SAS1440-2015

Want an Early Picture of the Data Quality Status of Your Analysis Data? SAS® Visual Analytics Shows You How

Gerhard Svolba, SAS Institute Inc. Austria

ABSTRACT

When you are analyzing your data and building your models, you often find out that the data cannot be used in the intended way. Systematic patterns, incomplete data, and inconsistencies from a business point of view are often the reason. You wish you could get a complete picture of the quality status of your data much earlier in the analytic lifecycle. SAS® analytics tools like SAS® Visual Analytics help you to profile and visualize the quality status of your data in an easy and powerful way. In this session, you learn advanced methods for analytic data quality profiling. You see case studies based on real-life data, where we look at time series data from a bird's-eye view and interactively profile GPS trackpoint data from a sail race.

INTRODUCTION

Analysts often get stuck in the analysis process when they find out that they cannot use the data in the intended way. At first appearance, everything looks fine. The data quality checks do not reveal any problems with the data. But analytic methods often have additional requirements on data quality that go beyond simple data validation checks. Advanced methods to profile the quality of the data are needed to fulfill these requirements.

Analytic methods, however, not only pose additional requirements on the data. They also provide the following advanced methods to profile and improve the data quality of your data.

- Special reports for missing values exhibit systematic patterns in the data.
- Plots of principal components detect groups of variable that have the same missing value structure.
- Imputation methods find the most appropriate replacement value for missing data.
- Predictive models make it possible to analyze whether outliers or missing values occur most often for certain value combinations.

Analytic methods are not able to solve every data quality problem. You as an analyst should not just consider the data as a set of technical measurements. You must also consider the business processes that deal with the data and collect the data, since these factors have a strong impact on the interpretability of the data. Data quality checks must consider the underlying business assumptions.

Therefore, the people in a company or organization who understand the business background should get access to the data. The business experts should be able to verify their business assumptions by analyzing and drilling into the data.

This process requires an intuitive user interface that offers interactive access to the data. SAS® Visual Analytics provides a large set of data exploration options in a point-and-click style and thus, perfectly meets these requirements.

This paper covers several aspects of data quality profiling:

- How SAS and analytic methods can help to profile your analysis data before you invest a lot of time in the statistical analysis.
- Methods to detect systematic and random missing values in your cross-sectional data.
- How advanced profiling and visualization techniques can be used to show you the shape of your time series data in a single picture.
- Interactive profiling to spot inconsistencies in your data from a business point of view.

DATA QUALITY IS AN ANALYTIC TOPIC

GOOD TECHNICAL DATA QUALITY MIGHT NOT BE ENOUGH

You as an analyst often receive analysis data that is said to have good data quality. After taking a closer look, however, you find that you still cannot use the data as intended in your analysis. It does not meet the data quality requirements for your analysis.

From a technical perspective, as long as the data values correspond to value lists and validation limits and the percentage of missing values does not exceed a certain threshold, the data is considered to be of sufficient quality. Technical data quality also includes checks for dependencies in a relational schema. A record in the ACCOUNTS table, for example, can exist only if there is a corresponding record in the CUSTOMER table for the same customer ID (Svolba 2007, p. 46, and Svolba 2012, p. 160).

In the reporting world, it might be enough that data meet these technical requirements. In the analytic world, the requirements often go beyond these items. You as an analyst, statistician, data miner, or data scientist have to answer business questions with your analysis. Your requirements are influenced by the features of the business question that you have to answer and the analytical method that you use.

HISTORIC SNAPSHOTS, CORRELATION, AND SYSTEMATIC PATTERNS

When building analytic models, you might have already encountered one or more of the following data quality problems. These examples show that analytic methods need more than correct values.

Historic Data and Historic Snapshots of the Data Are Not the Same

Operational systems are usually not interested in "old news." Because they focus on preserving only the current version of the data, they do not typically save data values for each historic point in time. Consider the following two examples.

- A tariff change that takes effect when a mobile phone customer crosses over from tariff "Standard" to tariff "Advanced" is usually treated by the billing system as an update to the tariff value at the particular day. The goal of the billing system is to make sure that the price of the customer's services after that day are based on this new tariff. You as a marketing analyst have to analyze which tariff-change patterns frequently lead to certain customer events, such as product upgrades or cancellations. In order to perform these analyses, you need to be able to retrieve the actual value of the tariff for each customer at certain historic points in time.
- Assume you are in charge of forecasting the number of rented cars for the next four weeks for a car rental agency. In order to perform this analysis, you need not only the actual number of cars rented for each day but also the bookings that have already been received for a particular rental day. For the rental date April 29, 2015, the statistical model might use the following data:
  - Number of cars effectively rented on April 29th, 2015
  - Number of bookings for the rental date April 29, 2015, that are known as of April 28, 2015 (the day before)
  - Number of bookings for the rental date April 29, 2015, that are known as of April 27, 2015 (two days before)

These historic snapshots can be provided for the analysis only if data is historicized in a data warehouse. Typically, the operational system continuously overwrites the historic booking numbers for a particular rental day or updates the tariff change with the actual value.

Correlation between Variables Can Help but Might Cause a Headache

Headache: Multicollinearity

In a strict mathematical sense, the variables in the input matrix (or design matrix) in predictive models should be independent and exhibit no correlation among each other. Otherwise, the variables are considered to be multicollinear.

If you have multicollinear variables in your final model, the estimates of the regression parameters can be very unstable. Another consequence of multicollinear variables is “sign inversions” of the regression parameters. Here the sign (or direction) of a particular input parameter differs between the multivariate and the univariate model. This situation often leads to problems with the business interpretation of the results (Svolba 2012, p.98).

Note that multicollinear variables might occur not only because your source data already contains them. You often induce multicollinearity into your analysis data when you create several derived variables from the same source variable in the data preparation step. For example, if you create derived variables by aggregating the variable Duration of Calls in different ways, often the result is highly correlated variables.

Data with multicollinear variables can produce meaningful reports, but analytic methods can incur problems when using the same data. Technical data quality checks do usually not consider multicollinearity.

Benefit: Substitution Effects

When building an analytic model, the correlation between variables can also have advantages. Correlation might enable you to substitute the effect of variable A with variable B.

For example, the correlation between the variable Number_of_Customer_Visits and the variable Purchase_Amount implies that you can select the Number_of_Customer_Visits instead of the Purchase_Amount as a predictor variable in your model. You might choose this approach because the Purchase_Amount has too many missing values. You can use the correlation of the two variables to help you tell a little bit of the story even without the benefit of the missing values.

In many cases you benefit from correlation unintentionally. Some variables are assumed to be important for the analytical model. However, they are not available in your data.

The variable Customer_Age, for example, might not be available, because the Date_of_Birth field does not contain reliable values. Consequently, other variables such as Duration_of_the_Customer_Relationship (implication: long-term customers are usually older) and Product_Portfolio (implication: younger customers choose a different product mix) might stand in as the predictor variable for the unavailable Customer_Age variable. This substitution is possible because a correlation between these variables exists.

