SAS Tipps und Tricks:

Sie wollen rechtzeitig ein Bild über die Datenqualität Ihrer Analysedaten haben?

SAS und JMP helfen Ihnen dabei.

Dr. Mihai Paunescu

sas THE POWER TO KNOW.

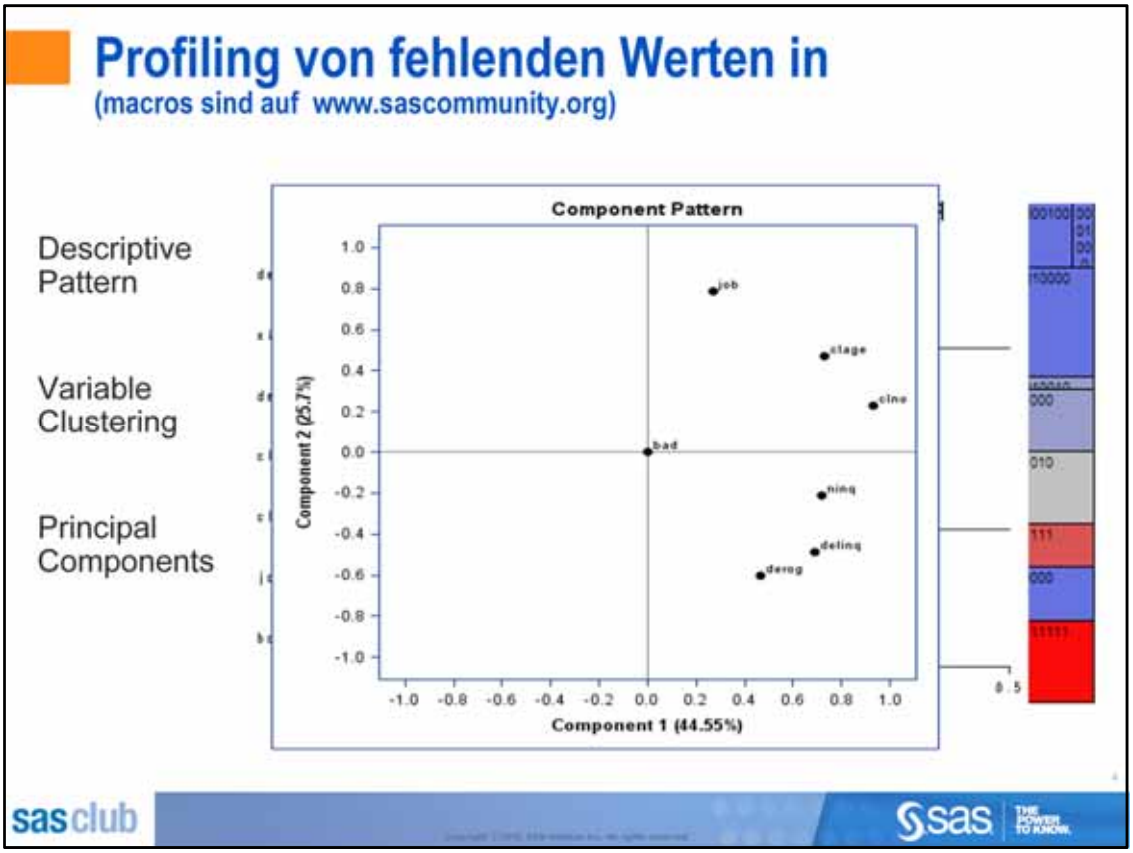
SAS unterstützt beim Verbessern der Datenqualität

- Erkennen von
  - Zusammenhängen zwischen fehlenden Werten
  - Ausreißern
- Behandlung von
  - fehlenden Werten durch individuelle Ersetzungswerte
  - Ausreißern
- Ähnlichkeits-Maße für Standardisierung und Record-Matching
- Methoden für seltene Ereignisse
- Stichprobenplanung

