Zeitschrift: Studies in Communication Sciences : journal of the Swiss Association of Communication and Media Research

Herausgeber: Swiss Association of Communication and Media Research; Università della Svizzera italiana, Faculty of Communication Sciences

Band: 4 (2004)

Heft: 2

Artikel: A methodology for data quality assessment on financial data

Autor: Amicis, Fabrizio de / Batini, Carlo

DOI: https://doi.org/10.5169/seals-790977

Studies in Communication Sciences 4/2 (2004) 115-136

Fabrizio De Amicis & Carlo Batini*

A METHODOLOGY FOR DATA QUALITY ASSESSMENT ON FINANCIAL DATA[1]

This paper proposes a methodology for data quality assessment that has been defined and applied in practice on financial data, in particular for registry data used to describe financial instruments. The methodology encompasses five major phases that define the recognition and classification of primary variables, the data quality analysis techniques and data quality rules used for the inspection of selected data quality dimensions. In the quantitative objective assessment, the measurement of erroneous observations considers the correlation between data quality dimensions. Three independent experts define a qualitative subjective assessment. The results of the two assessments are compared in order to detect discrepancies that are useful for the data quality experts to select actions for data quality improvement. The examples reported in this paper have been selected from a real case.

Keywords: financial data classification, inspection of data quality dimensions, data quality rules, data quality measurement, subjective and objective data quality assessment.

* Università di Milano Bicocca, Dipartimento di Informatica, Sistemistica e Comunicazione, fv.deamicis@libero.it, batini@disco.unimib.it

[1] The work presented in this paper has been partially supported by FIRB MAIS project – Multi-channel Adaptive Information Systems: models, methodology, qualifying object-oriented platform and architectures for the flexible on-line information systems.

1. Introduction

The business of practically all medium and large financial institutes is supported by information systems. Such systems manage databases characterised by high-dimensionality and large volume of data coming from heterogeneous sources and complex data download processes. Evaluation of the quality of an entire financial database is a difficult and expensive task and thus, it is rarely done. The development of practical, low cost measures related to financial information quality is therefore fundamental for business performance, operational risk and improvement of information management.

Financial data, considered in this paper, refers to variables that characterize the financial instruments, i.e. instruments used by financial institutions to perform their business; it can be classified into four main categories: *a) registry data* used to describe financial instruments (see section 4.1); *b) daily data* that refers to prices and exchange rates; *c) historical data* mainly related to time series, and *d) theoretical data,* that corresponds to the output of financial models such as e.g. the beta coefficient. In the present paper we will consider only registry data.

This paper proposes a methodology designed to obtain standard measurements and assessments for financial registry variables stored in the internal operational databases of a bank with feasible and low cost tools. The aim of the methodology is the definition of quantitative objective and qualitative subjective data quality assessment of financial registry variables. The methodology is based on the experience gained in data quality projects developed in different banks and financial institutes. For reasons of confidentiality, we will present only examples that cannot identify the financial institutes.

2. Related work

Previous work on methodologies for data quality assessment appears in Pipino et al. (2002), Lee et al. (2001), and Kahn et al. (2002). Common to our methodology and Pipino et al. (2002) is the idea of comparing quantitative objective assessment and qualitative subjective assessment in order to detect discrepancies and take actions for data quality improvements. Our methodology is more detailed in two perspectives, i.e. the identification and classification of variables and data analysis techniques that precede the assessment, and the definition of appropriate indices,

data quality rules, measurements and strategies for quantitative and qualitative assessments. In Lee et al. (2001) the AIMQ methodology for assessing information quality in organizations is defined. In Kahn et al. (2002), the Product and Service Performance (PSP) model, a component of the AIMQ methodology, is used as the base for the assessments of information quality.

From our knowledge, no complete methodology for data quality assessment on financial data has been proposed in the relevant literature. Contributions on specific data dimensions come from important international banks. Consistency for customer databases and accuracy of the reports analysed by account managers are considered in Matsumura and Shouraboura (1996). The use of artificial intelligence to benchmark organizational data flow is discussed in McKeon (2003). In Klein et al. (1996) the analysis results on the expectations about the base rate of errors on municipal bond data, expressed by five municipal bond analysts, are reported. Dasu and Johnson (2003) provide several analysis techniques based on Exploratory Data Mining.

3. Research Methodology

As mentioned in the related work, several methodologies are proposed in literature to assess the quality of data. Our methodology has been designed abstracting from data quality assessment projects developed for a major Italian bank, an important Italian asset management institute and a private Swiss bank. The methodology makes use of data quality analysis techniques available in literature and tailors such techniques to the financial data domain. Most of the techniques have been inspired from real needs, results and successful experiences.

4. A Methodology for Data Quality Assessment of Financial Registry Data

The inputs and outputs of the methodology concern:

a. *Financial variables and observations* - Typically, the number of registries in a financial database is usually around 50. In the present version we apply the methodology to a part of the complete set of the financial registry variables.
b. *Financial context* - The design of a financial database is related to a specific context of a bank and therefore a meaningful classification of registry variables depends on the financial context.

c. *Business rules* - A given database might be subject to any number of integrity constraints, also known as *business rules or data edits* (see Ross 1994; Redman 1996) of arbitrary complexity.
d. *Data quality analysis techniques* They comprise descriptive statistics, hypothesis testing, cluster analysis, detection of outliers, and statistical methods based on data visualizations.
e. *Process description*. The knowledge on data loading and updating processes has an important impact on the selection of data quality dimensions. For example, when a data loading process is not optimised, then timeliness and uniqueness as data quality dimensions are affected by errors.
f. *Business and data quality expertise* – Expertise is important for successful data quality analysis and it is fundamental for the definition of an appropriate qualitative subjective assessment.

