Business Intelligence I
Supplement
Statistical Methods for Data Preprocessing

W. Grossmann

Content

- Data Editing
- Data Imputation
- Data Integration

Data Editing – Definition

- Data Editing is the task to find erroneous values in the data
- Other denotation: Data Profiling
- Two types of processing in data editing
  - Replace the wrong value by a missing value and replace afterwards the missing value by methods of data imputation
  - Combine finding and replacing in one step
- If data editing is combined with the correction of the erroneous values the term data cleaning is also used

Data Editing – Reasons for errors

- Reasons for erroneous values:
  - Missing values
  - Wrong specification of variable type (string, numeric, date, ….)
  - Special characters
  - Outliers in numerical values
  - Numerical inconsistencies: Some variables have to comply with specific rule, e.g. VAT is a percentage of price, geocodes are within a well defined range
  - Logical inconsistencies: Variables have to fulfill soe logical rules, e.g. ***if*** (person is in compulsory military service)
    then (gender = male)

Data Editing – Methods

- Due to the different reasons for erroneous data there exist many different approaches
  - Deductive editing
  - Selective editing
  - Automatic editing
  - Manual Editing

Data Editing – Deductive editing

- Deductive editing is used for detecting and correcting of errors which have a structural cause (systematic errors)
  - For such errors one can formulate general logical rules like ***if***(condition) ***then***(correction) e.g.
    - ***If*** (number of employees > 0 and Costs of Temporary employees = 0) ***then*** Number of temporary employees = 0
- Problems:
  - A rule must be specified for all errors
  - Definition of the transitive closure of all errors
    (A => B) and (B => C) => (A => C)
- Software for deductive editing:
  R-package *deducorrect*

Data Editing – Automatic editing

- This method changes values in such a way that the number of necessary overall corrections is as small as possible
- Example:
  - Suppose there are four constraints for four attributes:

$$x_1 - x_2 > 0, \quad x_3 - x_4 > 0, \quad x_1 - x_3 > 0, \quad x_4 - x_2 > 0$$

  - The record (5,1, 6, 4) violates this rule; how should we correct the values?
- Software for deductive editing:
  R-package *editrules*

Data Editing – Outlier detection

- There are many different descriptive statistical methods for outlier detection (boxplot, scatterplot, cluster methods, regression trees)
- Important questions to be answered in advance:
  - Is the outlier an edit or correct value

Data Editing – Other methods

- Selective editing: Correction of those values which are of importance for further analysis
- Manual editing: Editing base on the expertise of persons (only for small scale problems)

Missing Values – Definition

- In practical BI applications one is frequently confronted with missing values
- Negligence of the observations with missing values is often not appropriate
  - Comparison of results for different attributes is not possible
  - In the case of calculation of correlations there may be problems due to loss of a considerable amount of data
- We are interested in methods for imputing the missing information

Missing Values – Models

- Imputation of missing values is based on the understanding of the mechanism which generates the missing values
- We consider only cases of so called item non-response which means that for each observation some information is available

Missing Values – Models

- Missing completely at random (MCAR): A missing value is MCAR if the probability of occurrence of a missing value is the same for each observation
  - Neglecting the missing value has no influence on the bias of an estimators but increases the variance of the estimator

Missing Values – Models

- Missing at random (MAR): The probability of occurrence of a missing value depends on the values of an attribute which is available
- Examples:
  - Certain person groups (age groups, gender groups, political groups, ….) are often not willing to give certain information (e.g. income, political preferences)
- Neglecting the missing values increases the bias of the results

Missing Values – Models

- Dependence of the missing values on not observed characteristics
- Examples:
  - The willingness to answer a certain question depends on the acceptance of the study
- In such cases it is necessary to estimate the unknown information
- The worst case is the dependence for a missing value depends on the variable itself (e.g. income)
  - Correction of bias almost impossible

Missing Values – Models

- We consider in the following only the cases MCAR and MAR
- These assumptions have to be justified from problem understanding
- Substitution of the missing values by using the mean is usually not a recommended strategy (tendency towards the mean, loss of variability)

Models for Missing Values - Adjustmet

- In the case of MCAR one could estimate the probability for the occurrence of missing values, for example by using logistic regression, and reweighting the date according to these probabilities
- This method can be used for one variable
- It influences the variance of the estimator