Profiling von fehlenden Werte in Querschnittsdaten

- Querschnittsdaten:

	id	NETINCOME	MWORK	EDUCATION	JOBPOSITION	COMPTYPE
1	1	15700	49	HIGH	OFFICER/WORK...	PUBLIC
2	2	26190.37	150	HIGH	OFFICER/WORK...	PUBLIC
3	3	28621.02	12	HIGH	OFFICER/WORK...	PRIVATE
4	4	16790	12	UNIVERSITY	MIDDLE	PUBLIC
5	5	18747.44	235	HIGH	OFFICER/WORK...	PUBLIC
6	6	31000	100	ACADEMY	OFFICER/WORK...	PUBLIC
7	7	34905	153	HIGH	SELFEMPLOYED	PRIVATE
8	8	25961.76	66	UNIVERSITY	MIDDLE	PUBLIC
9	9	35295.2	33	ACADEMY	OFFICER/WORK...	PUBLIC
10	10	59582.29	30	UNIVERSITY	OFFICER/WORK...	PRIVATE
11	11	17300	96	HIGH	MIDDLE	PRIVATE
12	12	21412	99	UNIVERSITY	SELFEMPLOYED	PRIVATE
13	13	36444	101	UNIVERSITY	MIDDLE	PUBLIC
14	14	25000	201	HIGH	TOP	PRIVATE
15	15	51965.1	110	HIGH	OFFICER/WORK...	PRIVATE
16	16	17230	129	HIGH	OFFICER/WORK...	PRIVATE
17	17	39000	291	ACADEMY	OFFICER/WORK...	PUBLIC
18	18	41183	291	HIGH	OFFICER/WORK...	PUBLIC
19	19	27500	51	ACADEMY	OTHER	OTHER
20	20	29956	346	UNIVERSITY	OFFICER/WORK...	PUBLIC



Variable Name
Description
BAD Defaulting or repaying the loan
DELINQ Number of delinquent trade lines
DEROG Number of major derogatory reports
NINC Number of recent credit inquires
CLAGE Age (in months) of the oldest trade line
CLNO Number of trade lines
JOB Current job, six categories

Ersetzen von fehlenden Werten mit dem SAS Enterprise Miner

Method	Description
Mean	Arithmetic mean of the non-missing observations
Median	Median of the non-missing observations
Distribution	Replacement values are calculated on the random percentiles of the variable's distribution of the non-missing values. This method does not change the distribution of the data very much.
Tree	A decision tree for each variable is run that predicts the imputation value based on other values in the dataset. The variables that are used as input variables for the tree can be selected. This method produces more accurate imputation values, but is also more time consuming.
Tree Surrogate	Same method as the TREE method. Additionally surrogate splitting rules are created for the case that an input variable is missing.
Midrange	(Minimum+Maximum)/2
Mid-Minimum	A certain proportion of the data is trimmed, by default the lower and upper 5 %. From the remaining data the MIDRANGE is calculated as imputation method.
Tukey's Biweight	Tukey's Biweight robust M-estimator
Huber	Huber's robust M-estimator
Andew's Wave	Andrew's Wave robust M-estimator
Default constant	Default constant to replace the missing values