As outputs, the methodology provides a qualitative subjective and quantitative objective data quality assessment of financial registry variables; these are standard measurements that can be used in the benchmarking of financial registry variables among different banks.

In figure 1, we provide a high level description of the five phases of the methodology and of their relationships. Initially, financial variables are selected according to their importance, and classified in order to pilot further analyses and evaluate the final results. The analysis phase evaluates data quality dimensions through appropriate metrics, and business rules, in order to identify errors. Phase 3, starting from measures of errors, provides an objective assessment. Business expertise is the resource used for performing qualitative subjective assessment in phase 4. Comparative analysis between the two assessments is finally performed in phase 5. In the following sections we examine each phase in detail.

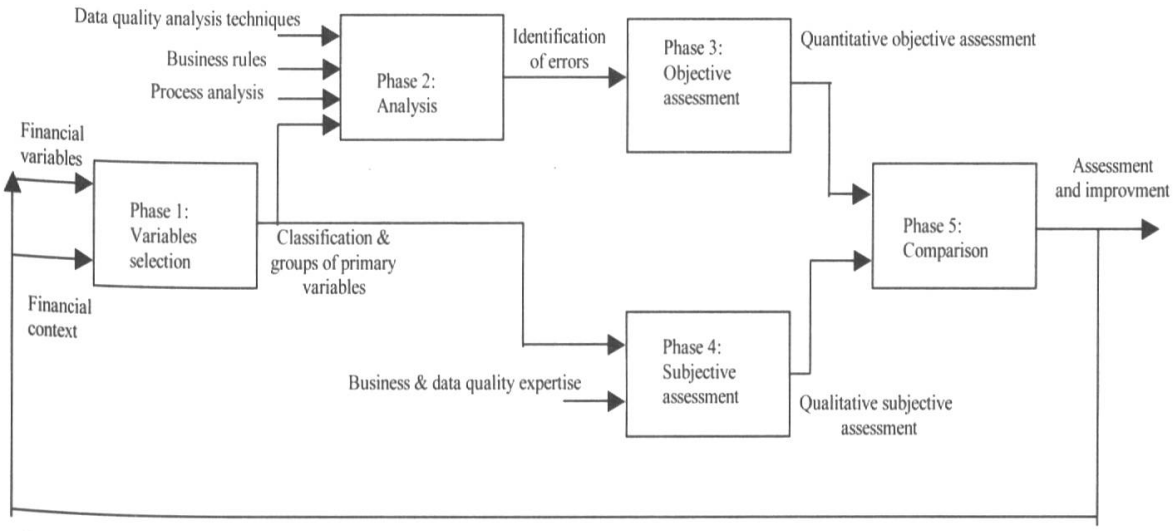

Figure 1: Methodology phases 1 to 5

4.1. Variable Selection: Recognition, description and classification of primary variables of financial registries

Figure 2 presents the organization of this phase. The objective is initially the identification of the most relevant financial registry variables. In the absence of knowledge from previous assessments, the selection is performed according to the importance of variables, while, in presence of knowledge from previous assessments, according to their significance and effectiveness. Outputs of phase 5 are important for data quality improvements and for tuning previous phases. For example, when the output reveals that a specific variable has an acceptable data quality level according to objective and subjective assessments, then in phase 1 the set of selected variables can be modified including other variables.

In Table 1 we show a sample output of the selection step. In the experiences performed in real life assessments, about 30 variables were considered.

Table 1: List of registry variables.

Variable	Description
Contract size	Size of a future contract in terms of quantity of quotes.
Coupon frequency	Number of times a year bond coupons are paid.
ISIN code	Identifies a financial instrument.
Market currency	ISO code of the official trade currency used in the market where the financial instrument is traded
Maturity date	Date on which financial instrument ceases its life
Price of conversion	Price due to purchase a single unit of the security we are going to convert in.
Moody's Rating	Rating code provided by Moody's
S&P Rating	Rating code provided by Standard & Poor

The goal of the classification step is to guide data quality analysts and business experts in applying in phase 3 and phase 4 more effective and correct data quality analysis techniques and data quality subjective assessments. In the classification step, primary variables are characterised, according to their meaning and role, as *qualitative/categorical (C), quantitative/numerical (N) or dates (D).* Classification is also performed by creating groups of variables that form "related issues" groups which affect the behaviour of investors and consumers, characterized by risk, business and descriptive factors.

Figure 2: Schematic presentation of phase 1: selection and classification of variables

Table 2 presents the classification of variables of Table 1 by variable typology and issue group.

Table 2: Classification of registry variables.

	Categorical variables	*Numerical variables*	*Date variables*
Risk (Group 1)	Moody's Rating S&P Rating		
Business (Group 2)	Market currency	Contract size Price of conversion	Maturity date
Description (Group 3)	ISIN code	Coupon frequency	

4.2. Analysis: Inspection of data quality dimensions

This phase identifies data quality dimensions and business rules to be measured and makes use of practical techniques used for inspection of financial data. Selection and inspection of data quality dimensions is related to process analysis, with the final goal of discovering the main causes of erroneous data, such as unstructured and uncontrolled data loading and data updating processes. The final result of data quality analysis - on selected data quality dimensions - is the identification of errors. Figure 3 presents inputs and output of phase 2, and the list of data quality dimensions considered, that we discuss in the rest of the section.