Simulation studies based on real data have shown that the removal of the most important variable in a predictive model usually does not reduce the overall predictive power by the amount that has been contributed by this variable. Other variables “jump in” and compensate the predictive contribution of the former top variable, to a certain extent (Svolba 2012, p. 227).

Systematic Patterns

For many analytical models, you must be able to rely on the randomness of the data. There should be no systematic pattern in the selection of the subjects of analysis, the occurrence of missing values, or the direction and the amount of bias in the data. Note that the prerequisite of complete randomness is rarely met. A small deviation, however, usually has a negligible effect on the systematic pattern.

In the next section, you meet my aunt Susanne and you learn more about systematic patterns, especially in the case of missing values.

Systematic patterns affect not only analytical models. Simple analyses such as descriptive reports are also influenced by systematic features in the data. The consequences for analytic methods, however, are often greater, especially if the results of these models are used as the basis for new strategies and decisions.

CONSEQUENCES OF BAD DATA QUALITY

The effects of bad data quality are manifold. If the quality of the underlying data is considered insufficient, analysis projects are often postponed, not started, or run only in part. This means that no results are available and important decisions are either not made or have to be postponed.

Bad data quality can also cause a lack of trust in the analysis results. Reduced confidence in the results can lead to a loss of confidence in your abilities as an analyst. At a higher level, this situation can also lead to a loss of brand image and can trigger regulatory fines or imprisonment. Regulations such as the Basel Accords or the Solvency Directives contain strict requirements for the handling and the provision of the analysis data.

If data with poor quality is used in our analytical models, important relationships are not identified and statistical significance is not reached. Data quality problems can result in fuzzy data and can hinder the ability of analytical models to reveal a true picture of the data. Worse yet, in more extreme situations, the picture might be seen as distorted and biased, rather than simply vague.

WHY MY AUNT SUSANNE GIVES ANALYSTS A HARD TIME

IS A SMALL PERCENTAGE OF MISSING VALUES IN THE AGE VARIABLE A PROBLEM?

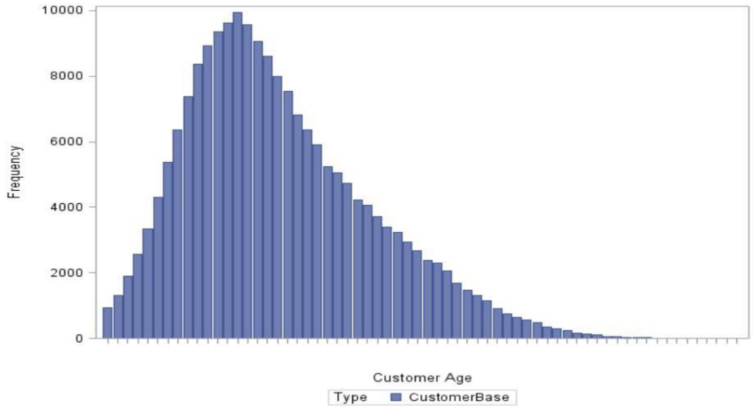
The Story of My Aunt Susanne

My aunt Susanne is an elderly lady, who lives at the countryside and looks forward to celebrating her 80th birthday soon. Since the 1960s she has had a telephone connection that was installed by her fixed-line provider. In the country where my aunt lives, at that time (and for many years later), you had to apply for a telephone contract and hope that you received one. This was long before topics such as "customer relationship management" or "customer care" became important. There was almost no personal data (date of birth and demographics, for example) collected during the application process, because there was no need for it. The most important details were the postal address of the telephone line, so that the provider could send the bill.

In the 1990s, topics such as "customer segmentation" and "know your customers" became more and more important to businesses, including Aunt Susanne’s phone provider. Since then, a customer must provide the date of birth with every new contract or contract change. My aunt, however, never changed or extended her phone contract. (She says, “A simple phone is enough!”) She never participated in customer surveys or marketing campaigns. Thus, no additional data was collected from her. And she is not the only one in this situation. In her circle of friends there are many people with a similar “data history.”

The Statistician in the Cubicle

If the statistician in the analysis department of Aunt Susanne’s phone provider now looks in the customer database and creates an analysis of customer age, he or she might see the picture as shown in Display 1.



Display 1. Histogram for the Age Variable

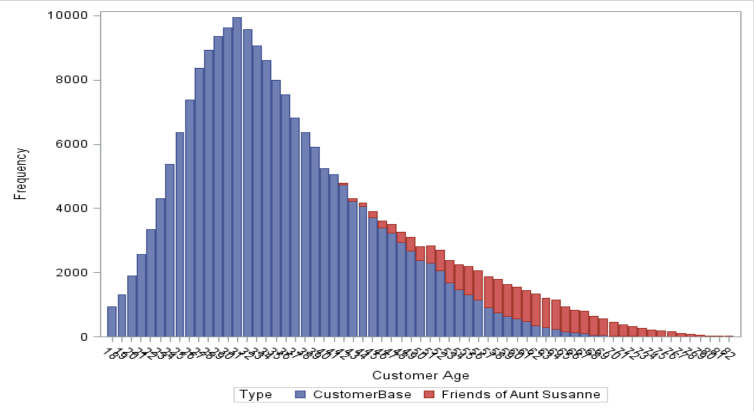
The age distribution by years shows how many customers are in which customer age groups. Based on that information, it is possible to define priorities for product bundles and selections for marketing campaigns. In addition, in this diagram, the statistician sees the proportion of missing values, where customer age could not be calculated because of a missing date of birth. In our case this proportion is 9.1 percent.

The statistician now must decide how to deal with the missing values.

- Shall a group "age unknown" be created?
- Shall the observations with missing values just be excluded from the analysis?
- Shall an average age of 42 years be assumed?
- Shall the imputation values be sampled from the true distribution?

The last three options assume implicitly that there is no pattern behind the fact that age is missing.

If, however, you now return to my aunt Susanne and her friends, you can assume that the missing values occur for customers in a higher age group, as shown in Display 2. After a certain year it was not possible to get a phone contract without providing the date of birth. Therefore, you can assume that the distribution of the missing age values does not cover the whole range of values, but is located at the right end of the distribution. The determination of an optimal replacement value for “age missing” has to consider this fact in the form of a business rule.



Display 2. Histogram for the Age Variable Including the True Values for the Missing Data

The red area in the histogram (shown in Display 2) is the "my aunt Susanne and her friends" group. In fact, they represent a specific customer segment: older, long-term customers, who do not show affinity for product upgrades or contract changes. Clearly, they should be treated differently in marketing actions! These customers probably have demands for specific hardware (phones with large keys and phones that are simple to use). Or they might need special assistance through the customer-care hotline.

Open Your Mind

What can you learn from this story? The data that you analyze has a history. It reflects not only the value that it measures, but it is also influenced by the business process and the type of data collection and data storage. To generate good results, it is mandatory that you look at the data not only from the statistical point of view.

You also have to observe the business background. For statistical analysis, you have to consider that things happen randomly only in very few cases. Think twice when you treat features in the data (such as missing values, outliers, and biases) as random. Consider whether you need to investigate the background and handle these features in your data on a case-by-case basis.

