Part I: Data Quality Defined

Introduction

General

The first part of this book focuses on the definition of data quality and the data quality characteristics that are important from an analytical point of view.

The first two chapters of this part extend the introduction by using example case studies and a definition of data quality for analytics.

- Chapter 1, "**Introductory Case Studies**," relates real-life examples to typical data quality problems, forming an example-oriented introduction to data quality for analytics.
- Chapter 2, "**Definition and Scope of Data Quality for Analytics**," defines data quality for analytics, discusses its importance, and provides examples of good data quality.

The next seven chapters discuss data quality characteristics that are at the heart of data quality for analytics:

- Chapter 3, "**Data Availability**," questions whether the data are available. Can the data needed for the analysis be obtained?
- Chapter 4, "**Data Quantity**," examines whether the amount of data are sufficient for the analysis.
- Chapter 5, "**Data Completeness**," deals with missing values for the available data fields from a data analysis perspective.
- Chapter 6, "**Data Correctness**," discusses whether the available data are correct with respect to their definition. Are the data what they claim to be and do they, in fact, measure what they are supposed to measure?
- Chapter 7, "**Predictive Modeling**," discusses special requirements of predictive modeling methods.
- Chapter 8, "**Analytics for Data Quality**," shows additional requirements of interdependences for analytical methods and the data.
- Chapter 9, "**Process Considerations for Data Quality**," shows the process aspect of data quality and also discusses considerations such as data relevancy and possible alternatives.

These chapters form the conceptual basis of the book (that is, the relevant features of data quality for analytics). The second part of the book uses this as a basis to show how the data quality status can be profiled and improved with SAS.

Chapter 1: Introductory Case Studies

1.1 Introduction

This chapter introduces data quality for analytics from a practical point of view. It gives examples from real-world situations to illustrate features, dependencies, problems, and consequences of data quality for data analysis.

Not all case studies are taken from the business world. Data quality for analytics goes beyond typical business or research analyses and is important for a broad spectrum of analyses.

This chapter includes the following case studies:

- In the first case study, the performance of race boats in sailing regattas is analyzed. During a sailing regatta, many decisions need to be made, and crews that want to improve their performance must collect data to analyze hypotheses and make inferences. For example, can performance be improved by adjusting the sail trim? Which specific route on the course should they sail? On the basis of GPS track point and other data, perhaps these questions can be answered, and a basis for better in-race decisions can be created.
- The second case study is taken from the medical research area. In a clinical trial, the performance of two treatments for melanoma patients is compared. The case study describes data quality considerations for the trial, starting from the randomization of the patients into the trial groups through the data collection to the evaluation of the trial.
- The last case study is from the demand forecasting area. A retail company wants to forecast future product sales based on historic data. In this case study, data quality features for time series analysis, forecasting, and data mining as well as report generation are discussed.

These case studies illustrate data quality issues across different data analysis examples. If the respective analytical methods and the steps for data preparation are not needed for the data quality context, they are not discussed.

Each case study is presented in a structured way, using the following six subsections:

- Short overview
- Description of the functional question and the domain-specific environment
- Discussion of practical questions of interest
- Description of the technical and data background
- Discussion of the data quality considerations
- Conclusion

1.2 Case Study 1: Performance of Race Boats in Sailing Regattas

Overview

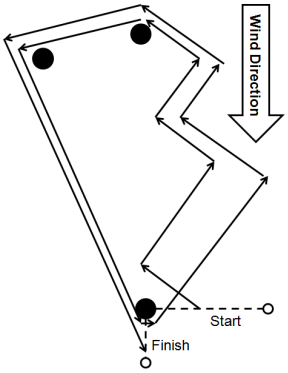
This case study explores a comprehensive data analysis example from the sailing sport area. Note that these characteristics of data quality not only apply to sailboat analysis, but they also refer to research- and business-related analysis questions. For a specific race boat, the GPS (global positioning system) track point data over different races and the base data (like the size of sails, crew members, and external factors) are collected for one sailing season. These data are then cleaned, combined, and analyzed. The purpose of the analysis is to improve the race performance of the boat by answering questions like the influence of wind and choice of sails or the effect of different tactical decisions.

Functional problem description

The name of the boat of interest is *Wanda*, and the team consists of a helmsman and two crew members. The boat participates in sailboat fleet races, where 10–30 boats compete against each other in 5–10 regattas per sailing season, and each regatta consists of 4–8 races. The race course is primarily a triangle or an "up-and-down" course, where the "up" and the "down" identify whether it is sailed against or with the wind.

The typical race begins with a common start of all participating boats at a predefined time. After passing the starting line, the boats sail upwind to the first buoy, then in most cases they go downwind to one or two other buoy(s), and then upwind again. This route is repeated two to three times until the finishing line is passed. Figure 1.1 illustrates an example race course.

Figure 1.1: Typical course in a sailboat regatta



Sailing is a complex sport. In addition to the optimal sailing technique, the state of the sailing equipment, and the collaboration and physical fitness of the crew, many factors have to be considered to sail a good race. The most important factors are listed here:

- When going upwind, sailboats can sail at an angle of about 45 degrees with the true wind. To reach the upwind buoys, the boats must make one or more tacks (turns). The larger the angle to the wind, the faster the boats sail; however, the distance that has to be sailed increases.
- Depending on the frequency and the size of wind shifts, it might be better to do more tacking (changing the direction when going upwind) to sail the shortest possible course. However, tacking takes time and decreases speed.
- The specific upwind course of a boat is typically planned to utilize the wind shifts to sail upwind as directly as possible.
- The sailboat itself offers many different settings: different sail sizes and different ways to trim the boat. An average race boat has about 20 trim functions to set (for example, changing the angle and shape of the sails).