Imputation of Missing Values - Methods

- Deductive Imputation: Similar to deductive editing
  - Applicable in cases of systematic errors
- In the case of MAR a number of methods are available for imputing missing values into the data
- The most frequently used are the following ones:
  - Hot deck imputation
  - kNN imputation
  - Regression imputation
  - Multiple imputation

Imputation of Missing Values – Hot deck imputation

- Hot deck imputation is a popular method used in surveys
- Basic idea: Looking in the existing data (hot deck) for complete observations which are candidates for donating a value and select one of the data sets as a donor for the missing observation
  - The quality of the method depends on the frequency of using a certain observation as donor; only few donors should be used many times

Imputation of Missing Values – Hot deck imputation

- Methods for selection of the donor
  - Sequential hot deck: Start with a first donor (value for imputation) and scan sequentially the donor candidates in the data set. If a value is missing substitute the missing value with this value; otherwise substitute the donor value with the value in the data
  - Stochastic hot deck: Select randomly a value from the donor candidates. The random selection is based on selection with replacement

Imputation of Missing Values – kNN imputation

- This method is similar to kNN classification
- First of all a distance measure between two observations is defined
  - The distance has to be defined in such a way that quantitative and qualitative attribute can be included in the distance calculation (cf. also clustering)
- Find the k nearest neighbors to the data and impute:
  - The mean or median in case of quantitative variables
  - The mode in case of qualitative variables
- Usually k = 5

Imputation of Missing Values – Regression imputation

- This method assumes that the mechanism for missing values is MAR
- In the case of a missing value for a quantitative variable define a regression model for the variable with missing observations:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

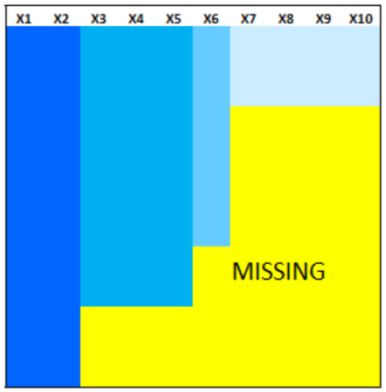
- Estimate the coefficients of the regression model

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k + \hat{\varepsilon}$$

- The missing value is defined as the prediction value of the regression model

Imputation of Missing Values – Regression imputation

- In the case of qualitative variables predict the values based on other models, for example a logistic regression model in the case of binary variables
- In the case of application for many variables it is necessary to sort the observations according to the number of missing values:



Imputation of Missing Values – Regression imputation

- The imputation is done sequentially for the cases with increasing number of missing values
- The method is sometimes also called stochastic regression imputation in contrast to methods which use the prediction of the mean value

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

Imputation of Missing Values – Multiple imputation

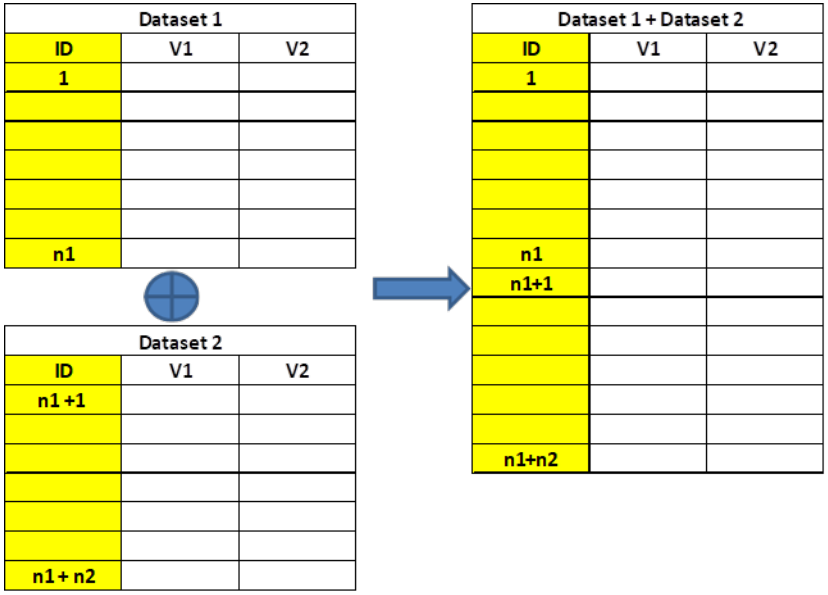
- Multiple imputation is a generalization of regression imputation
- First of all a regression model is estimated from complete observations
- From this model a number of predictors are generated
- The final imputation is defined as the mean of the candidates
- Multiple imputation takes into account the uncertainty of the imputation
- Software for imputation: R packages *VIM, mice, mix, mi*