sasclub

Profiling von fehlenden Werte in Zeitreihen

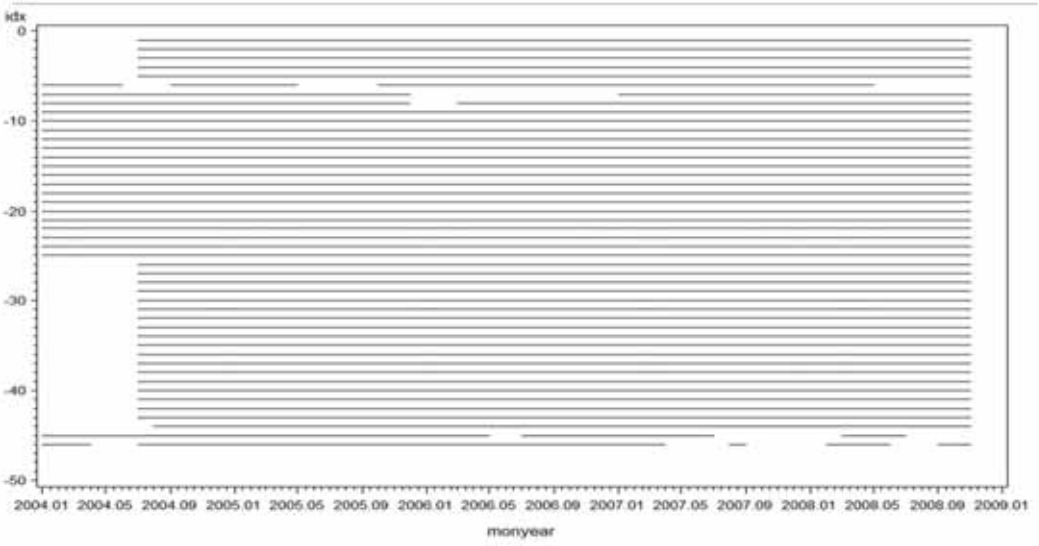
- Längsschnittdaten / Zeitreihen

	country_name	Jahr	GDP
549	Australia	1998	21442.833022
550	Australia	1999	20617.598688
551	Australia	2000	21768.042524
552	Australia	2001	19596.544212
553	Australia	2002	20214.306906
554	Australia	2003	23546.59072
555	Australia	2004	30569.074964
556	Australia	2005	34127.997291
557	Australia	2006	36202.533209
558	Australia	2007	40660.403928
559	Australia	2008	48348.261292
560	Australia	2009	42130.820325
561	Australia	2010	.
562	Austria	1960	935.39910695
563	Austria	1961	1031.7129439
564	Austria	1962	1087.8134937
565	Austria	1963	1167.6206439
566	Austria	1964	1270.9610375
567	Austria	1965	1377.5424695
568	Austria	1966	1489.830634
569	Austria	1967	1578.0092217
570	Austria	1968	1689.8431558
571	Austria	1969	1839.4905954
572	Austria	1970	2055.1117438
573	Austria	1971	2376.497732