There is much literature available on sailboat race tactics and sailboat trimming. To successfully compete with other teams, these two areas deserve as much attention as the proper handling of the boat itself.

Based on this situation, many practical questions are of interest to get more knowledge on the boat handling, the impact of different tactical decisions, and the reaction of the boat to different trim techniques.

Practical questions of interest

Based on the factors described earlier, there are many practical questions of interest:

- How can sailors better understand the handling of their boats?
- How does the boat react to trim decisions?
- What are the effects of different tactical decisions?
- Can the specific route that is sailed for a given course be improved?

A comprehensive list would go far beyond the scope of this book.

For this case study, let us focus on questions that are of practical interest for learning more about boat speed, effects of trim techniques, and tactical decisions. These questions are sufficient to describe the case study from a data quality perspective:

- Tacking: how much time and distance are lost when tacking? During a tack, the boat must turn through the wind and, therefore, loses speed. Only when the boat reaches its new course and gets wind from the other side is speed regained. Depending on the time and distance required for a tack under various conditions, the tactical decision to make many or few tacks during a race can be optimized.
- How does the upwind speed of the boat depend on influential factors like wind speed, wind direction, and sail size? On various settings of the trim functions or on the crew itself? The boat, for example, gains speed if it is sailed with only 55 degrees to the wind. The question is whether this additional speed compensates for the longer distance that has to be sailed to get to the same effective distance upwind. What data are needed to optimize the angle for sailing to the wind?
- How does the maximum possible course angle to the true wind depend on influential factors like wind speed, sail size, and trim function settings? Different trim functions allow changing the shape of the foresail and the mainsail. The effective course angle and speed in setting these trim functions is of special interest. Given the crew members, their physical condition, the boat, its sailing characteristics, the weather conditions, and the sea conditions, what are the optimal trim settings over the route chosen for the given course?
- How do different tactical decisions perform during a race? When sailing upwind, for example, tactical decisions can include making only a few tacks and sailing to the left area of the course and then to the buoy, sailing to the right area of the course, or staying in the middle of the course and making many tacks.
- How does the actual sailing speed or the angle to the true wind deviate from other boats competing in the race? Comparing the effect of different course decisions between the participating boats is of special interest. We can then see which areas of the course have the best wind conditions or whether different boats perform in a different way under similar conditions.
- By comparing the performance across boats in a race, can the sailing abilities of the individual crews and boats be further analyzed and improved?

Technical and data background

The boat *Wanda* uses a Velocitek SC-1 device, which is a GPS device that collects the coordinates from different satellites in 2-second intervals. Based on these data, the device displays in real time the average and maximum speeds and the compass heading. This information is vital during a race to track boat performance. The GPS device also stores the data internally in an XML format. These data can then be transferred to a computer by using a USB cable.

The following data are available in the XML file with one row per 2-second interval: timestamp (date and time), latitude, longitude, heading, and speed. A short excerpt is shown in Figure 1.2.

Figure 1.2: Content of the XML file that is exported by the GPS device

```
<?xml version="1.0" encoding="utf-8"?>
<VelocitekControlCenter xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:xsd="http://www.w3.org/2001/XMLSchema" createdOn="2009-05-
25T18:29:02.65625+02:00"
xmlns="http://www.velocitekspeed.com/VelocitekControlCenter">
 <MetadataTags>
 <MetadataTag name="BoatName" value="Wanda" />
 <MetadataTag name="SailNo" value="0000" />
 <MetadataTag name="SailorName" value="xxxx" />
 </MetadataTags>
 <CapturedTrack name="090521_131637" downloadedOn="2009-05-
25T18:23:46.25+02:00" numberTrkpts="8680">
 <MinLatitude>47.773464202880859</MinLatitude>
 <MaxLatitude>47.804649353027344</MaxLatitude>
 <MinLongitude>16.698064804077148</MinLongitude>
 <MaxLongitude>16.74091911315918</MaxLongitude>
 <DeviceInfo ftdiSerialNumber="VTQURQX9" />
 <SailorInfo firstName="xxxx" lastName="yyyy" yachtClub="zzzz" />
 <BoatInfo boatName="wwww" sailNumber="0000" boatClass="Unknown" hullNumber="0"
/>
 <Trackpoints>
 <Trackpoint dateTime="2009-05-21T13:49:24+02:00" heading="68.43" speed="5.906"
latitude="47.792442321777344" longitude="16.727603912353516" />
 <Trackpoint dateTime="2009-05-21T13:49:26+02:00" heading="59.38" speed="5.795"
latitude="47.7924690246582" longitude="16.727682113647461" />
 <Trackpoint dateTime="2009-05-21T13:49:28+02:00" heading="65.41" speed="6.524"
latitude="47.792495727539062" longitude="16.727762222290039" />
 <Trackpoint dateTime="2009-05-21T13:49:30+02:00" heading="62.2" speed="6.631"
latitude="47.792518615722656" longitude="16.727849960327148" />
 <Trackpoint dateTime="2009-05-21T13:49:32+02:00" heading="56.24" speed="6.551"
latitude="47.792549133300781" longitude="16.727928161621094" />
 <Trackpoint dateTime="2009-05-21T13:49:34+02:00" heading="60.56" speed="5.978"
latitude="47.792579650878906" longitude="16.728004455566406" />
 <Trackpoint dateTime="2009-05-21T13:49:36+02:00" heading="61.57" speed="7.003"
latitude="47.792606353759766" longitude="16.728090286254883" />
 <Trackpoint dateTime="2009-05-21T13:49:38+02:00" heading="52.03" speed="7.126"
latitude="47.792636871337891" longitude="16.728176116943359" />
```