Data Integration – Definition

- The goal of data integration is the provision of one dataset from a number of different datasets
- For applications BI methods we are mainly interested in integration of one data matrix where the rows represent the observed cases and the columns the attributes (variables) which are used in further analysis
- In case of time stamped data the structure is a more general because we can have for each observed case a number of rows indexed by an additional time stamp
  - The R package *tidyverse* is a versatile environment for producing different structures of time stamped data
- We consider here only data without temporal structure

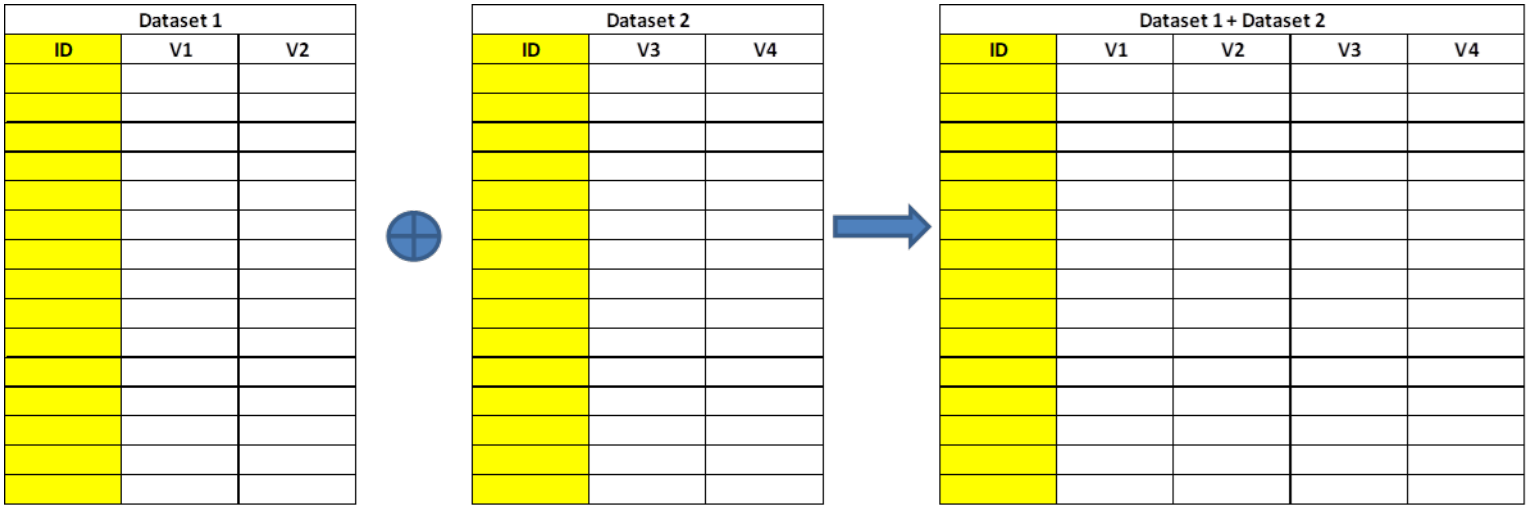
Data Integration – Scenarios

- One can distinguish the following scenarios for integration of two datasets:
  - Data adding:



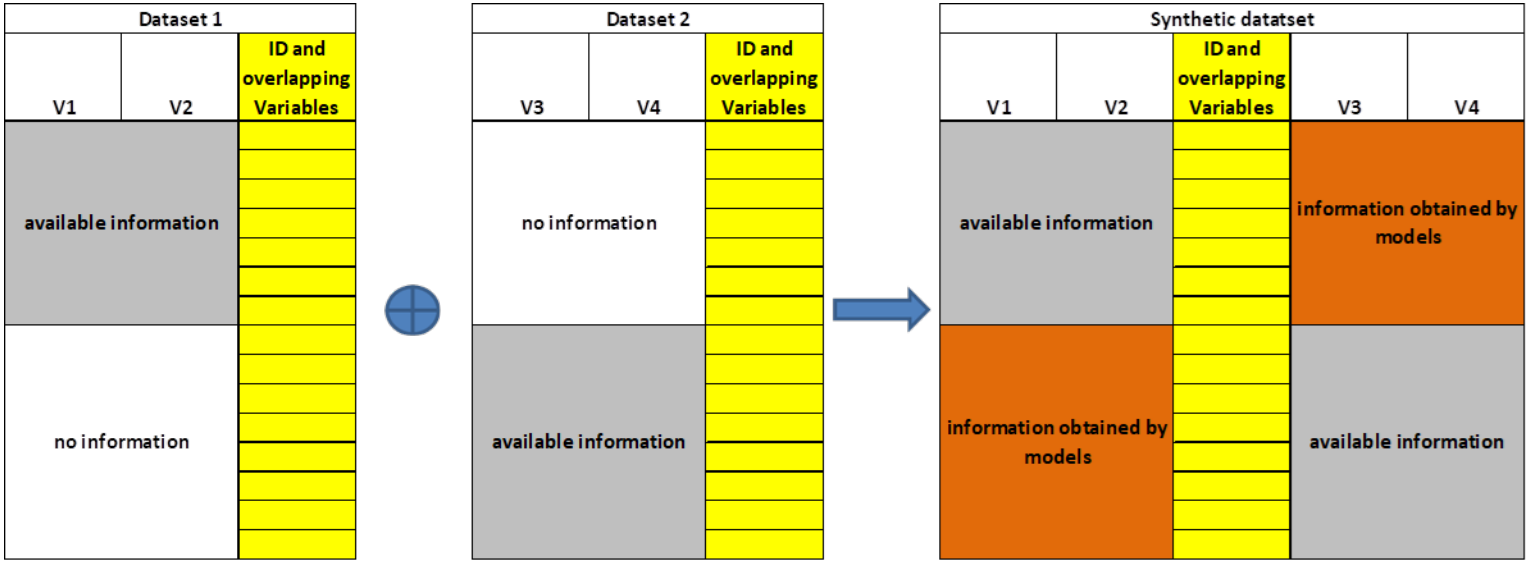
Data Integration – Scenarios

– Data linkage:



Data Integration – Scenarios

– Statistical Matching:



Data Integration – Data Adding

- In the case of data adding the most important pre-processing steps are:
  - Data editing of both datasets
  - Duplicate detection, i.e. finding the records in the two datasets which represent identical observations
- Finding of duplicates is known under the name of record matching
  - In the case of unique identifiers for all case this is rather easy

Data Integration – Data Adding

- A formal model for data matching
  - Given two datasets A and B with attributes which allow identification of identity of the data define the following partition of the data

$$M = \{(a,b);\ a = b,\ a \in A,\ b \in B\}$$
$$U = \{(a,b);\ a \neq b,\ a \in A,\ b \in B\}$$

  - In the case of data adding the set M should be empty, otherwise some editing is necessary
- In any case there should be an additional editing for the new dataset
- If necessary an imputation should be done

Data Integration – Data Linkage

- Data linkage is the process of combination of the records of two datasets which contain information of the same observation unit
- The combination of the records is called record linkage, also called record matching
- More general is the term entity resolution which is used in case of matching more complex structures

Data Integration – Data Linkage

- Processing steps in record linkage:
  - Identify with record matching the records in both datasets which represent the same observed cases
  - Define for each observation unit a new record which keeps all the attributes of both datasets
- Two scenarios for record linkage can be distinguished:
  - Deterministic (exact) record linkage
  - Probabilistic (stochastic) record linkage

Data Integration – Record Linkage

- In deterministic record linkage the identification of the matching record is explicit possible by a unique identifier (key) or by a number of attributes which allow unique identification;
  - The linkage corresponds to a join operation
- In probabilistic record linkage such an identification is not possible, we can define only a probability for the records that they belong to the same observation unit

Data Integration – Probabilistic Record Linkage

- From a methodological point of view probabilistic record linkage can be seen as a classification task
- The main challenge is the determination of the probability that two record represent the same observation unit
- Usually this probability is calculated by using similarity measures for the attributes under consideration
- This similarity depends on the type of attributes and uses often some preprocessing of the attributes