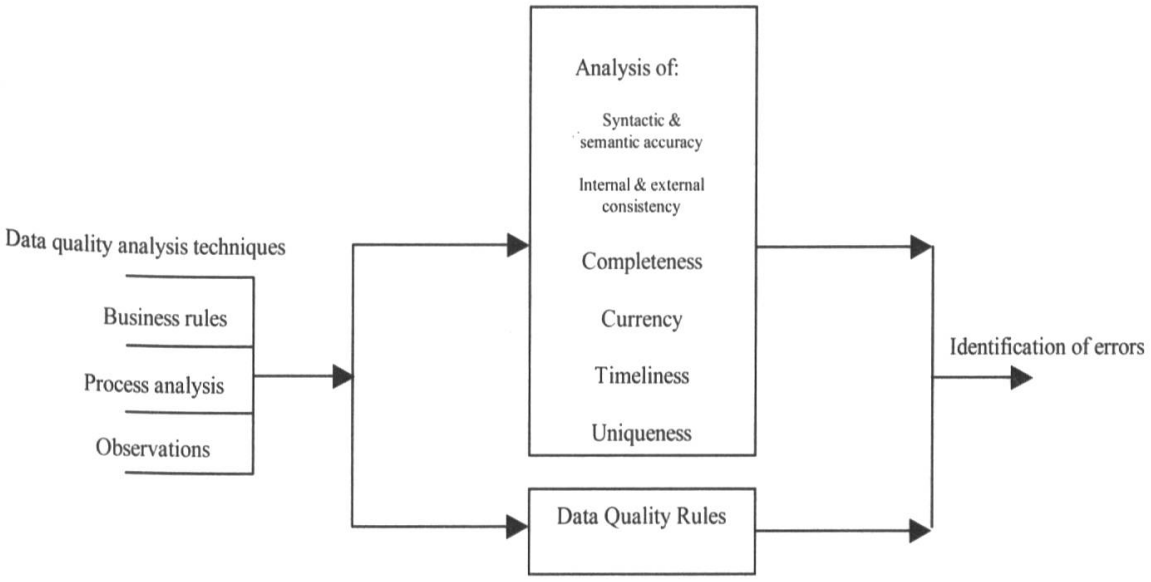

Figure 3: Schematic presentation of phase 2

Syntactic accuracy (D_1)

Syntactic accuracy is defined in Redman (1996) as the proximity of the data item value v to a domain data item v' considered correct. Errors related to syntactic accuracy can be categorized mainly as mismatched errors, insertion or deletion errors. To detect this kind of errors it is often necessary to compare two or more text strings and measure the difference (see Finkelsein et al. (1994)). Many functions have been developed to identify approximate matching (agrep, soundex, fonix, edit, etc., see Zobel and Dart (1995)). Our methodology proposes the following syntactic accuracy analysis technique.

- when a reference data dictionary or a lookup table is available, syntactic accuracy can be easily checked by comparing data values with the lookup table. For example, we may assume to have at disposal for categorical variables a domain table containing all the categories.
- If the lookup table is not available, a criterion for automating the identification of an erroneous category is the low frequency of the suspected category and its similarity to a frequent category; i.e., a rare category which is "highly similar" to a more frequent category is a good candidate for a new codification.

Semantic accuracy (D_2)

Semantic accuracy of values is defined as the distance between v and v', being v' the value corresponding to v considered semantically correct. As explained in Fugini et al. (2002), semantic accuracy is difficult to be quantified and the verification can be expensive. For numerical variables, a possible approach to verify semantic accuracy is to analyse descriptive statistics calculated by different sources. Relevant differences between descriptive statistics for the same variable can be due to semantic accuracy. For example, the variable Contract size has been used in a real life context with two different meanings: in a first data source it is used to represent the size of a future contract in terms of quantity of quotes, in another data source, to represent the value of a future contract. Such homonymy can be seen as a semantic inaccuracy both at the value and at the schema level. The detection of this error was possible when analysing the great difference between descriptive statistics associated to the two different sources, as shown in Table 3. Descriptive statistics depend on variable typology.

Table 3: Descriptive statistics of contract size variable

Variable observed: Contract size	Max value	Mean value	Standard deviation
Source 1	25	1	3
Source 2	164845	1153	13286

Internal and external consistency (D_3 and D_4)
Consistency indicates that two or more values do not conflict with each other. In our context, *internal consistency* refers to the consistency of a data value item within the same financial instrument; internal inconsistency can be detected and controlled using ad-hoc rules. *External consistency* refers to the consistency of a data value item in different types of business information. When applicable, database bashing is a technique used to detect and control external inconsistency, see Redman (1996); it involves comparing records from two or more databases.
Inconsistency is often related to redundancy. In particular, there is a negative relationship between data consistency and data redundancy, see Rutra et al. (1999). For example suppose that a financial instrument is represented by two distinct entries in the database. When one of the two entries has been updated and the other has not, a case of inconsistency occurs.

Completeness (D_5)
Completeness refers to the extent of presence of data values in a variable. In addition to the variable missing values, data quality analyst has also to check the presence of non-informative values that have to be classified as missing values. Moreover, it is important to distinguish between variables for which completeness is essential (for example, the ISIN code), variables for which completeness is only partially required (for example the Moody's or S&P rating is not provided for all the financial instruments), and variables for which completeness is not relevant, because of the variable meaning (e.g, in case of multiple response variables). Figures 4.1 and 4.2 present the histograms of missing value frequencies for registry variables, and for all the variables of a financial database; note that a large number of variables are only nominally present in the database.