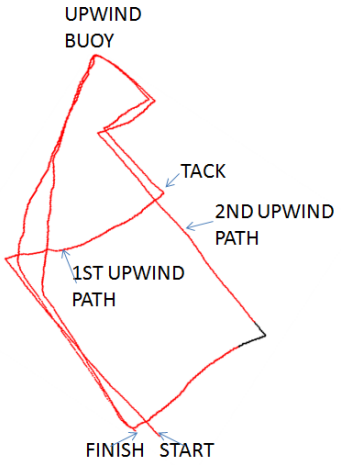
These data can be analyzed by using Velocitek software to visualize the course, speed, and heading of a boat and to perform simple analyses.

Other data processing systems can use these data to perform specific analyses. In this case study, the data have been imported into SAS by using a SAS DATA step to prepare the data for analysis. Different graphical and statistical analyses can be performed to answer the practical questions listed earlier.

Figure 1.3 is a line chart that has been produced using SAS/IML Studio. It shows the race course and the specific route that was sailed. The course is similar to the one shown in Figure 1.1. After the start, the boat goes upwind to the first buoy and then downwind to the second buoy. The circuit is repeated a second time, and then the finish line is reached. Note that some annotations are included to identify some features of the specific route that was sailed.

On the first upwind path, the boat obviously experienced a wind shift that slowed the progress to the upwind buoy. From an ex-post tactical viewpoint, the boat should have tacked again or it should have been farther to the right in the course area.

Figure 1.3: Line chart of one race



The **second data source**, in addition to the GPS track point data, is a logbook that contains crew-recorded data for each race. For example, it includes the names of crew members, the sailing area, the general wind direction, the general wind strength, and other meteorological values as well as the size and type of the sails.

Data quality considerations

Based on the practical and technical background, many aspects of the analysis can be discussed, but our focus is data quality:

- The GPS track point data are only available for two boats: the boat whose crew wants to perform analyses and for one additional boat. Most of the remaining boat teams either did not save the GPS track point data or they were unwilling to share the data with potential competitors. A few other teams did not use a GPS device. Thus, comparison between boats can only be performed in a limited way.
- The GPS device only collects data that are related to the position of the boat itself. Information about the wind direction and wind strength are not collected. In order to collect this information, a separate device is needed. Therefore, the questions that relate to the effect of wind strengths and wind direction shifts cannot be answered with the GPS track point data.
- Assuming a constant behavior of the boat itself and the way the helmsman pilots the boat, it is possible to infer the wind direction from the compass heading of the boat. However, if the wind shifts immediately before or during a tack, the analyst might not be able to identify if the tacking angle and the new heading after the tack are caused by a wind shift or by a different helmsman behavior.
- There is no timestamped protocol of the different settings of the trim functions of the boat. There is only a rough recording, in many cases based on personal memory, of the main trim function settings at the beginning of the race. It is therefore not possible to identify if a change in speed in the second upwind course is due to a different trim setting or to different wind or helmsman conditions.
- During the sailing season, only the GPS tracking data were recorded in a regular and structured way. Thus, at the end of the season, when the idea to perform this analysis arose, some of the other data, such as participating crew members, size of sails used, average wind speed, average wind direction, and other trim settings, were retrieved based on the memory of the crew members. So, clearly, some information was lost, and given the human factor some data were recorded with potential errors. (The probability of data accuracy and completeness is very high for data that are collected automatically through electronic systems. However, for data that are manually documented and entered into a system, the probability is lower—not because of systematic and malicious bias but due to environmental distractions and fatigue. In practice, the human source of error can be found in many other cases.)

These points reflect the situation of **data unavailability**. Some data that are desirable for analytical insights are simply not available, which means that some analyses cannot be done at all and other analyses can only be done in a reduced scope.

- To analyze the data using SAS software, the data must be exported to a PC from the GPS device as XML files. In this step, the data must have been correctly collected and stored in the GPS device itself and then exported correctly into an XML file.
- After the XML file is stored on the PC, it is read into SAS, which then validates that the file was imported correctly. Care has to be taken to correctly separate the individual fields and to correctly represent the date, time, and numeric values. Thus, before the data can be analyzed, there are multiple places where erroneous information can enter the data, but there are also multiple checks to ensure data quality.

The **correctness in data collection and the accuracy of their transfer** are vital to data preparation for analysis. Before data can be analyzed, data validation must be performed to ensure that the real-world facts, the data, are collected by the device correctly, stored correctly, and transferred correctly to the computer for use by the analysis software.