METHODS FOR DETECTING SYSTEMATIC MISSING VALUES

A Bird's-Eye View of the Missing Value Pattern

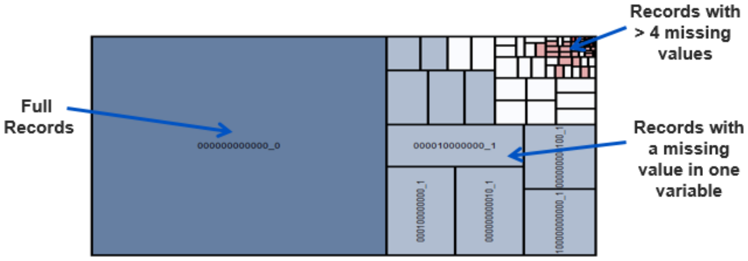
You will not be able to spot patterns of missing values as shown above, if you look only at reports that tabulate the number or percentage of missing values per variable. You gain more insight if you analyze the structure of missing values by using missing value pattern plots, as shown in Display 3.

An additional variable with values 0 and 1 is created for each analysis variable. This additional variable indicates whether the respective value is missing. These indicator values are then concatenated for each record. This action results in a string of 0s and 1s.

- A string of `00000000` shows a customer that has values for all variables.
- A string of `00101000` shows a customer that has a missing value in the 3rd and 5th variable.

Display 3 shows an illustrated example output.

- You see that around 60% of the records have no missing value. They are also called "full records."
- Another third of the records have a single missing value in one of the variables.
- A small proportion of records has missing values in four or more variables.



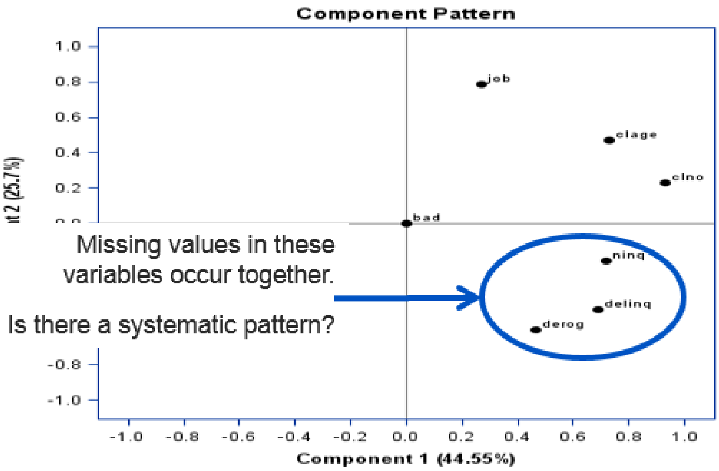
Display 3. Missing Value Pattern Plot as a Tile Chart

This picture gives you insight into the structure of the missing values. It tells you for which variables missing values occur together and how the missing values are distributed over records. You simply view your analysis data from the bird's perspective.

My aunt Susanne and her friends are represented as one segment in this chart. The missing values in the age variable most likely occur for records where the customer start date is also missing. Analyzing this segment in more detail, you might notice that customers with a missing age variable are subscribed at an old and basic tariff rate and have no traffic data for Internet connections.

In addition, you can use analytic methods to discover groups of variables that have missing values for the same records. Display 4 shows the results of a principal components plot. The missing value indicators that were created for the missing value patterns are used here as the analysis variable.

In the component plot, the variables that are close to each other have missing values for the same records. This arrangement enables you to detect groups of variables that are systematically missing together.



Missing values in these variables occur together.

Is there a systematic pattern?

Display 4. Missing Value Component Plot

The plots shown in Display 3 and Display 4 can easily be generated with the %MV_PROFILING macro. The macro can be downloaded (http://www.sascommunity.org/wiki/Data_Quality_for_Analytics_--_Download_Page). See the usage examples below. (For details, see Svolba, 2012, p. 127.)

Profile all variables from data set HMEQ:

```
%MV_Profiling(data=hmeq, vars = _all_)
```

Profile the variables JOB, REASON, DEBTINC, and VALUE from data set HMEQ:

```
%MV_Profiling(data=hmeq, vars = job reason debtinc value);
```

Sample the data for better performance to 10 %:

```
%MV_Profiling(data=hmeq, vars = _all_, sample=0.1);
```

Use only variables starting with “D.” Turn off the creation of the tile chart, the variable clustering, and the principal component analysis:

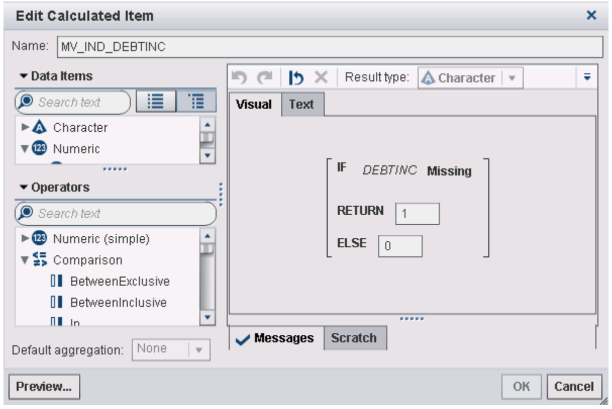
```
% MV_Profiling (data=hmeq, vars = d:,ods = NO, princomp=NO);
```

Using a Predictive Model to Explain the Missing Yes/No Indicator

Another method to detect a systematic pattern of missing values is to create an indicator variable Age_Missing YES/NO and then train a predictive model to detect whether there are differences between these two groups.

You might discover that in the previous example, the missing age values occur with old contract types or for customers who have a specific phone behavior. (Aunt Susanne is not making international phone calls or creating data traffic; she is just phoning her friends occasionally).

You can perform this task with SAS® Enterprise Miner™, with regression procedures in SAS/STAT® and (very conveniently) with SAS® Visual Analytics. Display 5 shows how to use Edit Calculated Item dialog box to simply create an indicator variable in SAS® Visual Analytics. In this example variable DEBTINC, the debt-income-ratio of the data set SAMPSIO.HMEQ is used.



Name: MV_IND_DEBTINC

Result type: Character

```
IF DEBTINC Missing
RETURN 1
ELSE 0
```

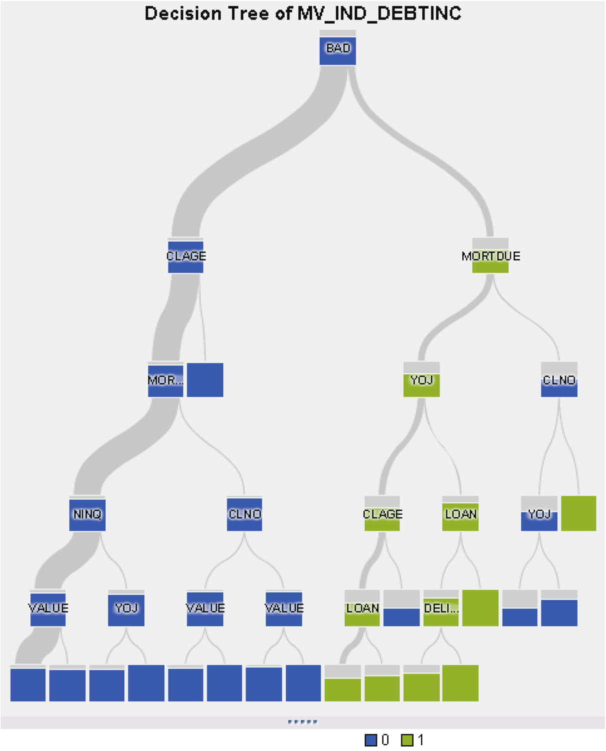
Default aggregation: None

Display 5. Creating a Missing Value Indicator in SAS Visual Analytics

In the next step, you create a decision tree in SAS Visual Analytics where the indicator variable is used as the target variable. The other variables in the data set are the input variables.