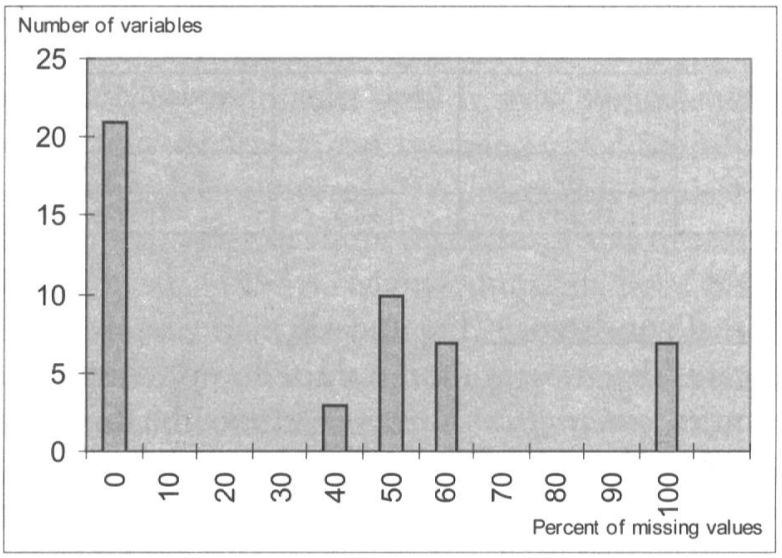

Figure 4.1: Histogram for the frequencies of missing values of registry variables

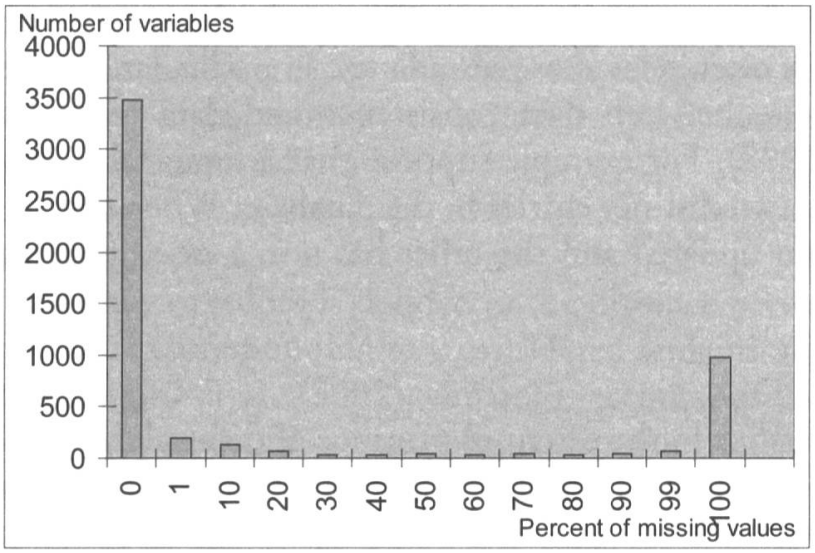

Figure 4.2: Histogram for the frequencies of missing values of financial variables

Currency (D_6)

The currency dimension refers to the temporal proximity of the data value result of the last update with respect to the current value. As an example, if the value of an address changes and it is not updated on time, then the value of the address is obsolete, and, as a consequence, incorrect. As another example, data of external financial providers contain errors; it may happen, for example, that a spot forex (the exchange value of e.g dollar vs euro), coming from an external provider, appears as not correct for several seconds during the day. If the loading process runs in the peri-

od of time in which the data is not correct, the internal database inherits and keeps the error until the loading process will run again.
This dimension can be analyzed examining the frequency distribution of selected variables over the time, and comparing values from difference sources.

Timeliness (D_7)
As reported in Fugini et al. (2002), timeliness is defined as the availability of data on time, or rather within the time constraints specified by the destination organization. Most of the outliers detected from the comparison between in-house ratings and ratings from an external independent source are related to timeliness problems. An example of a possible technique, used in practice, to detect and measure errors of timeliness dimension for rating variables is described here, based on comparing in-house data with external independent sources (for example Bloomberg). Suppose that $V^{External}$ is an external variable that represents the Moody's rating (or Standard & Poor's rating) and $V^{In\text{-}house}$ is the related in-house variable. The domain for Moody's rating is {Aaa, Aa1, Aa2, Aa3, A1, A2, A3, Baa1, Baa2, Baa3, Ba1, Ba2, Ba3, B1, B2, B3, Caa, Ca, C}. In general, let's define to be the domain of the two variables. A new numerical variable (R) is defined as follows:

$$(1) \qquad R_i = | i^{External} - i^{In-house} |$$

where i is the i-th value of the external/in-house variable.

The following example shows that external and in-house rating variables differ mainly for a single step. For this specific example, the cause of errors of the rating variable can be interpreted as an incorrect loading process of the internal rating variable.

Distribution of R values can be collapsed in a single index, on a 0-1 scale, in the following way:

$$(2) \qquad \frac{\sum_{i=1}^{n} R_i \times N_i}{\max(R_i) \times N}$$

where
- N_i is the number of observation related to R_i
- $max(R_i)$ is the maximum value of R_i
- N is the total number of observations.

If the distribution of the variable R is uniform and the frequency of errors is not significant, the data quality analyst should examine other possible causes of errors, such as for example typo errors. Figure 5 shows examples of the R distribution, for the comparison of in-house rating values (Moody's rating) and rating values of external provider.

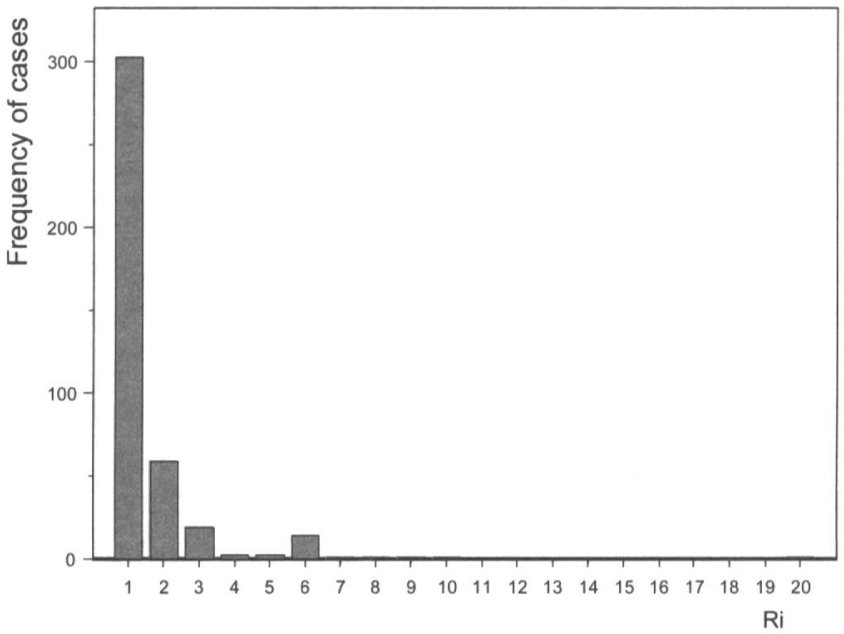

Fig. 5.1: A distribution of variable R that puts problems related to timeliness in evidence

