Data Quality in Agriculture Applications*

Joana E. Gonzales Malaverri[1], **Claudia Bauzer Medeiros**[1]

[1]Institute of Computing – State University of Campinas (UNICAMP)
13083-852 – Campinas – SP – Brasil

{jmalav09, cmbm}@ic.unicamp.br

Abstract. *Data quality is a common concern in a wide range of domains. Since agriculture plays an important role in the Brazilian economy, it is crucial that the data be useful and with a proper level of quality for the decision making process, planning activities, among others. Nevertheless, this requirement is not often taken into account when different systems and databases are modeled. This work presents a review about data quality issues covering some efforts in agriculture and geospatial science to tackle these issues. The goal is to help researchers and practitioners to design better applications. In particular, we focus on the different dimensions of quality and the approaches that are used to measure them.*

1. Introduction

Agriculture is an important activity for economic growth. In 2011, agricultural activities contributed approximately with 22% of Brazil's Gross National Product [CEPEA 2012]. Thus there are major benefits in ensuring the quality of data used by experts and decision makers to support activities such as yield forecast, monitoring and planning methods. The investigation of ways to measure and enhance the quality of data in GIS and remote sensing is not new [Chrisman 1984, Medeiros and de Alencar 1999, Lunetta and Lyon 2004, Congalton and Green 2009]. The same applies to data managed in, for instance, Information Manufacturing systems [Ballou et al. 1998]; Database systems [Widom 2005], Web systems [Hartig and Zhao 2009]; or Data Mining systems [Blake and Mangiameli 2011]. All of these fields are involved in and influence agriculture applications.

Despite these efforts, data quality issues are not often taken into account when different kinds of databases or information systems are modeled. Data produced and reported by these systems is used without considering the defects or errors that data contain [Chapman 2005, Goodchild and Li 2012]. Thus, the information obtained from these data is error prone, and decisions made by experts becomes inaccurate.

There are many challenges in ongoing data quality such as: modeling and management, quality control and assurance, analysis, storage and presentation [Chapman 2005]. The approach used to tackle each one of these issues depends on the application scenario and the level of data quality required for the intended use [U.S. Agency for International Development 2009]. Thus, understanding what attributes of quality need to be evaluated in a specific context is a key factor.

*Work partially financed by CNPq (grant 142337/2010-2), the Microsoft Research FAPESP Virtual Institute (NavScales project), CNPq (MuZOO Project and PRONEX-FAPESP), INCT in Web Science(CNPq 557.128/2009-9) and CAPES, as well as individual grants from CNPq.

This paper presents a brief review from the literature related to issues about data quality with special consideration to data managed in agriculture. The goal is to provide a conceptual background to become the basis for development of applications in agriculture.

2. Data for agriculture applications

Data in agriculture applications can be thematic/textual or geospatial, from primary to secondary sources, raw or derived. Thus, rather than just analyzing issues concerning the quality of geospatial data, this paper considers quality in all kinds of data, and provides guidelines to be applied for agriculture applications.

Research related to data quality in agriculture considers several issues. There are papers that concentrate on agricultural statistics data (e.g., production and consumption of crops) like [CountrySTAT 2012] and [Kyeyago et al. 2010]. The efforts that have been made to study the quality of geospatial data [FGDC 1998, ISO 19115 2003, Congalton and Green 2009, Goodchild and Li 2012] are also taken advantage of in the agriculture domain. However, there are other kinds of data that need to be considered such as files containing sensor-produced data, crop characteristics and soil information, human management procedures, among others [eFarms 2008].

This general scenario shows that agricultural activities encompass different kinds and sets of data from a variety of heterogeneous sources. In particular, the most common kinds of data are regular data and geospatial data. Regular data can be textual or numeric and can be stored on spreadsheets or text files (e.g., crop descriptions from official sources). Geospatial data correspond to georeferenced data sources and can include both raster and vector files, for example, satellite images using GeoTIFF format or a road network on shapefiles. Geospatial data may also come in data streams [Babu and Widom 2001] - packets of continuous data records - that can be obtained from aboard satellites, ground sensors or weather stations (e.g., temperature readings). All these data need different levels of access and manipulation and thus pose several challenges about data quality.