- GPS data for another boat are available, but before these data can be combined with the first boat, the time values must be realigned because the internal clock of the other boat was one hour behind. When the two data sets are aligned by the common factor of time, they can be merged, and the combined information can be used for analysis. Note that in many cases augmenting data can add important information for the analysis, but often the augmenting data must be prepared or revised in some way (for example, time, geography, ID values) so that they can be added to existing data.
- If three races are sailed during a day, the log file contains the data for the three races as well as the data for the time before, between, and after the races. To produce a chart as shown in Figure 1.3, the data need to be separated for each race and any unrelated data need to be deleted. Separating the races and clearing the non-race records is frequently quite complicated because the start and end of the race is often not separately recorded. To perform this task, the data need to be analyzed before as a whole and then post-processed with the start and end times.

These points show that prior to the analysis, **data synchronization and data cleaning** often need to be done.

- In a few cases, the GPS device cannot locate the position exactly (for example, due to a bad connection to the satellites). These cases can cause biases in the latitude and longitude values, but they can especially impact the calculated speeds. For example, if a data series contains a lost satellite connection, it can appear that the boat went 5.5 to 6 knots on average for over an hour and then suddenly went 11.5 knots for 2 seconds. These data must be cleaned and replaced by the most plausible value (for example, an average over time or the last available value).
- In another case, the device stopped recording for 4 minutes due to very low temperatures, heavy rain, and low batteries. For these 4 minutes, no detailed track point data were recorded. During this interval, the position graph shows a straight line, connecting the last available points. Because this happened when no tacking took place, the missing observations could be inserted by an interpolation algorithm.
- In another case, the GPS device was unintentionally turned off shortly before the start and turned on again 9 minutes later. Much tacking took place during this interval, but the missing observations cannot be replaced with any reasonable accuracy.
- Some of the above examples for “data unavailability” can also be considered as missing values similar to the case where information like sail types and settings and crew members were not recorded for each race.

These data collection examples show how some **values** that were intended to be available for the analysis can be **missing or incorrect**. Note that the wind direction and wind strength data are considered to be **not available** for the analysis because they were not intended to be collected by a device. The GPS track point data for the first 9 minutes of the race **are missing** because the intention was to collect them (compare also chapters 3 and 5).

- Practical questions to be answered by the analysis involve the consequences of tacking and the behavior of the boat when tacking. The data for all races contain only 97 tacks. If other variables like wind conditions, sail size, and trim function settings are to be considered in the analysis as influential variables, there are not enough observations available to produce stable results.

To answer practical questions with statistical methods, a representative sample of the data and a sufficient amount of data are required. The more quality data that are available, the greater the confidence we can have in the analysis results.

Case 1 summary

This example was taken from a non-business, non-research area. It shows that data quality problems are not limited to the business world, with its data warehouses and reporting systems. Many data quality aspects that are listed here are relevant to various practical questions across different analysis domains. These considerations can be easily transferred from sailboat races to business life.

Many analyses cannot be performed because the data were never collected, deleted from storage systems, or collected only in a different aggregation level. Sometimes the data cannot be timely aligned with other systems. Due to this incomplete data picture, it is often impossible to infer the reason for a specific outcome—either because the information is not available or because the effects cannot be separated from each other.

These aspects appear again in the following chapters, where they are discussed in more detail.

1.3 Case Study 2: Data Management and Analysis in a Clinical Trial

General

This case study focuses on data management and analysis in a long-term clinical trial. In general, the specifics of a clinical trial significantly impact data collection, data quality control, and data preparation for the final analysis. Clinical trials focus on data correctness and completeness because the results can critically impact patient health and can lead, for example, to the registration of a new medication or the admission of a new therapy method.

This case study only discusses the data quality related points of the trial. The complete results of the trial were published in the *Official Journal of the American Society of Clinical Oncology* 2005 [2].

Functional problem description

The clinical trial discussed in this case study is a long-term multicenter trial. More than 10 different centers (hospitals) recruited patients with melanoma disease in stages IIa and IIb into the trial that lasted over 6.5 years. Each patient received the defined surgery and over 2 years of medication therapy A or B. The trial was double-blind; neither the patient nor the investigator knew the actual assignment to the treatment groups. The assignment to the treatment group for each patient was done randomly using a sequential randomization approach.

During and after the 2 years of treatment, patients were required to participate in follow-up examinations, where the patient's status, laboratory parameters, vital signs, dermatological examinations, and other parameters that describe patient safety were measured. The two main evaluation criteria were the recurrence rate of the disease and the patient survival rate. Depending on their time of recruitment into the trials, patients were expected to participate in follow-up exams at least 3 years after the end of the therapy phase.

Patients were recruited into this trial in different centers (hospitals). All tasks in treatment, safety examinations, trial documentation into case-record forms (CRFs), and evaluation of laboratory values were performed locally in the trial centers. Tasks like random patient allocation into one of the two treatment groups (randomization), data entry, data analysis, and trial monitoring were performed centrally in the trial monitoring center.

The following tasks in the trial took place locally in the trial center:

- Recruitment of patients into the trial and screening of the inclusion and exclusion criteria.
- Medical surgery and dispensing of medication to the patients.
- Performance of the follow-up examinations and documentation in writing of the findings in pre-defined CRFs.
- Quality control of the accuracy and completeness of the data in the CRFs compared to patient data and patient diagnostic reports. This step was performed by a study monitor, who visited the trial centers in regular intervals.