Data Integration – Probabilistic Record Linkage

- In the case of string variables (names) the similarity is frequently calculated by the Jaro-Winkler distance

$$d(s_1, s_2) = \begin{cases} 0 & m = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m} \right) + & m \neq 0 \end{cases}$$

$m =$ Number of the identical symbol in den strings

$t =$ Number of transpositions for obtaining a corresponding symbol

Statistical Matching – Problem Description

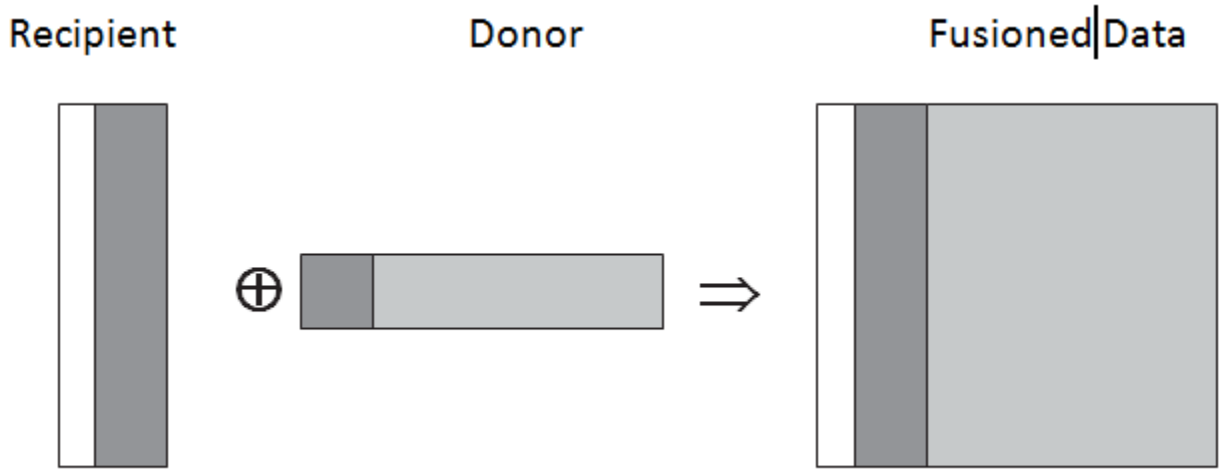
- Statistical matching using ideas similar to imputation in the following context:
  - Given two data sets A and B with the following structure:

    A = [X|Y] is called the recipient data set

    B = [X|Z] is called the donor data set
  - The variables X are called common variables
  - The task is to generate for each observation in the recipient data set information about the values of the variables Z

Statistical Matching – Problem Description

- Graphical representation:



Statistical Matching – Problem Description

- Statistical matching and data fusion are often understood as synonyms, but strictly speaking statistical matching is only a core method for data fusion.
- **Data fusion** is the process of integrating multiple data sources to produce more consistent, accurate, and useful information than that provided by any individual data source (Wikipedia)
  - Many applications in obtaining information from different sensor data

Statistical Matching – Methodology

- Statistical matching encompasses the following steps
  - Find variables X which are candidates for common variables.
    - These candidates should high correlation with the donor variables
  - Check dependencies between the Y variables and the Z variables
    - In ideal case the variables Y given X and Z given X should be independent
  - Check the common variables for completeness
    - If necessary complete the common variables using imputation

Statistical Matching – Methodology

- Find for each observation in the recipient set an observation in the donor set which is similar to the recipient. Two frequently used methods for defining the similarity are:
  - Similarity based on distances between observations; this is similar to kNN imputation
  - Define a propensity score, i.e. compute the probability that an observation in the joint data set of donors and recipients belongs to the data set is recipients using a logistic regression. This probability is called propensity score. The donor is now defined by the observation in the donor data which has a similar propensity as the donor. The justification of the method is that in case of selection of the donors the distribution of the missing values is in the recipient data set is the same as in the donor data set

Statistical Matching – Methodology

  - Validate the solution by
    - Internal validation: In how far reflect the imputed variables in the recipients the properties of the variables in the donors
    - External validation: Is the new data set useful for the intended analysis
- Statistical matching is useful because it can reduce the costs of additional data collection
- A prerequisite is that the donors and recipients are randomly selected from a population
  - In particular, for propensity scores this is important
  - Software tools for matching: R packages *matchingR, Matching*