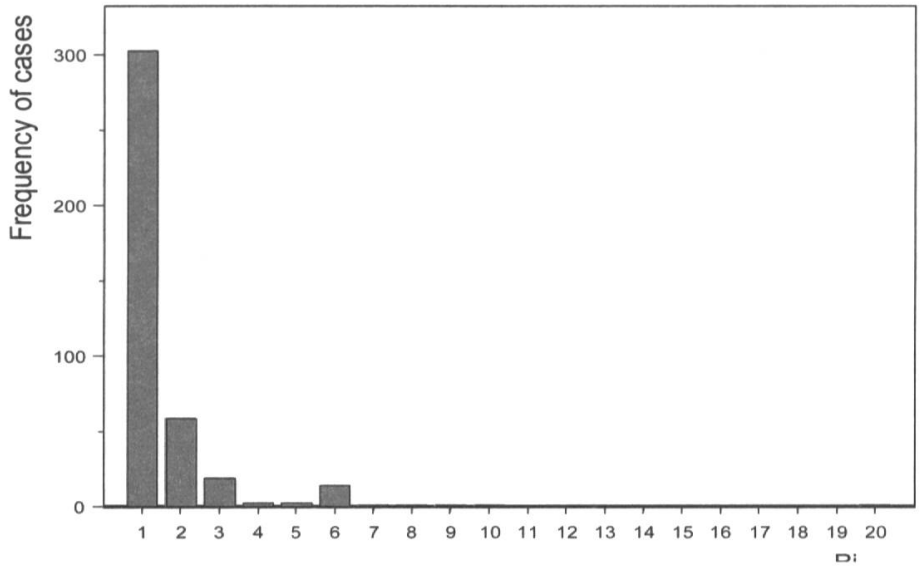

Fig. 5.2: A distribution of variable R that does not reveal problems of timeliness

Uniqueness (D_8)

The problem of duplicates is typical in many databases. If X and Y are two records (also called cases or observations) represented by vector (X_1, X_2,...X_v)

and $(Y_1, Y_2, \ldots Y_v)$ respectively, X and Y respect a symmetric relationship *duplicate* if they represent the same financial instrument. It is not always easy to detect and clean duplicates from a table with a large number of records, because it may happen that there is not a perfect match between two records (see Zobel and Dart (1995)). For this reason, it is necessary to verify if *similar cases* have duplicate information. The similarity among cases can be quantified by a score statistic that is related to the number of matches (see Gower (1971)). As an example (see Figure 6), in a financial database the table representing registry variables had a percentage of duplicates of 12%. Analysing the number of real loaded records over time, against expected ones, it is possible to detect possible loading errors and candidate duplicate records.

