
'2009-05-21T14:04:58+02:00"	heading="204.78"	speed="5.503"	latitude="47.
'2009-05-21T14:05:00+02:00"	heading="207.05"	speed="5.295"	latitude="47.
'2009-05-21T14:05:02+02:00"	heading="206.9"	speed="5.175"	latitude="47.7
'2009-05-21T14:05:04+02:00"	heading="210.27"	speed="5.721"	latitude="47.
'2009-05-21T14:05:06+02:00"	heading="204.1"	speed="5.468"	latitude="47.7
'2009-05-21T14:05:08+02:00"	heading="199.92"	speed="5.536"	latitude="47.
'2009-05-21T14:05:10+02:00"	heading="198.01"	speed="5.722"	latitude="47.

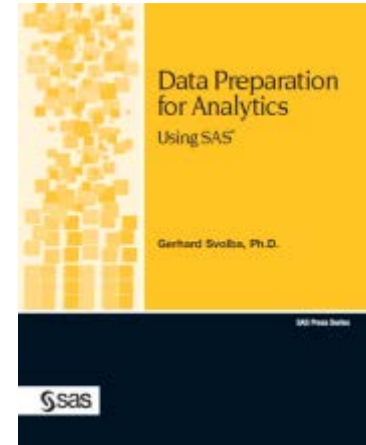
- „Datenqualität für Analytik“ ist mehr!
  - Mehr Anforderungen
  - Mehr Möglichkeiten
- SAS und JMP sind perfekt geeignet für das
  - Profiling,
  - Bewerten,
  - Verbessern,
  - Simulieren der Datenqualität.

## KONTAKT- INFORMATIONEN



### **Data Quality for Analytics Using SAS SAS Press 2012**

[http://www.sascommunity.org/wiki/Data\\_Quality\\_for\\_Analytics](http://www.sascommunity.org/wiki/Data_Quality_for_Analytics)



### **Data Preparation for Analytics Using SAS SAS Press 2006**

[http://www.sascommunity.org/wiki/Data\\_Preparation\\_for\\_Analytics](http://www.sascommunity.org/wiki/Data_Preparation_for_Analytics)



### **Gerhard Svolba**

Analytic Solution Architect

SAS-Austria

[Gerhard.svolba@sas.com](mailto:Gerhard.svolba@sas.com)

[http://www.sascommunity.org/wiki/Gerhard\\_Svolba](http://www.sascommunity.org/wiki/Gerhard_Svolba)

[LinkedIn](#)

[XING](#)



THE  
POWER  
TO KNOW<sup>®</sup>