The CRFs were then sent to the central data management and statistic center of the trial. This center was in charge of the following tasks:

- Performing the randomization of the patients into the treatment groups A and B with a software program that supports sequential randomization.
- Storing the randomization list, which contained the allocation patient number to treatment, in access-controlled databases.
- Maintaining the trial database that stored all trial data. This database was access-controlled and was logging any change to the trial records.
- Collecting the CRFs that were submitted from the trial centers and entering them into the trial database.
- Performing data quality reports on the completeness and correctness of the trial data.
- Performing all types of analyses for the trial: safety analyses, adverse event reports, interim analyses, and recruitment reports.

Practical question of interest

The practical question of interest here was the ability to make a well-founded and secure conclusion based on the trial data results.

- The main criterion of the trial in the per-protocol and in the intention-to-treat analysis was the comparison of the disease-free intervals between the treatment groups and the comparison of the survival between treatment groups.
- To achieve this, a sufficient number of patients, predefined by sample-size calculation methods, were needed for the trial. To check whether the recruitment of patients for the trial was on track, periodic recruitment reports were needed.
- Beside the main parameters, recurrence of disease and survival, parameters that describe the safety of the patients, was collected for the safety analysis. Here laboratory and vital sign parameters were analyzed as well as the occurrence of adverse events.

All these analyses demanded correct and complete data.

Technical and data background

The randomization requests for a patient to enter the trial were sent by fax to the monitoring center. The trial data were collected on paper in CRFs and entered into an Oracle database. This database did not only support the data entry, but it also supported the creation of data quality and completeness reports.

Data quality considerations

Based on the scope of a clinical trial presented here, the following aspects of data quality during data collection, data handling, and data analysis are of interest:

- To improve the correctness of the data provided through the CRFs, a clinical monitor reviewed and validated the records in each trial center before they were submitted to the monitoring center. In this case, very high data quality was established at the very beginning of the process as possible errors in data collection were detected and corrected before data entry for the records.
- Each information item was entered twice (that is, two different persons entered the data). Therefore, the data entry software had to support double data entry and verify the entered data against lists of predefined items, value ranges, and cross-validation conditions. It also had to compare the two entered versions of the data. This was achieved by online verification during data entry and by data quality reports that listed the exceptions that were found during the data checks.
- A crucial point of data quality in this clinical trial was the correctness of the values of the randomization lists in the clinical database. This randomization list translates the consecutive numeric patient codes into treatment A and treatment B groups. Obviously, any error in this list, even for a single patient number, would bias the trial results because the patient's behavior and outcome would be counted for the wrong trial group. Therefore, much effort was used in ensuring the correct transfer of the randomization list into the trial database.
- The randomization list was provided to the data monitoring center as hardcopy and as a text file in list form. Thus, the text file had to be manually preprocessed before it could be read into the database. Manual preprocessing is always a source of potential error and unintended data alteration. The final list that was stored in the database was manually checked with the originally provided hardcopy by two persons for correctness.
- As an additional check, two descriptive statistics were provided by the agency that assigned the double-blind treatments and prepared the randomization list, the mean and the standard deviation of the patient numbers. For each group A and B, these statistics were calculated by the agency from the source data and then compared with the corresponding statistics that were calculated from the data that were entered in the trial database. This additional check was easy to perform and provided additional confidence in the correctness of the imported data.

These practices indicate that in clinical trials there is an extremely strong emphasis on the correctness of the data that are stored in the clinical database. To achieve and maintain data correctness, the focus must be on validating and cross-checking the **data collection**, the **data transfer**, and the **data entry** of the input data.

- To trace changes to any field in the trial database, all data inserts, updates, or deletions of the trial database were logged. Based on this functionality, a trace protocol could be created for any field to track if, and how, values changed over time. An optional comment field enabled the insertion of comments for the respective changes. The commenting, logging, and tracing processes were very important in maintaining high data quality, especially for data fields that were critical for the study: the time until relapse, the survival time, and the patient status in general. The ability to perform an uncontrolled alteration of data does not comply with external regulations, and it is a potential source of intended or unintended biasing of the trial and the trial results.
- From a process point of view, it was defined that any change to the data, based on plausibility checks or corrections received at a later point, would only be made to the trial database itself. No alterations or updates were allowed at a later stage during data preparation for the analysis itself. This requirement was important to create and maintain a single source of truth in one place and to avoid the myriad coordination and validation problems of data preparation logic and data correction processes dispersed over many different analysis programs.

- Based on logging data inserts, updates, and deletions, it was also possible to rollback either the database or an individual table to any desired time point in the past. The historical database replication functionality is required by Good Clinical Practice (GCP) [10] requirements. It enables analysts to access the exact status of a database that was used for an analysis in the past.

For security and regulatory reasons, **tracing changes in the database** was very important. In addition to the support for double data entry, the trial database provided functionality for tracing changes to the data and for enabling the database rollback to any given date.