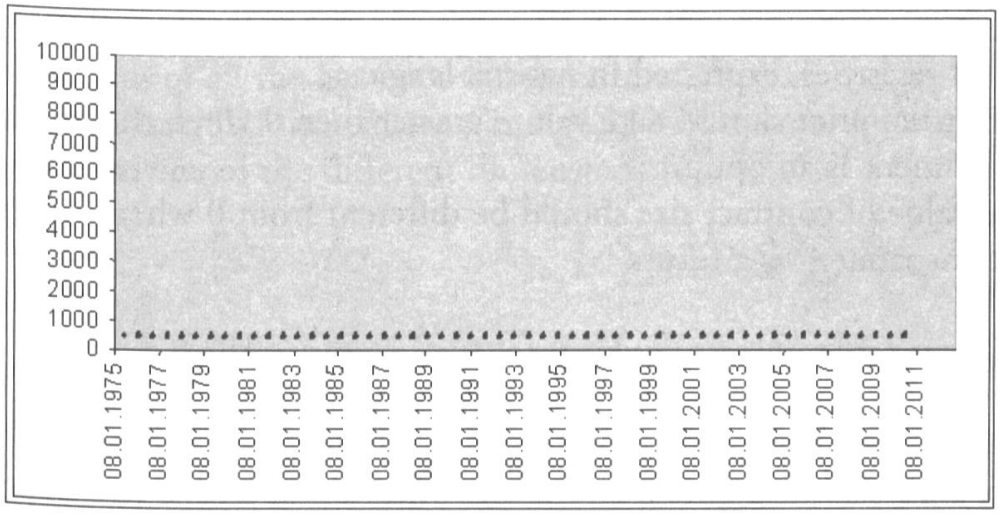

Figure 6.1: Expected number of loaded financial instrument records

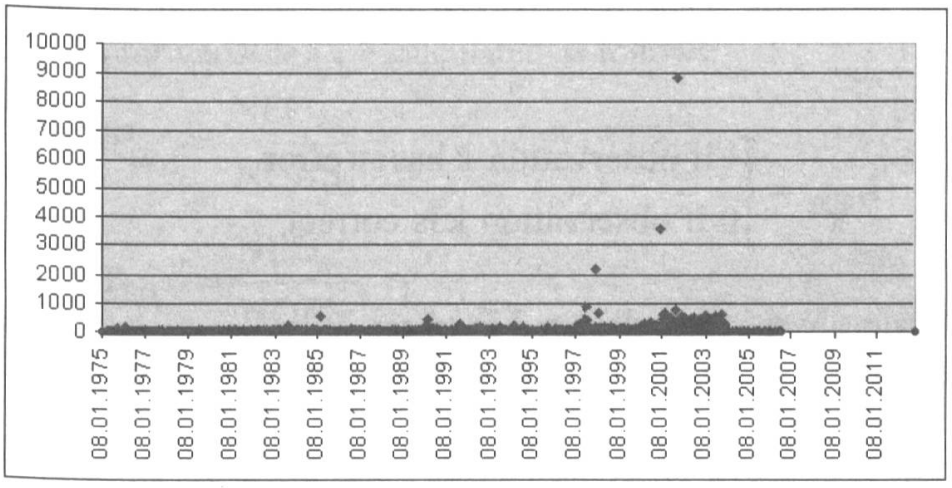

Figure 6.2: Real number of loaded financial instrument records

Data Quality Rules (D_9)
Data quality rules represent all the relevant semantic properties of variables not captured by the previous quality dimensions. They are defined starting from the business rules available as results of the database design process plus the observations and general rules resulting from the previous analysis on data quality properties. In our approach, data quality rules are a superset of rules commonly referred as business rules or integrity constraints, needed frequently in practical situations as a further check on the correctness of existing business rules. For example, a NOT NULL constraint can be overtaken using a non-informative value like "ZZZ" for categorical variables and zero for numerical variable). Outputs of data quality rules are erroneous observations.

We present two examples of data quality rules to be applied to financial registries, expressed in natural language.

1. The strike price should take values greater then 0 when the financial instrument is an option
2. The value of contract size should be different from 0 when the financial instrument is a future

4.3. Quantitative objective assessment

In phase 2, we have defined data quality dimensions and related data quality analysis techniques used for identification of errors. In this phase, we define appropriate indices for the evaluation and quantification of the global data quality level.

It is possible to count the *Number of Erroneous Observations* using the following indicator.

$$(3) \qquad I_{OBSk} = \begin{cases} 1 \text{ if observation k has an error} \\ 0 \text{ if observation k is correct} \end{cases}$$

where errors are the ones detected in phase 2.

The number of erroneous observations, for data quality dimension j and variable i, is:

$$(4) \qquad NEO_{ij} = \sum_{k=1}^{n} I_{OBSkij}$$

where n is the number of observations.

Note that the data quality dimensions are not statistically independent, since certain types of errors can influence more then one dimension. For example, if a financial instrument is duplicated and contains a syntactic accuracy error, then the erroneous observation will be counted in dimensions D_1 and dimensions D_8. As a consequence of the above discussion, in our methodology two different countings are defined.

The number of erroneous observations from data quality dimensions can be calculated in two different ways:

a. as the sum of all the erroneous observations (5);
b. using Poincarè's formula, see Chung (1974), we may thus excluding the *intersection* of the different dimensions (6);

$$(5) \qquad NEO_{+j} = N[EO_{1j} \cup EO_{2j} \cup ... \cup EO_{mj}] = \sum_{i=1}^{m} N[EO_{ij}] = \sum_{i=1}^{m} NEO_{ij}$$

$$(6) \qquad NEO_{+j}^{1} = N[EO_{1j} \cup EO_{2j} \cup ... \cup EO_{mj}] = \sum_{i=1}^{m} N[EO_{ij}] - \sum_{i<k} \sum N[EO_{ij} EO_{kj}] + \sum_{i<k<l} \sum \sum N[EO_{ij} EO_{kj} EO_{lk}] - ... + (-1)^{m-1} N[EO_{1j} EO_{2j} ... EO_{mj}]$$

where $\mathbf{EO}_{ij}$ is the set of erroneous observations related to the data quality dimension D_i and variable Varj.

According to (5) and (6), the percentage $P_{EO_{+j}}$ and $P_{EO_{+j}}{}^{1}$ of erroneous observations for variable j are calculated as follows:

$$(7) \qquad P_{EO_{+j}} = \frac{NEO_{+j}}{NOBS}$$

$$(8) \qquad P_{EO_{+j}}{}^{1} = \frac{NEO_{+j}^{1}}{NOBS}$$

where NOBS is the total number of observations.

Table 4 presents the distribution of the number of erroneous observations (NEO) from data quality dimensions and variables in a symbolic form.

Table 4: NEO Matrix

	Group 1		Group 2			Group 3	
	Var1	...	Varj	...	...	Varn	Total
Syntactic Accuracy (D_1)	NEO_{11}	...	NEO_{1j}	...	...	NEO_{1n}	NEO_{1+}
Semantic Accuracy(D_2)	...	...	...	...	...	....	...
....	...		....			...	
Total	NEO_{+1}	...	NEO_{+j}	...	...	NEO_{+n}	NEO

Table 5 presents the percentage of detected erroneous observations for three variables in a real case study.

Table 5: Experimental results on detection of erroneous observations

	Percent of detected erroneous observations (intersection is not considered)			Percent of detected erroneous observations (intersection is considered)		
	Moody's Rating	S&P Rating	Market Currency	Moody's Rating	S&P Rating	Market Currency
Syntactic Accuracy	1.7	1.5	2.1	1.7	1.5	1.2
Semantic Accuracy	0	0.1	1.4	0	0.1	0.7
Internal Consistency	2.7	3.2	1.3	0	0	1.3
External Consistency	1.6	1.1	0.1	0	0	0.1
Incompleteness	3.5	5.5	8.1	3.5	5.5	7.1
Currency	0	0	0	0	0	0
Timeliness	8.6	9.2	2	7.3	8	1.1
Uniqueness	4.9	4.9	9.3	4.9	4.9	9.3
Total	23	25.5	24.3	17.4	20	20.9

4.4. Qualitative subjective assessment

The goal of this phase is the definition of a qualitative subjective assessment for the selected financial variables. The methodology proposes a qualitative subjective assessment obtained merging three independent assessments, as shown in figure 7. A business expert analyses data from a business point of view, a financial operator (e.g. a trader) uses daily a wide amount of data and a data quality expert analyses data with the aim to improve its quality.