3. Dimensions of data quality

Data quality has various definitions and is a very subjective term [Chapman 2005]. A broad and consensual definition for data quality is "fitness for use" [Chrisman 1984]. Following this general concept, [Wang and Strong 1996] extended this definition as *data that are fit for use by data consumers*, i.e. those who use the data. [Redman 2001] complements the data quality concept by claiming that data are fit to be used if they are free of defects, accessible, accurate, timely, complete, consistent with other sources, relevant, comprehensive, provide a proper level of detail, and easy to read and interpret. Quality is context-based: often data that can be considered suitable for one scenario might not be appropriate for another [Ballou et al. 1998].

Data quality is seen as a multi-dimensional concept [Wang and Strong 1996, Ballou et al. 1998, Blake and Mangiameli 2011]. Quality dimensions can be considered as attributes that allow to represent a particular characteristic of quality [Wang and Strong 1996]. In particular, accuracy, completeness, timeliness and consistency have been extensively cited in the literature as some of the most important quality

dimensions to information consumers [Wang and Strong 1996, Parssian 2006]. Correctness, reliability and usability are interesting in areas like simulation modeling process, as discussed in [Scholten and Ten Cate 1999].

[Wang and Strong 1996] classified fifteen dimensions of quality grouped in four main categories - see Table 1(a). Dimensions accuracy, believability, objectivity and reputation are distinguished as *intrinsic data quality*. Timeliness and completeness are examples of *contextual data quality*. Interpretability and consistency describe features related to the format of the data and are classified as *representational data quality*. Accessibility and security are labeled as *accessibility data quality*, highlighting the importance of the role of information systems that manage and provide access to information.

Table 1.

(a) The 15 dimensions framework [Wang and Strong 1996]

Category	Dimensions
Intrinsic DQ	Believability
	Accuracy
	Objectivity
	Reputation
Contextual DQ	Value-added
	Relevancy
	Timeliness
	Completeness
	Appropriate amount of data
Representational DQ	Interpretability
	Ease of understanding
	Representational consistency
	Concise representation
Accessibility DQ	Accessibility
	Access security

(b) The PSP/IQ model [Lee et al. 2002]

	Conforms to Specifications	Meets or exceeds Consumer expectations
Product Quality	Free-of-error	Appropriate amount relevancy
	Concise representation	Understandability
	Completeness	Interpretability
	Consistent representation	Objectivity
Service Quality	Timeliness	Believability
	Security	Accessibility
		Ease of operation
		Reputation

The model of [Lee et al. 2002], Product Service Performance Information Quality (PSP/IQ), consolidates Wang and Strong's framework. Their goal is to represent information quality aspects that are relevant when decisions for improvement of information quality need to be made. Table 1(b) presents the PSP/IQ model showing that information quality can be assessed from the viewpoint of product or service and in terms of the conformance of data to the specifications and consumer expectations.

According to [Naumann and Rolker 2000] three main factors influence the quality of information: the user's perception, the information itself, and the process to retrieve the information. Based on these factors, the authors classify information quality criteria in 3 classes: *Subject-criteria*, *Object-criteria* and *Process-criteria*. Subject-criteria are those that can be determined by users' personal views, experience, and backgrounds. Object-criteria are specified through the analysis of information. Process-criteria are related to query processing. Table 2 shows their list of quality criteria grouped by classes, together with suggested assessment methods for each quality criterion.

USAID [U.S. Agency for International Development 2009] provides practical advices and suggestions on issues related to performance monitoring and evaluation. It highlights five quality dimensions: validity, reliability, precision, integrity, and timeliness.

In summary, the concept of quality encompasses different definitions and its dimensions (or attributes) can be generic or specific and this depends on the application

domain.