If the missing values in the DEBTINC variable occur randomly, none of the input variables have a meaningful association with the (target) indicator variable. Consequently, the decision tree does not create relevant splits.

If the missing values occur systematically for certain sub-segments, the splits of the decision tree describe the sub-segment. Display 6 shows a decision tree for the DEBTINC missing value indicator in SAS Visual Analytics.



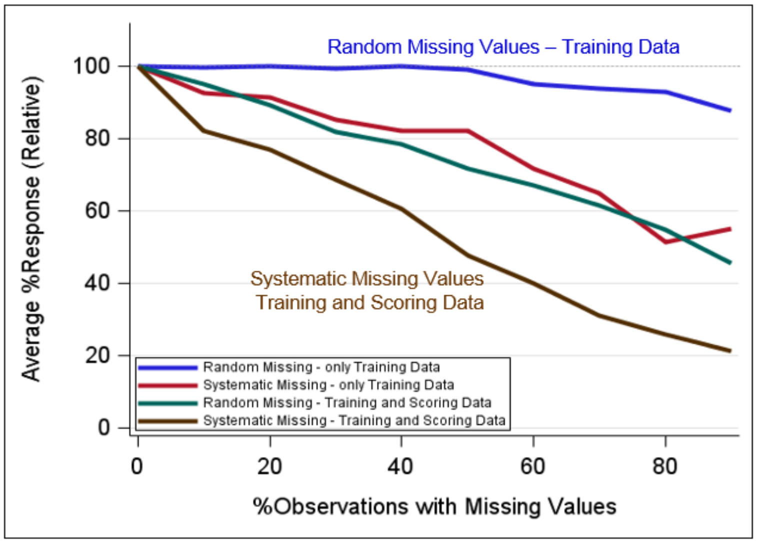
Display 6. Decision Tree for the DEBTINC Missing Value Indicator

You see that the tree has splits for a number of variables. Variable BAD is used for the first split and already differentiates between leaves with a large proportion of missing DEBTINC values on the right and a small proportion on the left. You can assume that these missing values do not occur randomly.

DOES QUANTITY MATTER? NOT ALWAYS.

The previous examples show that it is important to detect systematic patterns in the occurrence of missing values. Simulation studies have been performed to quantify the effect on predictive modeling of the presence of systematic missing values compared to random missing values (Svolba 2012, p. 207).

In these simulation scenarios, missing values have been inserted in the data. Then the missing values are imputed with the mean and the resulting data is used in a predictive model. Different proportions of missing values and different types of missing values, systematic and random, have been studied in these scenarios. The results are shown in Display 7.



Display 7. Line Chart Showing the Results from Simulation Studies for Missing Values

You see that an increasing proportion of missing values reduces the rate of correct predictions. However, you also see that there is a remarkable difference between the different types of missing values. The blue and green lines represent the scenarios with random missing values. Random missing values result in better model quality than the scenarios with systematic missing values (represented as red and brown lines).

How often have you been in situations where you discussed whether to impute values when 30% of the observations have missing values? These discussions usually concentrate on the sufficient number of nonmissing observations (“Do we have enough of them?”). However, the more important issue is that systematic missing value patterns have a much higher effect on model quality than random missing values.

Many people feel insecure about imputing values for the 30% of observations that have missing values, yet they feel comfortable imputing values for the 10% of observations that have missing values. A better approach is to investigate the reason for the missing values. Only then do you know whether to impute your missing values and if so, how to impute them.

PROFILING TIME SERIES DATA

WHAT’S SPECIAL ABOUT TIME SERIES DATA?

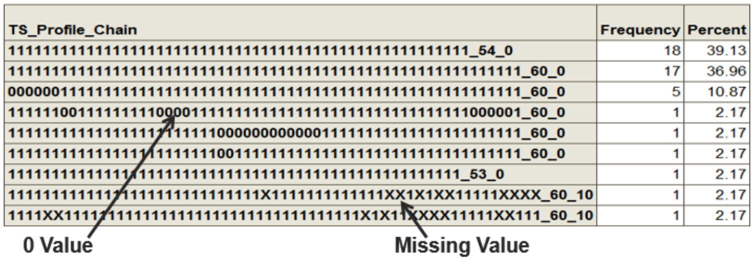
It is important to understand the structure of your time series data before you use it to build forecast models. Time series data sets usually have only a few variables and a very large number of records. It is often hard to get an overview of the quality status of your data.

- Due to the volume of data, it is almost impossible to browse through the data to see whether there are sections with an accumulation of zeros or missing values.

- Simply analyzing the total number of missing values per variable is not enough. You need this information broken out by time series.

You need analysis techniques that show you the profile of the time series in the analysis data set. When checking the data quality status, you are usually not interested in the course of the time series. You are interested in the distribution of the length of the time series, the occurrence and co-occurrence of zero values, missing values, and negative values.

Display 8 shows the representation of time series as a concatenation of 1 (value exists), 0 (zero value) and X (missing value).



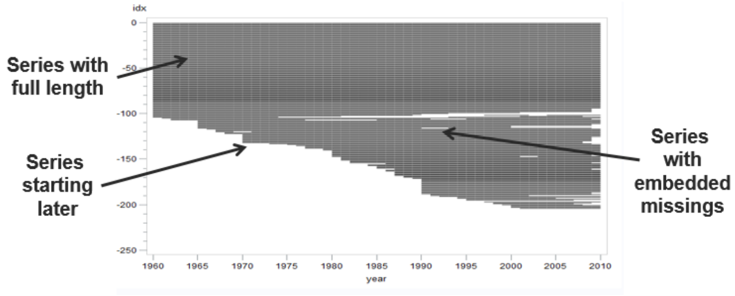
TS_Profile_Chain	Frequency	Percent
111111111111111111111111111111111111111111111111111111_54_0	18	39.13
111111111111111111111111111111111111111111111111111111111111_60_0	17	36.96
000000111111111111111111111111111111111111111111111111111111_60_0	5	10.87
111110011111111110000111111111111111111111111111111111000001_60_0	1	2.17
111111111111111111111111000000000000011111111111111111111111_60_0	1	2.17
111111111111111111111111001111111111111111111111111111111111_60_0	1	2.17
11111111111111111111111111111111111111111111111111111_53_0	1	2.17
111111111111111111111111111111X1111111111111XX1X1XX11111XXXX_60_10	1	2.17
1111XX11111111111111111111111111111111111X1X11XXXX11111XX111_60_10	1	2.17

Display 8. Time Series Profile-Chain in Tabular Form

This tabular representation gives you an overview over the structure and completeness patterns of your time series.