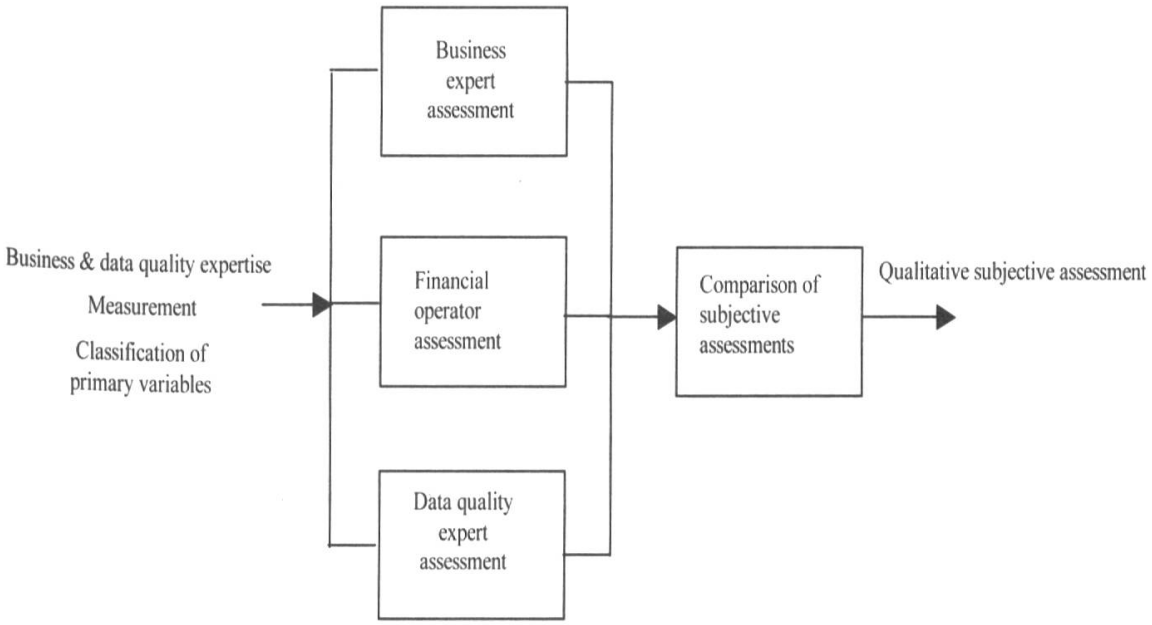

Figure 7: Schematic presentation of phase 4

Every subject matter expert, based on her/his own experience, marks out the Quality Level (QL) expected for each data quality dimension and variable as defined in Table 6 and 7. The percentages reported in Table 6 result from experiences performed in real life assessments.

Table 6: Definition of quality levels

Quality Level	***Related percentage of error***
Low (L)	More than 5%
Medium (M)	1%-5%
High (H)	0-1%

Table 7: Table of subjective quality assessments

	Group 1		Group 2			Group 3
	Var1	…	Varj	…	…	Varn
Syntactic Accuracy (D_1)	QL_{11}	…	QL_{1j}	…	…	QL_{1n}
Semantic Accuracy(D_2)	…	…	…	…	…	…
…	…	…	…	…	…	…
Total	QL_{m+}	…	QL_{m+}	…	…	QL_{m+}

The final QL assessment is the QL expressed by the majority of independent experts. For example, if two out of three experts express a low

QL, then the final QL is low. In the case of three different independent assessments, the worst-case scenario is considered (low QL). This is because when the experimental result is medium or high, this discrepancy is significant and it is worthwhile to investigate further.

Due to their importance in critical business processes, variables belonging to group 1 and 2 (see table 2 for the definition of the groups) should have a percentage of errors not greater than the low quality level. Variables belonging to group 3 should have a percentage of errors larger greater then the medium quality level.

As an example, in Table 8 we report the results of subjective qualitative assessments on three variables referring to rating and market currency, provided by three experts.

Table 8: Example of subjective qualitative assessment

	Moody's Rating	S&P Rating	Market Currency
Syntactic Accuracy	H	H	H
Semantic Accuracy	H	H	M
Internal Consistency	H	H	H
External Consistency	H	H	M
Incompleteness	L	L	M
Currency	H	H	H
Timeliness	M	M	H
Uniqueness	H	H	H
Total	H	H	H

4.5. Comparison between objective and subjective assessment

In the final phase of the methodology, objective and subjective assessment are compared. For each variable and quality dimension, we calculate the difference between the percentage of erroneous observations obtained from quantitative analysis, and the quality level defined by the judgement of the three experts. A possible quantitative outcome of the comparison is the following.

$$\Delta = \begin{cases} 1 & \text{if percentage of errors is greater then QL defined by the experts} \\ 0 & \text{if percentage of errors agrees with QL defined by the experts} \\ -1 & \text{if percentage of errors is less then QL defined by the experts} \end{cases}$$

Discrepancies (values of Δ different from 0) will drive the data quality expert to the correction of data errors. In particular, the data quality expert should carefully analyse values of Δ equal to 1 because positive discrepancies reveal an underestimation, by business experts, of data errors. More in general, every data quality dimension characterised by a positive discrepancy should be analysed in order to improve the data correctness. On the other hand, we do not consider that values of Δ equal to 1 indicate false positives of the objective technique because all the errors detected in the analysis phase are reproducible or documented.