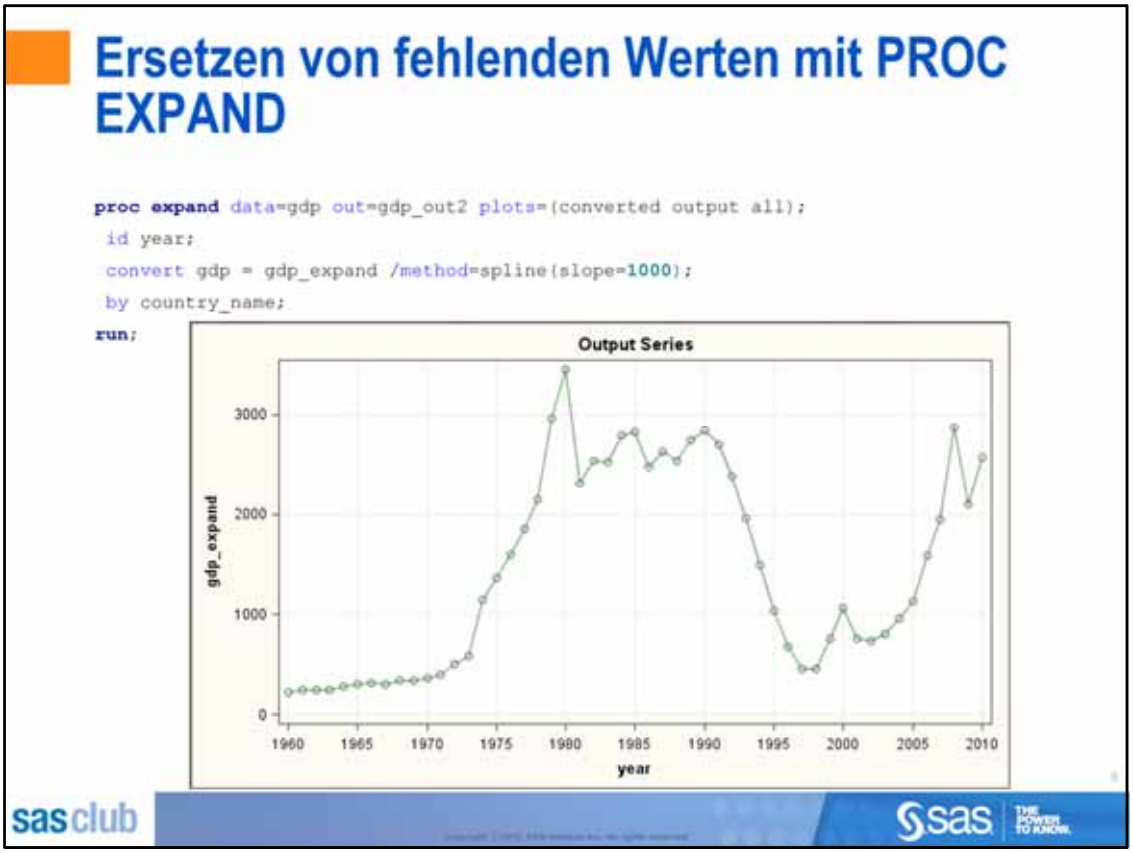
sasclub

Profiling von fehlenden Werten in Zeitreihen

(macros can be downloaded from www.sascommunity.org)