- You find 18 time series with a length of 54 time points that have no missing or zero values.
- There is another group of 5 time series with a length of 60 months. They, however, begin with a sequence of six zero values.

Display 9 shows a time series pattern plot for a different data set. This type of graphical representation allows you to detect the group of times series with full length, those that start later and those with embedded missing values.



Display 9. Line Chart in SAS/GRAPH® Showing the Pattern of the Time Series

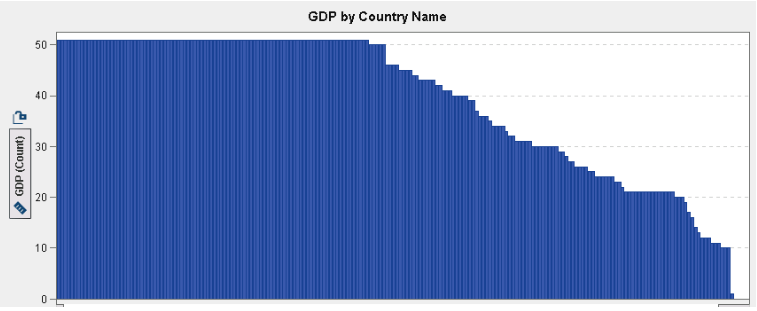
The output shown in Displays 8 and 9 can be created with the %PROFILE_TS_MV macro. The macro can be downloaded and is explained in more detail in Svolba 2012, p. 146. Here is an example invocation of this macro:

```
%Profile_TS_MV (data=ts.ts_demo46, id=tsid, date=year, value=GDP,
                w=0.1, plot=YES);
```

The variable GDP is then analyzed over the time variable YEAR for each time series ID (TSID). The macro orders the data so that similar time series appear next to each other. This arrangement creates a clearer picture.

USING SAS® VISUAL ANALYTICS TO PROFILE TIME SERIES DATA

The time series pattern plot can also be created in SAS Visual Analytics. You see the distribution of the time series lengths, if you create a bar-chart for the time series ID with a frequency count of the measurement variable GDP as the height. The resulting plot is shown in Display 10.



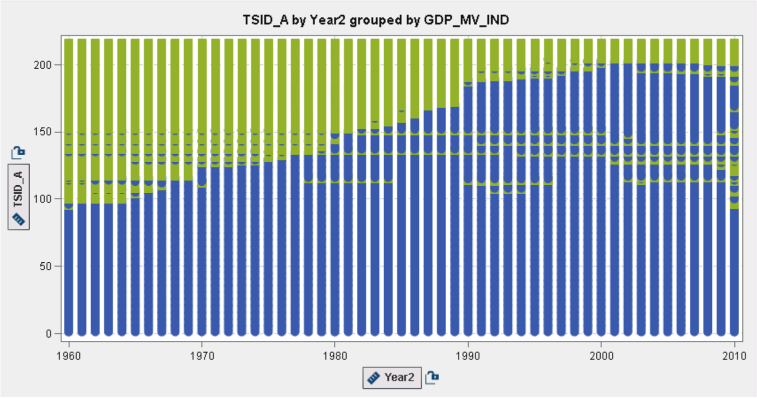
Display 10. Bar Chart Showing the Length of the Time Series

You see how many time series have full length. The chart, however, does not reveal whether the shorter time series start later, end earlier, or have embedded missing values.

You can also create a time series pattern plot as shown in Display 9 in SAS Visual Analytics.

- You simply create a missing value indicator for the measurement variable GDP in the same ways as shown above in Display 5. You call this variable GDP_MV_IND.
- Next, you create a scatter plot with the time ID variable YEAR on the x-axis and the time series ID (TSID) variable on the y-axis. Note that these variables have to be INTERVAL scaled variables. In this example the TSID is already pre-sorted by the similarity of the time series pattern to show a clear picture.
- Finally, you add the missing value indicator variable GDP_MV_IND as a group variable to the plot. Thus, you get different colors for the missing and nonmissing points. In some cases you might want to change the size of the dots in the scatter plot in order to improve the display of the chart.

The resulting plot can be seen in Display 11.



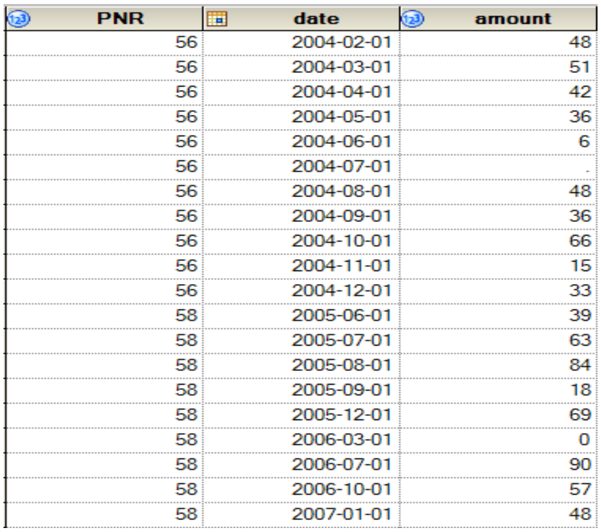
Display 11. Time Series Pattern Plot Created in SAS Visual Analytics

Note that it is very simple to perform these steps in SAS Visual Analytics. This method also works very well on big data. The plot helps you see the structure and pattern of your time series data and gives you good insight that you might not see, otherwise.

NO MISSING VALUES DOES NOT NECESSARILY MEAN COMPLETE DATA

A data set without missing values does not necessarily mean that you have complete time series data. Different to cross sectional data, time series data sometimes “hides” missing data. This situation occurs if, instead of having missing values for the variable, the whole record is missing (Svolba, 2014).

In Display 12 a monthly time series of purchase amounts for a customer is shown. In July 2004 you see a missing value for variable amount. What you do not see on the first look is that the records for the months Jan and Feb 2006 are missing as well. You detect these missing records only if you scan through the DATE variable.



PNR	date	amount
56	2004-02-01	48
56	2004-03-01	51
56	2004-04-01	42
56	2004-05-01	36
56	2004-06-01	6
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

Display 12. Table with Time Series Data and Missing Values

This continuity of the time series records is also referred to as “contiguity.” For mid-size and large data, you cannot manually check every single time series for contiguity. With PROC TIMESERIES, SAS provides a powerful method to detect the missing records and insert them into the data. The following syntax example shows you how to use PROC TIMESERIES for this task.

```
PROC TIMESERIES DATA = air_missing OUT = timeid_inserted;
  ID date INTERVAL =MONTH SETMISS=0;
  VAR qty;
  BY prod_id;
RUN;
```

From the ID statement, PROC TIMESERIES knows that the date variable should be on a monthly basis. Thus, it can check the sequence in which the months should appear. The SETMISS=0 option determines that for the inserted records the analysis variable QTY should be set to 0. Note that there are many other options available with PROC TIMESERIES. For more details refer to Svolba 2012, p. 149.