A Δ equal to -1 indicates that:

a. the objective part of the methodology is not suitable for detection of data quality problems, or
b. the data quality expert has to verify that all the inspection techniques in the analysis phase have been successfully implemented and all the erroneous observations have been detected.

The following table presents the Δ results obtained comparing in our real life scenario quantitative results of Table 5 with qualitative results of Table 8.

Table 9: Examples of discrepancies in the real case scenario

	Δ values when intersection among quality dimensions is not considered			Δ values when intersection among quality dimensions is considered		
	Moody's Rating	S&P Rating	Market Currency	Moody's Rating	S&P Rating	Market Currency
Syntactic Accuracy	1	1	1	1	1	1
Semantic Accuracy	0	0	0	0	0	-1
Internal Consistency	1	1	1	0	0	1
External Consistency	1	1	-1	0	0	-1
Incompleteness	-1	0	1	-1	0	1
Currency	0	0	0	0	0	0
Timeliness	1	1	1	1	1	1
Uniqueness	1	1	1	1	1	1
Total	1	1	1	1	1	1

From the above table, we may notice that for all three variables the total percentage of detected errors has been underestimated by the subjective

assessments. The histogram of D values, in figure 8, shows that the number of discrepancies is greater when, in the counting of erroneous observations, the intersection among quality dimensions is not considered.

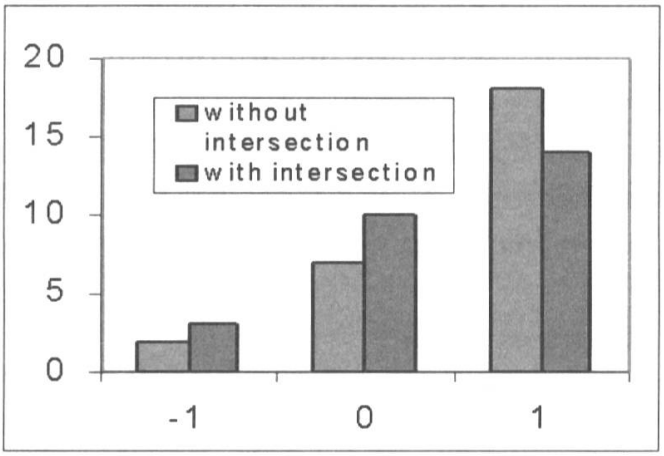

Figure 8: Histogram of Δ values.

5. Conclusion and Futher Research

In this paper we have proposed a methodology for the quantitative and qualitative data quality assessment of financial data. In the quantitative assessment, a variety of measures has been introduced in order to evaluate the data quality level for selected data quality dimensions. On the other hand, for the qualitative approach, the method proposed here is based on the assessments of three types of financial experts. Discrepancies in the results between quantitative and qualitative approaches are further investigated since their presence highlights possible problems in specific data quality dimensions or variables. Data quality analysis techniques proposed in the paper are easy to develop and do not require the usage of expensive solution packages. Although techniques proposed in the present work are the fruit of many years' experience in the data quality financial area, the procedures developed in the paper can also be applied to other typologies of data, providing ad-hoc adaptation.

Future research will focus on the establishment of a benchmark that should provide comparison with a reference universe, detection of best practices, and identification of areas where fruitful improvement efforts are required. Such a benchmark will be based on the development of data quality assessments of various financial databases, thus creating historical data on data quality with the aim of performing comparison for contexts with similar characteristics.

Acknowledgments

We thank Dr. Vasiliki Alexandrou, and Mr. Marco Beozzi for their contribution to the draft of this paper and for providing useful technical documentations.

References

Chung, K. L. (1974). Elementary Probability Theory with Stochastic Processes. New York: Springer-Verlag.

Dasu, T. & Johnson, T. (2003). Exploratory Data Mining and Data Cleaning. Wiley series in Probability and Statistics.London: Wiley.

Fugini, M.G. et al. (2002). Data Quality in Cooperative Web Information Systems. Technical Report. Dipartimento di Informatica e Sistemistica, Rome: Università La Sapienza.

Gower, J.C. (1971). A General Coefficient of similarity and some of its Properties. *Biometrics* 27: 857-74.

Kahn, B.K.; Strong, D.M. & Wang, R. Y. (2002). Information Quality Benchmarks: Product and Service Performance, *Communications of the ACM* 45.

Klein, B.D.; Goodhue, D.L. & Davis, G.B. (1996). Conditions for the Detection of Data Errors in Organizational Settings: Preliminary Results from a Field Study Proceedings of 1996 International Conference on Information Quality. Cambridge, MA: MIT.

Lee, Y.W. et al. (2001). AIMQ: A Methodology for Information Quality Assessment. *Information & Management.*

Matsumura, A & Shouraboura, N. (1996). Competing with Quality Information. Proceedings of 1996 International Conference on Information Quality. Cambridge, MA: MIT.

McKeon, A. (2003). Barclays Bank Case Study: Using Artificial Intelligence to Benchmark Organizational Data Flow Quality. Proceedings of the Eighth International Conference on Information Quality. Cambridge, MA: MIT.

Pipino, L.; Lee, Y. & Wang, R. (2002). Data Quality Assessement. *Communications of the ACM* 45.

Redman, T.C. (1996). Data Quality for the Information Age. Norwood, MA: Artech House.

Ross, R. (1994). The Business Rule Book: Classifying, Defining and Modeling Rules, Boston: Database Research Group.

Wang, R.; Storey, V. & Firth, C. (1997). A Framework for Analysis and Data Quality Research. *IEEE Transactions on Knowledge and Data Engineering* 7/4.

Zobel, J. & Dart, P. (1995). Finding Approximate Matches in Large Lexicons. *Software-Practice and Experience* 25/3: 331-345.