- Because there was no central laboratory for the trial, the determination of the laboratory parameters was done locally in each hospital. But the nonexistence of a central laboratory led to two problems.
  - Some laboratories did not determine all the parameters in the measurement units that were predefined in the CRF, but they did define them in different units. Thus, to obtain standardized and comparable measurements, the values had to be recalculated in the units specified in the CRF.
  - The normal laboratory values for the different laboratories differed. Frequently different laboratories have different normal laboratory values. To perform plausibility checks for the laboratory values based on normal laboratory values, a different lookup table for each trial center may have been needed.
- As it turned out, the usage of normal laboratory values was not suitable for plausibility checks because roughly 15% of the values fell outside of these limits. If the normal laboratory values had been used, the validation effort required would have been much too high and would result in the acceptance of the slightly out of limit value. The purpose of data validation was not to highlight those values that fell out of the normal clinical range but to detect those values that could have been falsely documented in the CRF or falsely entered into the database. Thus, it was decided to compute validation limits out of the empirical distribution of the respective values and to calibrate the values that way so that a reasonable amount of non-plausible values were identified.
- The primary evaluation criterion of the trial was the time until relapse. For each treatment group, a survival curve for this event was calculated and compared by a log rank test. To calculate this survival curve, a length of the period is needed, which is calculated from the patients' trial start until the date of their last status. In the survival analysis, the status on the patient's last date, relapse yes or no, was used to censor those observations with no relapse (yet). The important point here is the correct capture of the patient status at or close to the evaluation date. In a long-term trial, which continues over multiple years and contains a number of follow-up visits, the patients' adherence to the trial protocol decreases over time. Patients do not show up to the follow-up visits according to schedule. The reasons can be from both ends of the health status distribution. For some, their health status is good, and they see no importance in attending follow-up meetings; for others, their health status is bad, and they cannot attend the follow-up meetings. Therefore, without further investigation into the specific reason for not adhering to the trial protocol, identifying the patient's exact status at the evaluation snapshot date is complicated. Should the status at their last seen date be used? That is an optimistic approach, where if no relapse has been reported by those not adhering to the trial protocol, then no relapse has occurred. Or should it be based on the pessimistic assumption that a relapse event occurred immediately after their last seen date?
- Also, determining the population to be used for the per-protocol analysis is not always straightforward. The per-protocol analysis includes only those patients who adhered to all protocol regulations. A patient, for example, who did not show up at the follow-up visits for months 18 and 24 might be considered as failing to follow-up at an interim analysis, which is performed after 2.5 years. If, however, they showed up at all consecutive scheduled visits in months 30, 36, and 42, then they might be included in the final analysis after 4 years.

These points focus on the **correctness of the data** for the analysis. In the following, plausibility checks and rules on how to define a derived variable play an important role:

- In the respective study, a desired sample size of 400 patients was calculated using sample-size calculation methods. This number was needed to find a difference that is statistically significant at an alpha level of 0.05 and a power for 80%. Recruitment was planned to happen over 4 years (approximately 100 patients per year).
- After 9 months of recruitment, the clinical data management center notified the principal investigator that the actual recruitment numbers were far below the planned values and that the desired number of patients would only be achieved in 6.5 years. Continuing the study at this recruitment pace for the desired sample size would delay the trial completion substantially, about 2.5 years. But stopping recruitment and maintaining the 4-year schedule would result in too few patients in the trial. Based on this dilemma, additional hospitals were included in the trial to increase the recruitment rate.

In clinical research, much financial support, personal effort, and patient cooperation are needed. It is, therefore, important to ensure there is a reasonable chance to get a statistically significant result at the end of the trial, given that there is a true difference. For this task, sample-size planning methods were used to determine **the minimum number of patients (data quantity)** in the trial to prove a difference between treatments.

Case 2 summary

This case study shows the many data quality problems in a very strict discipline of research, clinical trials. There are two strong focuses: the correctness of the data and the sufficiency of the data. To obtain sufficient correct and complete data, substantial effort is needed in data collection, data storage in the database, and data validation. The financial funding and personal effort to achieve this result need to be justified compared to the results. Of course, in medical research, patient safety—and, therefore, the correctness of the data—is an important topic, which all clinical trials must consider. In other areas, the large investment of effort and funding might not be easily justified.

From this case study, it can be inferred that in all analysis areas, there is a domain-specific balancing of costs against the analysis results and the consequences of less than 100% correct and complete data.

1.4 Case Study 3: Building a Data Mart for Demand Forecasting

Overview

This last case study shows data quality features for an analysis from the business area. A global manufacturing and retail company wants to perform demand forecasting to better understand the expected demand in future periods. The case study shows which aspects of data quality are relevant in an analytical project in the business area. Data are retrieved from the operational system and made available in analysis data marts for time series forecasting, regression analysis, and data mining.

Functional problem description

Based on historic data, demand forecasting for future periods is performed. The forecasts can be sales forecasts that are used in sales planning and demand forecasts, which, in turn, are used to ensure that the demanded number of products is available at the point of sale where they are required. Forecast accuracy is important as over-forecasting results in costly inventory accumulation while under-forecasting results in missed sales opportunities.

Demand forecasts are often created on different hierarchical levels (for example, geographical hierarchies or product hierarchies). Based on monthly aggregated historic data, demand forecasts for the next 12 months can be developed. These forecasts are revised on a monthly basis. The forecasts are developed over all levels of the hierarchies; starting with the individual SKU (stock keeping unit) up to the product subgroup and product group level and to the total company view.