USING SAS VISUAL ANALYTICS TO FIND HIDDEN DATA QUALITY PROBLEMS

INTERACTIVE PROFILING OF ANALYTICAL DATA QUALITY

For analytical data quality profiling, SAS Visual Analytics offers two main benefits: its interactivity and its data visualization. The interactivity enables you to get your hands directly on the data. The data visualization functions make the access to your data and to the analysis methods very easy.

SAS Visual Analytics enables you to work directly on the data. You filter, subset, and group your data while you explore it. You are able to analyze value combinations from the business point of view and you can trace your findings directly to the source data.

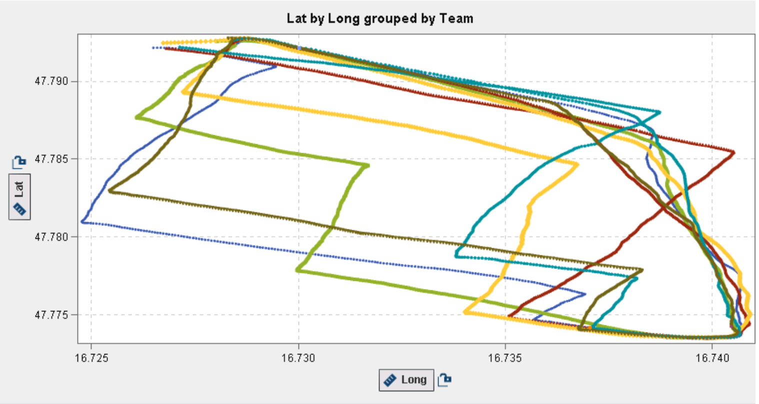
SAS Visual Analytics offers easy and visual access to your data and analyses. It enables different types of users to analyze and quality check the data. Applying data quality routines is now no longer limited to those who can program or use sophisticated tools. Business experts in their field find an intuitive access to their data. They are able to verify the data using their business expertise and run their individual plausibility checks.

USING DECISION TREES TO FIND THE REASON FOR DATA QUALITY PROBLEMS

Profiling GPS Track Point Data

The data in the following examples are taken from a sail race (Svolba, 2013). The participating boats use GPS tracking devices that display speed and compass heading. The device also stores the longitude and latitude coordinates, the compass heading, the speed in knots, and the timestamp in two-second intervals.

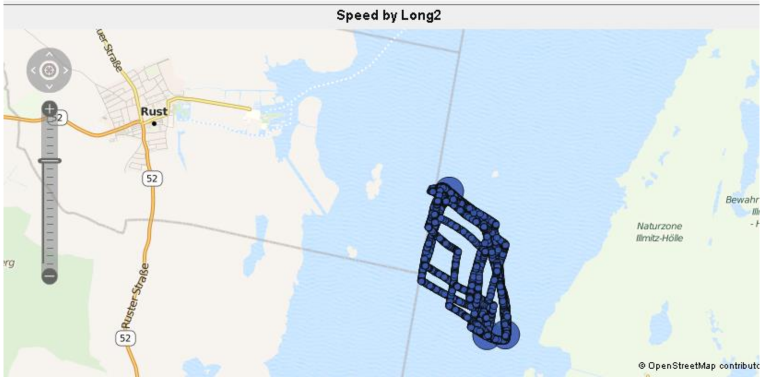
Display 13 shows the race course of the six participating boats. You can create such a chart easily in SAS Visual Analytics by dragging the longitude, latitude, and the team variable into the chart area. Visual Analytics automatically detects the chart type that should be created from the variable types.



Display 13. Race Course by Team as a Scatterplot of Longitude and Latitude

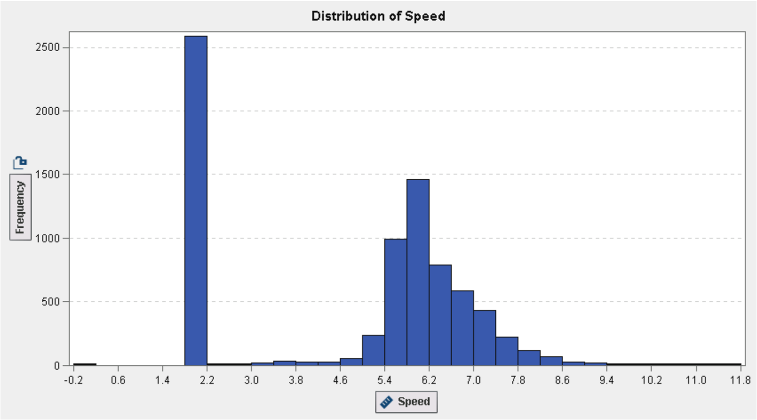
This chart shows the layout of the race area and the course of the six boats. The GPS coordinates seem to be collected without major deviations as the courses exhibit a smooth, non-erratic behavior.

Display 14 shows the course of the boats in a geo-map. This visual check also reveals that the GPS measurements seem to be of good quality. All data points are located in the blue area (the lake), which makes sense for a sail race.



Display 14. Race Course on a Map Chart

In the next step the individual variables are analyzed in more detail. A histogram for speed in knots is created. In Display 15 you see a reasonable distribution of speed between 4 and 9 knots, which makes sense for this type of sailboat. However, you also see a large accumulation of data points at 2 knots.



Display 15. Histogram of Speed in Knots

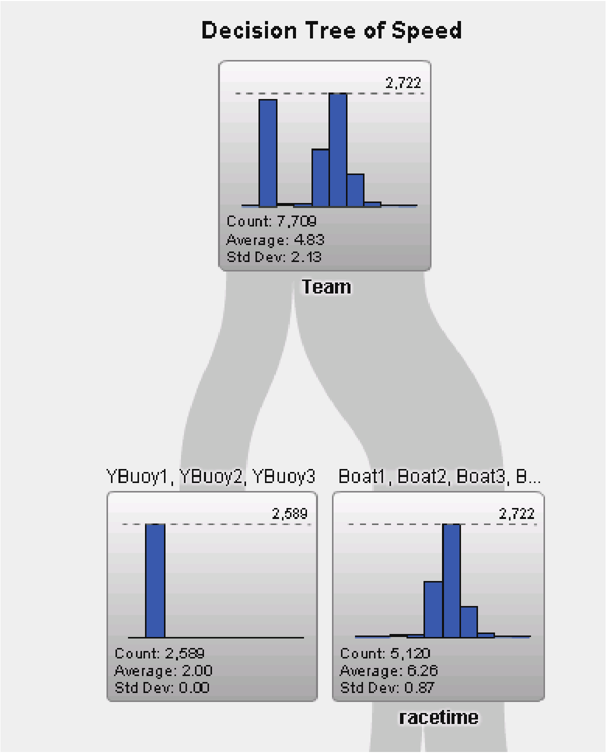
The distribution of speed in this graph raises doubts about whether the speed values are collected correctly. You want to analyze in detail why there is such an accumulation at 2 knots and whether this accumulation occurs only for certain subgroups.