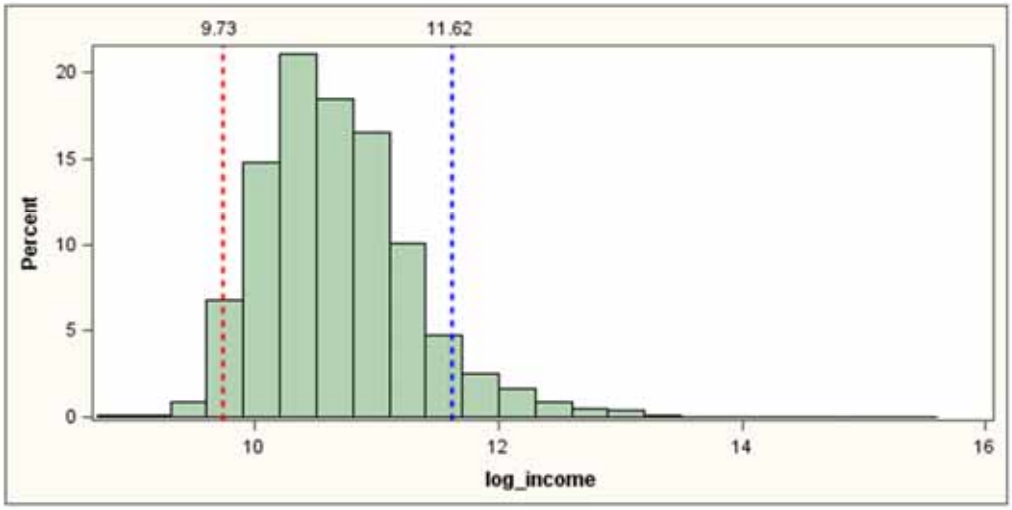


sasclub

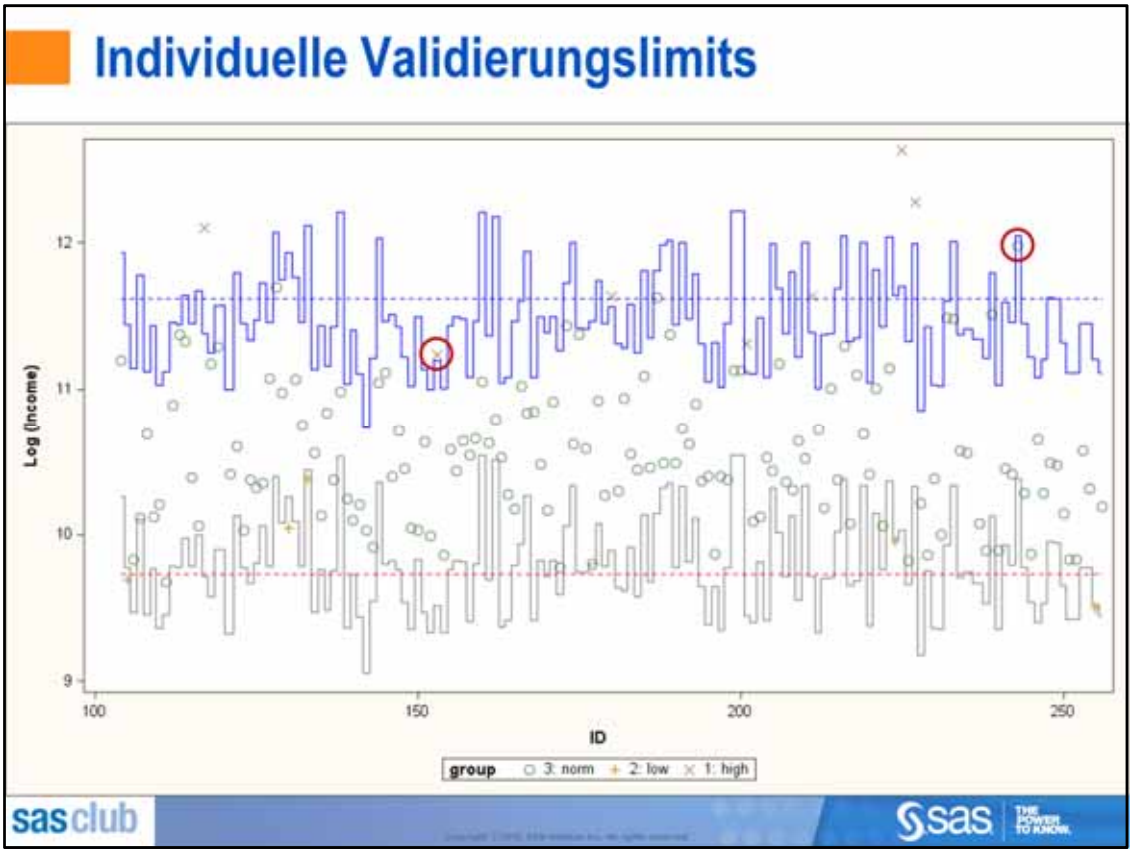



```
proc expand data=gdp out=gdp_out2 plots=(converted output all);
 id year;
 convert gdp = gdp_expand /method=spline(slope=1000);
 by country_name;
 where country_name in ('Afghanistan', 'Iraq', 'Iran, Islamic Rep.',
'Equatorial Guinea');
ruN;
```