Some of the products have a short history because they were launched only during the last year. These products do not have a full year of seasonal data. For such products, the typical methods of time series forecasting cannot be applied. For these products, a data mining model is used to predict the expected demand for the next months on product base data like price or size. This is also called *new product forecasting*.

A data mining prediction model has been created that forecasts the demand for the future months based on article feature, historic demand pattern, and calendar month. For products that have a sufficient time history, time series forecasting methods like exponential smoothing or ARIMA models are employed. For many products, the times series models provide satisfactory forecasts. For some products, especially those that are relatively expensive, if they have variables that are known to influence the quantities sold, then regression models can be developed, or the influential variables can be added to ARIMA models to form transfer function models.

Functional business questions

The business questions that are of primary interest in this context are as follows:

- On a monthly basis, create a forecast for the next 12 months. This is done for items that have a long data history and for items that have a short data history.
- Identify the effect of events over time like sales promotions or price changes.
- Identify the correlation between item characteristics like price, size, or product group and the sales quantity in the respective calendar month.
- Identify seasonal patterns in the different product groups.
- Beyond the analytical task of time series forecasting, the system also needs to provide the basis for periodic demand reporting of historic data and forecast data and for planning the insertion of target figures for future periods into the system.

Technical and data background

In this case study, the company already had a reporting system in place that reports the data from the operational system. Data can be downloaded from this system as daily aggregates for a few dimensions like product hierarchy or regional hierarchy. These data have two different domains, the order and the billing data. Time series forecasting itself was only performed on the order data. For additional planning purposes, billing data also were provided.

Another important data source was the table that contains the static attributes (characteristics) for each item. This table contained a row for each item and had approximately 250 columns for the respective attribute. However, not all variables were valid for each item. Beside a few common attributes, the clusters of attributes were only relevant to items of the same item group.

Some additional features for each item were not yet stored in the central item table, but they were available in semi-structured spreadsheets. These spreadsheets did contain relevant information for some product groups that could be made available for the analysis.

Data quality considerations

The following features of the project had a direct relation to data quality:

- Historic order data and historic billing data for the last 4 years were transferred from the operational system to the SAS server. Given all the different dimensions over millions of rows, the data import was several gigabytes in size.
    - This amount of data cannot be checked manually or visually. To verify correctness of the data that were imported into the SAS system, a checksum over months, weeks, product hierarchies, and so forth was created. The checksum shows the number of rows (records) read in, the number of rows created, and so on. While in SAS virtually any checksum statistic can be calculated, only those

statistics that are also available in the original system (for example, a relational database) can be used for comparison. For some dimensions of the data, the checksums differed slightly.

- Usually it is a best practice rule to investigate even small differences. In this case, however, most of the small deviations were due to a small number of last-minute bookings and retrospective updates that were shown in the life system on a different day than in the export files. This also made the comparison difficult between the exported data from the life system and the values in the life system itself. There is the possibility that immediately after the data was exported, the numbers had already changed because of new transactions. In a global company, it is not possible to export the data during the night when no bookings are made. From a global perspective, it is never "night."

These points reflect the **control of the data import process and correctness check** after transfer and storage in the source system.

- In the case described here, the order and billing data were complete. It is reasonable for the billing data to be complete in order to bill customers; otherwise, revenue would be lost.
- Static data like item features, other than product start date and price, were not as well-maintained because they are not critical for day-to-day business operations. However, from an analytical perspective, the characteristics of various items are of interest because they can be used to segment items and for product forecasting. The table containing item characteristics had a large number of missing values for many of the variables. Some of the missing values resulted when variables were simply not defined for a specific product group. The majority of the missing values, however, occurred because the values were not stored.
- The non-availability of data was especially severe with historic data. Orders from historic periods were in the system, but in many cases it was difficult to obtain characteristics from items that were not sold for 12 months.
- Because the company was not only the manufacturer of the goods but also the retailer, point-of-sale data were also available. Point-of-sale data typically provide valuable insight, especially for the business question on how to include short-term changes in customer behavior in the forecasting models. However, capturing these data for the analysis was complicated because only the last 12 months of point-of-sale data were available in the current operational system. For the preceding time period, data were stored in different systems that were no longer online. To capture the historic data from these older systems, additional effort in accessing historic backup files from these systems was required.
- Another source of potential problems in data completeness was that for some item characteristics, no responsibility for their maintenance and update was defined. This is especially true for data provided in in the form of spreadsheets. Therefore, in the analyses, because it was uncertain whether an updated version of these data would be available in a year, care had to be taken when using some characteristics.

These points refer to the **availability and completeness of the data**. While it is important to emphasize that existing data were transferred correctly into the system for analysis, it is also important to clearly identify the completeness status of the data. In this case, what was observed was typical for many data collection situations: Transactional data that are collected by an automatic process, like entering orders, billing customers, and forwarding stock levels, are more complete and reliable. Also, data that control a process are typically in a better completeness state. Data that need to be collected, entered, and maintained manually are, in many cases, not complete and well-maintained.

- For demand forecasting of products with a shorter history, classical time series forecasting methods could not be applied. Here, a repository of items was built with the respective historic data and item characteristics. As described earlier, predictive data mining models were built based on these data to forecast demand. The repository initially contained hundreds of items and increased over time as more and more items became available in the database. At first view, it might seem sufficient to have hundreds of items in the database. But in a more detailed view, it turned out that for some product categories only around 30 items were available, which was insufficient to build a stable prediction model.