Using a Decision Tree to Find the Reason

One option is to use the interactive facilities in SAS Visual Analytics and search manually for the source of the 2 knots values.

A more elegant and more effective method is to use the decision tree to automatically detect these groups. In the decision tree you use the variable Speed_in_Knots as a target variable and all other variables as input variables.

The resulting decision tree is shown in Display 16. You see that the first split is made for variable TEAM. The categories BUOY1 to BUOY3 are segmented in a separate leaf that all have a speed value of 2. The other categories BOAT1 to BOAT6 are segmented to a separate leaf. This reveals the source of the 2-knot values in a very straightforward way.



Display 16. Decision Tree for Speed Showing the Buoys in a Separate Leaf

You see that you can use the predictive modeling feature in SAS Visual Analytics to find the reason for the accumulation of 2-knot values. Obviously records for the buoys in the race course have been inserted into the data. Beside their longitude and latitude values, these records have been assigned an artificial value of 2 knots.

Most likely you would have been also able to spot this relationship by interactively analyzing your data. If you have a larger number of variables and more complex (hidden) relationships, you have to invest much more time to find them manually.

You do not have to explicitly ask and analyze every single question when you use decision trees for that purpose. Decision trees enable you to pose the overall question and get back the findings.

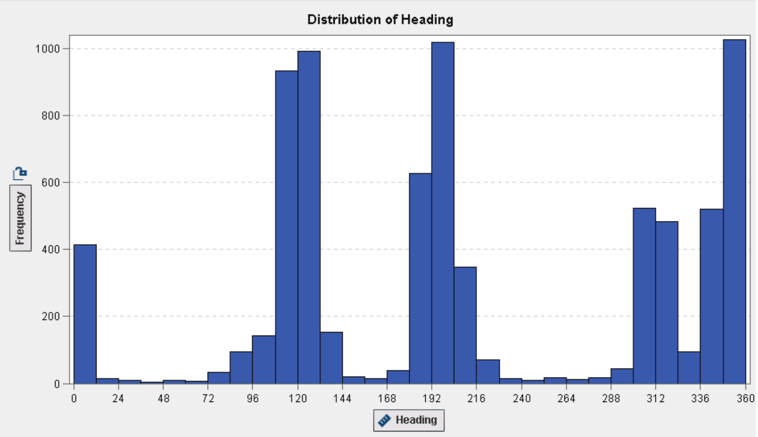
USING INTERACTIVE DATA ANALYSIS TO UNCOVER IMPOSSIBLE VALUE COMBINATIONS

Univariate Analysis of the Compass Heading

In Display 17 you see the distribution of the compass heading values. A first analysis reveals that the values range from 0° to 360°. This range makes sense from a business perspective.

The true wind direction in the race was around 155° (SSE). Consequently, you also see an accumulation of values around 110° and 200°. This represents the upwind course where the boats sail in an approximate 45° angle to the true wind direction.

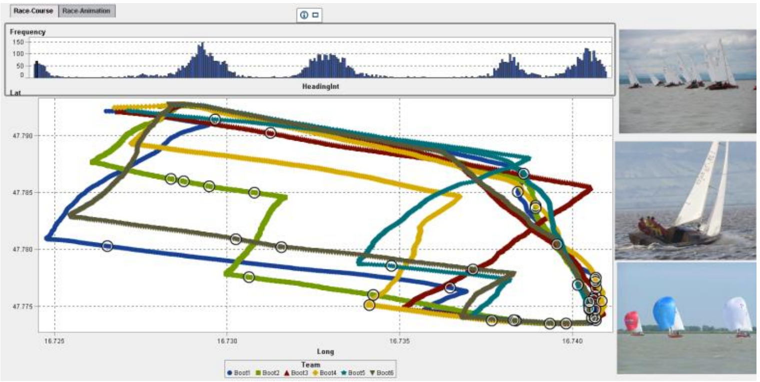
Another accumulation of values can be seen around 310° and 360°. This represents the downwind courses. Looking at this graph, everything makes sense from a business perspective.



Display 17. Histogram of the Compass Heading

Business Insight with Interactive Data Analysis

You can also interactively link the histogram for compass heading with the race course. The diagram in Display 18 is a combination of the charts shown in Displays 13 and 17. The link between the two charts makes it possible to select values in the compass heading histogram and see the respective points highlighted in the race course plot.



Display 18. Interactive Plot of Compass Heading and the Race-Course

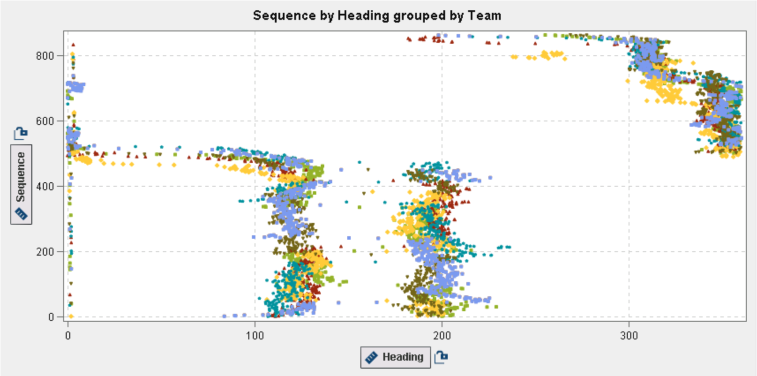
If you select the bars around compass course 110°, the upwind paths heading to lower right of the display are highlighted. If you select the compass courses around 200°, the upwind paths heading to the lower left are highlighted.

The other two peaks in the histogram represent the downwind paths. The compass headings of 2° are selected and highlighted in Display 18. It is surprising that you see highlighted data points not only in the downwind section where boats are sailing northwards. You also see a few highlighted data points in the upwind section.

This data seems to imply that the boat is sailing upwind with a compass heading around 200°, then it turns for two seconds to the compass heading around 2°, and then it immediately turns back. This course is not physically possible for a sailboat. Rather, we have to assume that there is a data quality problem. Perhaps the data reported by the GPS device is biased.

Using a Sequence Plot

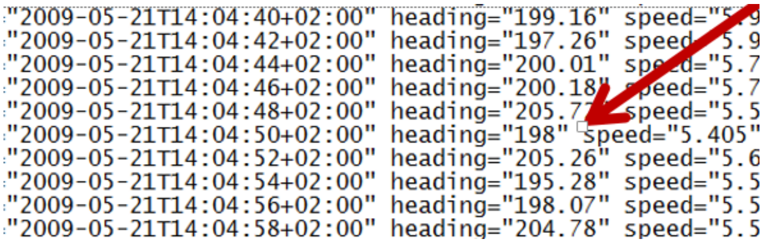
When looking at the data in a sequence plot, you see a similar picture. The compass heading is plotted on the x-axis and the sequence of observations is plotted on the y-axis in Display 19. The values for different boats are shown in different colors. The first 400 data points belong to the upwind course. You see that there are a few implausible values in the interval (0, 3). From the colors you detect that the implausible values do not only occur for a particular boat, but for all boats.