Table 2. The classification of [Naumann and Rolker 2000]

Class	Quality Criterion	Assessment Method
Subject Criteria	Believability	User experience
	Concise representation	User sampling
	Interpretability	User sampling
	Relevancy Continuous	User assessment
	Reputation	User experience
	Understandability	User sampling
	Value-Added	Continuous user assessment
Object Criteria	Completeness	Parsing, sampling
	Customer	Support Parsing, contract
	Documentation	Parsing
	Objectivity	Expert input
	Price	Contract
	Reliability	Continuous assessment
	Security	Parsing
	Timeliness	Parsing
	Verifiability	Expert input
Process Criteria	Accuracy	Sampling, cleansing techniques
	Amount of data	Continuous assessment
	Availability	Continuous assessment
	Consistent representation	Parsing
	Latency	Continuous assessment
	Response time	Continuous assessment

4. Data Quality Measurement

A significant amount of work addresses the measurement of the quality of data and information. The distinction between data and information is always tenuous. Although there is a tendency to use information as data that has been processed and interpreted to be used in a specific context - e.g., economics, biology, healthcare - data and information are often used as synonymous [Pipino et al. 2002]. According to [Naumann 2001], information quality measurement is the process of assigning numerical values, i.e. scores, to data quality dimensions. Related work differentiate between manual and automatic measurement of data quality. Manual approaches are based on the experience and users' point of view, i.e. a subjective assessment. Automatic approaches apply different techniques (e.g., mathematical and statistical models) in order to compute the quality of data. There follows an overview of work that investigates these topics.

4.1. Manual approaches

[Lee et al. 2002] measure information quality based on 4 core criteria to classify information: soundness, dependability, usefulness, and usability. Each class includes different quality dimensions. For instance, soundness encompasses: free-of-error, concise and consistent representation and completeness. The authors apply a survey questionnaire to the users to obtain scores for each criterion ranging from 0 to 1. The interpretation of the quality measure is made using gap analysis techniques. [Bobrowski et al. 1999] suggest a methodology also based on questionnaires to measure data quality in organizations. Quality criteria are classified as direct or indirect. Direct criteria are computed applying software metrics techniques and these are used to derive the indirect criteria.

While [Lee et al. 2002] and [Bobrowski et al. 1999] rely on questionnaires and users' perspective to obtain quality criteria scores, the methodology of [Pierce 2004] uses control matrices for data quality measurement. The columns in the matrix are used to list data quality problems. Rows are used to record quality checks and corrective processes. Each cell measures the effectiveness of the quality check at reducing the level of quality

problems. Similarly to [Lee et al. 2002] and [Bobrowski et al. 1999], this methodology also requires users' inputs to identify how well the quality check performs its function.

Volunteered geographic information (VGI) is a mechanism for the acquisition and compilation of geographic data in which members of the general public contribute with geo-referenced facts about the Earth's surface to specialist websites where the facts are processed and stored into databases. [Goodchild and Li 2012] outline three alternative solutions to measure the accuracy of VGI – crowd-sourcing, social, and geographic approaches.

The crowd-sourcing approach reflects the ability of a group of people to validate and correct the errors that an individual might make. The social approach is supported by a hierarchy of a trusted group that plays the role of moderators to assure the quality of the contributions. This approach may be aided by reputation systems as a means to evaluate authors' reliability. The geographic approach is based on rules that allow to know whether a supposed geographic fact is true or false at a given area.

4.2. Automatic approaches

Examples of work that use automatic approaches to measure data quality include [Ballou et al. 1998] and [Xie and Burstein 2011]. [Ballou et al. 1998] present an approach for measuring and calculating relevant quality attributes of products. [Xie and Burstein 2011] describe an attribute-based approach to measure the quality of online information resources. The authors use learning techniques to obtain values of quality attributes of resources based on previous value judgments encoded in resource metadata descriptions.