- Individual product-group forecasting was necessary because products in different groups had different sets of descriptive variables. Also, products in different groups were assumed to have different demand patterns. Thus, separate models were needed. For the whole set of items, only six characteristics were commonly defined, and so the analyst had to balance data quantity against individual forecasting models. The analyst could decide to build a generic model on only the six characteristics, or they could build individual models with more input variables on fewer observations.
- In time series forecasting, for products with a long time history, each time series is considered and analyzed independently from the other series. Thus, increasing or decreasing the number of time series does not affect the analytical stability of the forecast model for a single series. However, the number of months of historic data available for each time series has an effect on the observed performance of the model.

These points refer to **data quantity**. For stable and reliable analytic results, it is important to have a sufficient number of observations (cases or rows) that can be used in the analysis.

- Deferring unavailable data. In some cases, data did not exist because it was not stored or it was not retained when a newer value became available or valid. Sometimes it is possible to recalculate the historic versions of the data itself. A short example demonstrates this:
  - To analyze and forecast the number of units expected to be sold each month, the number of shops selling the items is an influential variable, but it is often unavailable.
  - To overcome the absence of this important variable, an approximate value was calculated from the data: **the number of shops that actually sold the article**. This calculation can easily be performed on the sales data.

The content of the variable is, however, only an approximation; the resulting number has a deceptive correlation with the number of items sold because a shop where the item was offered but not sold is not counted. The inclusion of this variable in a model results in a good model for past months, but it might not forecast well for future months.

Case 3 summary

This case study shows features of data quality from a time series forecasting and data mining project. Extracting data from system A to system B and creating an analysis data mart involve data quality control steps. This process is different from a set of a few, well-defined variables per analysis subject because there are a large number of observations, hierarchies, and variables in the product base table. Given this large number of variables in the data, it is much more difficult to attain and maintain quality control at a detailed level. The case study also shows that for accuracy of results in analytical projects, the data quantity of the time history and the number of observations is critical.

1.5 Summary

Data quality features

This chapter discusses three case studies in the data quality context. These case studies were taken from different domains, but they share the fact that the results depend directly on the data and, thus, also on the quality of the data.

The quality of data is not just a single fact that is classified as good or bad. From the case studies, it is clear that there are many different features of data quality that are of interest. Some of these features are domain-specific, and some depend on the individual analysis question. The different features can be classified into different groups. For each case study, an initial grouping was presented.

These case studies are intended not only to whet your appetite for the data quality topic but also to highlight typical data quality specifics and examples of analyses.

A classification of data quality features that were discussed in the case studies follows. This classification is detailed in chapters 3 through 9.

Data availability

- GPS data were only available for two boats. No wind data were collected.
- No recording of the trim setting on the sailboat was done.
- Static information on the items to be forecasted was not entered into the system or maintained over time.
- Historic data or historic versions of the data (like the static information on the items from 12 months ago) were not available.
- Point-of-sale data from historic periods that were captured in the previous operational system could not be made available or could only be made available with tremendous effort.

Data completeness

- Some of the GPS data were missing because the device was turned on late or it did not record for 4 minutes.
- For a number of patients, no observations could be made for some follow-up visits because the patients did not return.
- Static information on the items to be forecasted was not completely entered into the system or maintained over time.

Inferring missing data from existing data

In some cases, an attempt was made to compensate for the unavailability of data by inferring the information from other data, for example:

- Estimating the wind direction from the compass heading on the upwind track.
- Approximating the unavailable number of shops that offered an item for sale from the number that actually sold them.

In both cases, a substitute for the unavailable data was found, which should be highly correlated with the missing data. This enables reasonable approximate decisions to be made.

Data correctness

- In a few cases, the value of the calculated boat speed from the GPS data appeared to be wrong.
- For the sailboat case study, when data for sail size and composition of crew members were captured post-hoc, the sail sizes could not be recaptured with 100% certainty. In this case, a most likely value was entered into the data.
- The source data on the CRFs were manually checked by an additional person.
- Data entry in the clinical trial was performed twice to ensure accuracy.
- The transfer of the randomization list for the trial followed several validation steps to ensure correctness.
- Transferring data from a GPS device to a PC text file and importing the file into the analysis software are potential sources of errors if data change.
- Any change in clinical trial data was recorded in the database to provide a trace log for every value.
- For each laboratory parameter, a plausibility range was defined to create an alert list of potential outliers.

- Transferring millions of rows from the operational system to the analysis system can cause errors that are hard to detect (for example, in the case of a read error when a single row is skipped in the data import).

Data cleaning

- After the GPS were imported, the values of the XML file needed to be decoded.
- The GPS data for the individual races needed to be separated.
- Implausible laboratory values were output in a report that the monitor used to compare with the original data.

Data quantity

- The database did not contain enough tacking data to analyze the tacking behavior of the boat in detail.
- In the clinical trial, a measurement in study control was taken to increase the number of participating clinical centers. Otherwise, not enough patients for the analysis would have been available.
- In the data mining models for the prediction of the future demand for items that have only a short history, only a few items had a full set of possible characteristics.