Display 19. Dot-Plot for Compass Heading over the Data Collection Sequence

A closer data investigation reveals that in sections where most of the compass heading values are around 200°, a few values around 2° also occur. Tracing these cases back to the individual data record enables you to identify the respective lines in the source data.

In Display 20 you see that this problem occurs only if the compass heading appears as an integer value in the source data. If the true compass heading has two zeros (.00) after the decimal point, the GPS device outputs integer values. The trailing zeros are obviously truncated from the compass heading values.



```
"2009-05-21T14:04:40+02:00" heading="199.16" speed="5.9
"2009-05-21T14:04:42+02:00" heading="197.26" speed="5.9
"2009-05-21T14:04:44+02:00" heading="200.01" speed="5.7
"2009-05-21T14:04:46+02:00" heading="200.18" speed="5.7
"2009-05-21T14:04:48+02:00" heading="205.77" speed="5.5
"2009-05-21T14:04:50+02:00" heading="198" speed="5.405"
"2009-05-21T14:04:52+02:00" heading="205.26" speed="5.6
"2009-05-21T14:04:54+02:00" heading="195.28" speed="5.5
"2009-05-21T14:04:56+02:00" heading="198.07" speed="5.5
"2009-05-21T14:04:58+02:00" heading="204.78" speed="5.5
```

Display 20. Source Data from the GPS Tracking Device

The data integration program that reads this data into a SAS data set does not consider such a situation and forces 2 digits to be added to the right of the decimal point. Therefore, the compass heading values of 198° are shifted to 1.98°. This leads to the strange picture of compass heading values close to 0°, even if the boat sails in the opposite direction.

Such cases remain undetected if you just profile your data with simple cross tables or univariate charts. You need interactive profiling to spot these situations.

CONCLUSION

It is important to differentiate between regular data quality and data quality for analytics. Analytic methods have additional requirements on data quality. But they also offer methods to profile and improve data quality. SAS analytic procedures and SAS Visual Analytics offer a rich set of methods to get insight into the quality status of your data.

SAS macros and SAS sample programs help you to profile your data in a very powerful way. It is very important to find out whether there are systematic patterns in your data, such as the systematic patterns that can occur in the case of missing values. Simulation studies give important insight into the consequences of using poor data quality for predictive analytics.

SAS® Visual Analytics offers powerful methods to interactively profile your data. It allows you to get closer to the data values to find out where inconsistencies or strange value combinations occur.

REFERENCES

Svolba, Gerhard. 2007. *Data Preparation for Analytics Using SAS®*. Cary, NC: SAS Institute Inc. Available http://www.sascommunity.org/wiki/Data_Preparation_for_Analytics.

Svolba, Gerhard. 2012. *Data Quality for Analytics Using SAS®*. Cary, NC: SAS Institute Inc. Available http://www.sascommunity.org/wiki/Data_Quality_for_Analytics.

Svolba, Gerhard. 2013. “Is your data ready for analytics?” *IT Briefcase*. Available http://www.itbriefcase.net/is-your-data-ready-for-analytics.

Svolba, Gerhard. 2014. “Missing Values.” *Analytics*. Available http://viewer.zmags.com/publication/6af8b2ed#/6af8b2ed/58.

APPENDIX

The following data quality profiling macros are available for you:

- %COUNT_MV: Shows the frequency and percentage of missing values for both numeric and character variables.
- %MV_PROFILING: Shows missing value patterns for cross-sectional data in a tile chart and analyze the structure of missing values in a principal components plot and variable clustering tree plot.
- %PROFILE_TS_MV: Shows the missing value patterns for time series data in a table and in a line plot (bird’s-eye view).
- %CHECK_TIMEID: Checks the continuity of time series data and reports the missing records.

The macros can be downloaded here: http://www.sascommunity.org/wiki/Data_Quality_for_Analytics_--_Download_Page.

ACKNOWLEDGMENTS

Many people have helped and inspired me to write and to complete this paper: Udo Sglavo, Sascha Schubert, Mike Gilliland, Anne Milley, Rainer Sternecker, Diane Hatcher, Anette Almer, Anne Baxter, and Robin Langford.

RECOMMENDED READING

- Svolba, Gerhard. 2006. “Efficient ‘One-Row-per-Subject’ Data Mart Construction for Data Mining.” *Proceedings of the Thirty-First Annual SAS Users Group International Conference*. Cary, NC. SAS Institute Inc. Available http://www2.sas.com/proceedings/sugi31/078-31.pdf.
- Svolba, Gerhard. 2013. “’The hungry statistician’ or why we never can get enough data.” *Subconscious Musings*. SAS blog. Available http://blogs.sas.com/content/subconsciousmusings/2013/12/03/the-hungry-statistician-or-why-we-never-can-get-enough-data/.
- Svolba, Gerhard. 2013. “Why my aunt Susanne and her friends give us a hard time in statistical analysis.” *Subconscious Musings.* SAS blog. Available http://blogs.sas.com/content/subconsciousmusings/2013/10/13/why-my-aunt-susanne-and-her-friends-give-us-a-hard-time-in-statistical-analysis/.
- SAS Institute Inc. 2014. *SAS® Visual Analytics 7.1: Users Guide*. Cary, NC: SAS Institute Inc. Available http://support.sas.com/documentation/onlinedoc/va/.
- SAS Institute Inc. *SAS/ETS® 13.2: Users Guide - Procedures*. Cary, NC: SAS Institute Inc. Available http://support.sas.com/documentation/onlinedoc/ets/index.html.
- Svolba, Gerhard. 2014. *Data Preparation for Analytics and Data Quality for Analytics on the Road*. SAS picture blog. Available http://www.sascommunity.org/wiki/%22Data_Preparation_for_Analytics%22_and_%22Data_Quality_for_Analytics%22_on_the_Road.
- Barkaway, David. 2010. “dATa qWaliti 4 Analytics.” *Proceedings of the SAS Global Forum 2010 Conference.* Available http://support.sas.com/resources/papers/proceedings10/328-2010.pdf.
- SAS Institute Inc. 2012. Changes and enhancements for the book *Data Quality for Analytics Using SAS.* Cary, NC: SAS Institute Inc. Available http://www.sascommunity.org/mwiki/images/8/89/DQFA_Enhancements_and_Corrections.pdf
- SAS Institute Inc. 2014. Corrections for the book *Data Preparation for Analytics Using SAS.* Cary, NC: SAS Institute Inc. Available http://www.sascommunity.org/mwiki/images/f/fb/Enhancements_and_Corrections_A60502_DataPreparationForAnalytics.pdf

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Gerhard Svolba
SAS Institute Inc. Austria.
Mariahilfer Strasse 116, A-1070 Wien
Email: mailto: Sastools.by.gerhard@gmx.net
Web: http://www.sascommunity.org/wiki/Gerhard_Svolba