Erkennen von Ausreißern in Querschnittsdaten



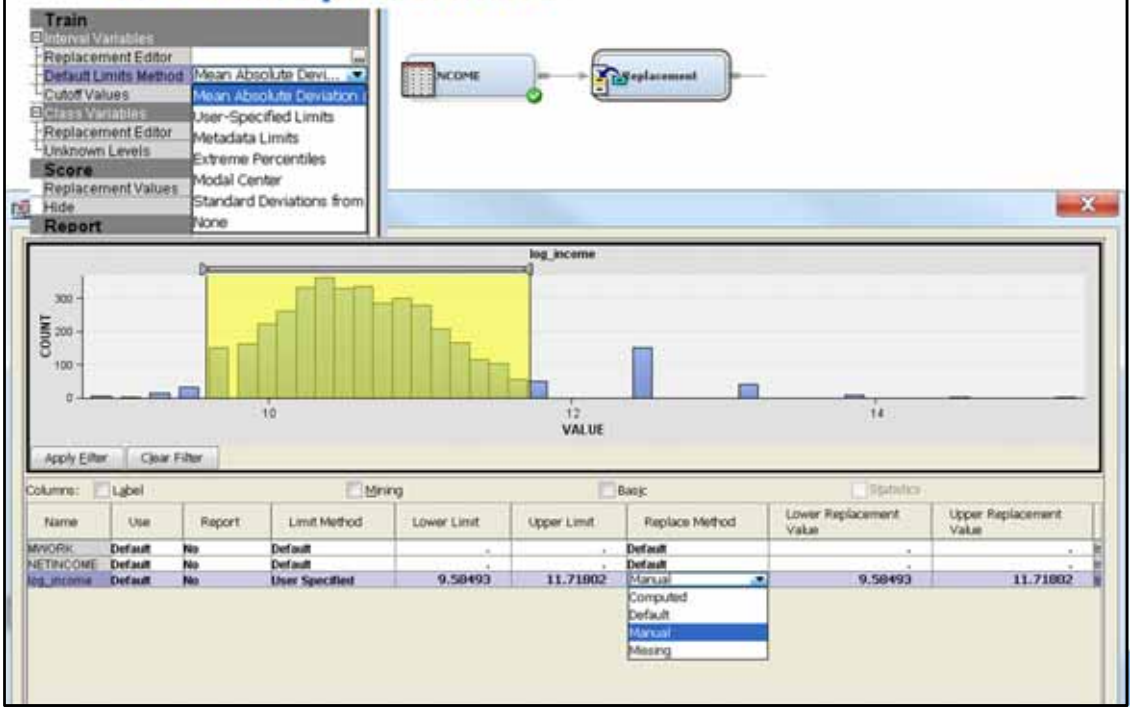
sasclub



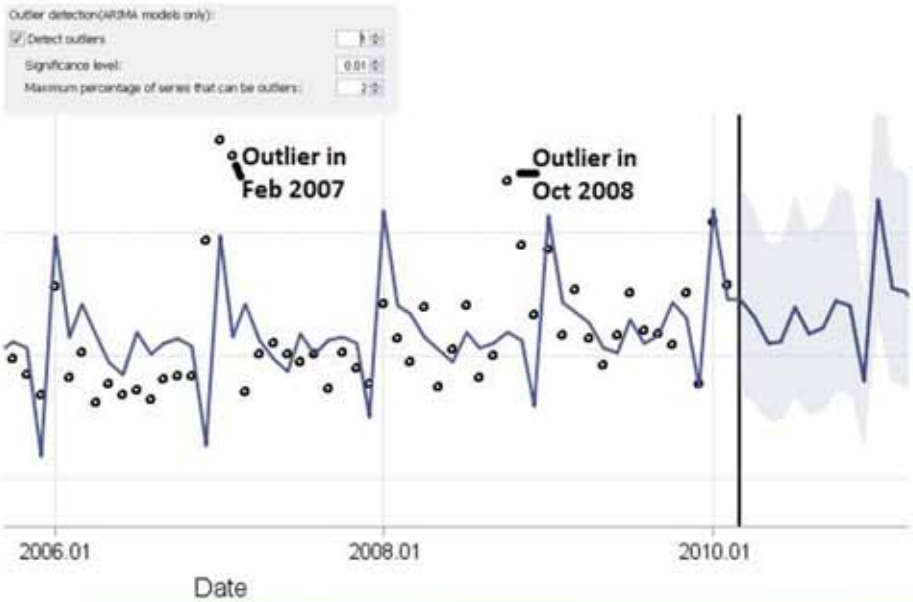

```
/*
                Modell erstellen für die Vorhersage des logarithmiertes
Einkommens
                in Abhängigkeit vom persönlichen Merkmalen
*/

proc glm data=tmp.income;
 class education jobposition comptype ;
 model log_income= MWork education  jobposition comptype /solution;
 where Netincome > 0 ;
 output out=pred_inc p=reference r=residual stdi=stdi;
run;


/* Ermitteln von allgemeinen Limits für Ausreißern */
proc sql noprint;
 select mean(log_income)+1.5*std(log_income) as upper_limit,
mean(log_income)-1.5*std(log_income) as lower_limit
 into :upper, :lower
 from pred_inc2;
quit;
```

```
/* Ermitteln von individuellen Limits für Ausreißern */
data pred_inc2(keep=id group residual reference log_income netincome upper:
lower: stdi);
 set pred_inc;
 ID=_N_;
 upper_i=reference+1.5*stdi;
 label upper_i='Log (Income)';
 lower_i=reference-1.5*stdi;
 if reference ne .;
 if log_income > upper_i then group='1: high';
 else if log_income < lower_i then group='2: low';
 else group='3: normal';
 upper=&upper;
 lower=&lower;
run;

/* Darstellung der Ausreißer für 150 Beobachtungen */

ods graphics on / height=600px width=1000px;
proc sgplot data=pred_inc2(firstobs=100 obs=250);
 scatter x=id y=log_income /group=group markerattrs=(size=9);
 step x=id y=upper_i /justify=center lineattrs=(color=blue);
 step x=id y=lower_i /justify=center lineattrs=(color=grey);
 series x=id y=upper / lineattrs=(color=blue pattern=2 );
 series x=id y=lower / lineattrs=(color=red pattern=2 );
run;
```

Behandlung von Ausreißern in SAS Enterprise Miner



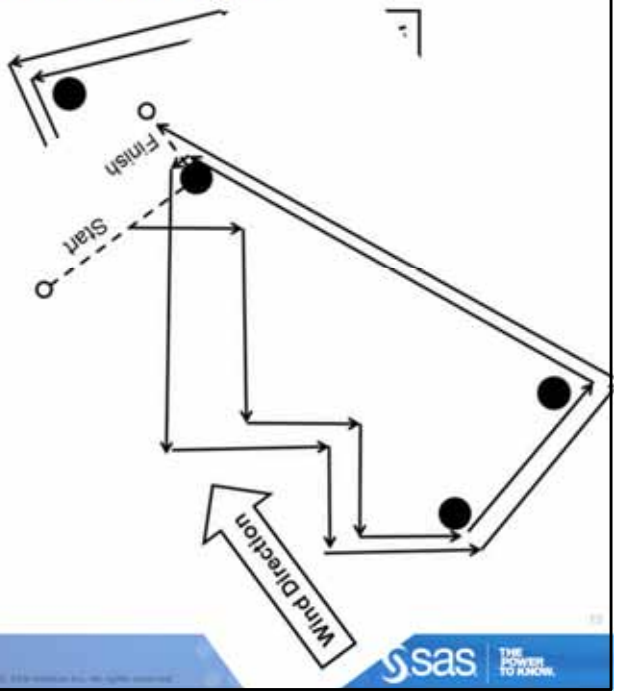
Erkennen von Ausreißern in Zeitreihendaten im SAS Forecast Server



sasclub

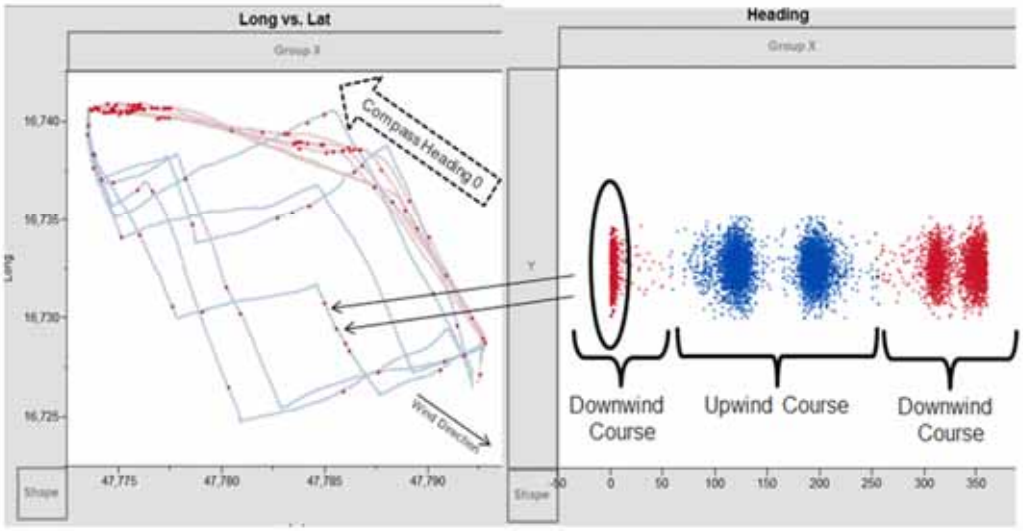
Graphisch interaktive Datenqualitätskontrolle mit JMP

Skizze einer Regattabahn mit 3 Bojen.



sasclub

Graphisch interaktive Datenqualitätskontrolle mit JMP



Graphisch interaktive Datenqualitätskontrolle mit JMP

- Aufdecken eines Fehlers beim Einlesen der Daten

```
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.
```

sasclub

sas THE POWER TO KNOW.