In order to evaluate the impact of data quality in the outcomes of classification - a general kind of analysis in data mining - [Blake and Mangiameli 2011] compute metrics for accuracy, completeness, consistency and timeliness. [Shankaranarayanan and Cai 2006] present a decision-support framework for evaluating completeness. [Parssian 2006] provides a sampling methodology to estimate the effects of data accuracy and completeness on relational aggregate functions (*count, sum, average, max*, and *min*). [Madnick and Zhu 2006] present an approach based on knowledge representation to improve the consistency dimension of data quality.

Although not always an explicit issue, some authors present the possibility to derive quality dimensions using historic information of data, also known as provenance. For instance, the computing of timeliness in [Ballou et al. 1998] is partially based on the time when a data item was obtained. Examples of work that have a direct association between quality and data provenance are [Prat and Madnick 2008], [Dai et al. 2008] and [Hartig and Zhao 2009]. [Prat and Madnick 2008] propose to compute the believability of a data value based on the provenance of this value. The computation of believability has been structured into three complex building blocks: metrics for measuring the believability of data sources, metrics for measuring the believability from process execution and global assessment of data believability. However, the authors only measure the believability of numeric data values, reducing the applicability of the proposal.

[Dai et al. 2008] present an approach to determine the trustworthiness of data integrity based on source providers and intermediate agents. [Hartig and Zhao 2009] present a method for evaluating the timeliness of data on the Web and also provide a

solution to deal with missing provenance information by associating certainty values with calculated timeliness values. Table 3 shows a summary with the quality dimensions studied in automatic approaches together with the application domain where the dimensions are considered.

Table 3. Summary of quality dimensions covered by automatic approaches

Work	Quality Dimension studied	Data managed by
[Ballou et al. 1998]	Accuracy and timeliness	Information Manufacturing System
[Shankaranarayanan and Cai 2006]	Completeness	Decision support system
[Parssian 2006]	Accuracy and completeness	Databases
[Madnick and Zhu 2006]	Consistency	Databases
[Prat and Madnick 2008]	Believability	Databases
[Dai et al. 2008]	Trustworthiness	Databases (data integrity)
[Hartig and Zhao 2009]	Timeliness	Web
[Xie and Burstein 2011]	Reputation	Web (Health Information Portals)
[Blake and Mangiameli 2011]	Accuracy, completeness, consistency and timeliness.	Databases

5. Data Quality in Applications in Agriculture

Considering the impact that agriculture has on the world economy, there is a real need to ensure that the data produced and used in this field have a good level of quality. Efforts to enhance the reliability of agricultural data encompass, for example, methodologies for collection and analysis of data, development of novel database systems and software applications.

Since prevention is better than correction, data collection and compilation are some of the first quality issues that need to be considered in the generation of data that are fit for use [Chapman 2005]. For instance, non-reporting data, incomplete coverage of data, imprecise concepts and standard definitions are common problems faced during the collection and compilation of data on land use [FAO 1997].

Statistical techniques and applications are being used to produce agricultural statistics such as crop yield production, seeding rate, percentage of planted and harvested areas, among others. One example is the [CountrySTAT 2012] framework. This is a web-based system developed by the Food and Agriculture Organization of the United Nations [FAO 2012]. It integrates statistical information for food and agriculture coming from different sources. The CountrySTAT is organized into a set of six dimensions of data quality that are: relevance and completeness, timeliness, accessibility and clarity, comparability, coherence, and subjectiveness.

Other example is the Data Quality Assessment Framework (DQAF) [International Monetary Fund 2003] that is being used as an international methodology for assessing data quality related to the governance of statistical systems, statistical processes, and statistical products. It is organized around a set of prerequisites and five dimensions of data quality that are: assurance of integrity, methodological soundness, accuracy and reliability, serviceability, and accessibility.

Based on both the CountrySTAT and the DQAF frameworks, [Kyeyago et al. 2010] proposed the Agricultural Data Quality Assessment Framework (ADQAF) aiming at the integration of global and national perspectives to

measure the quality of agricultural data. It encompasses quantifiable (e.g., accuracy and completeness) and subjective (e.g., relevance and clarity) quality dimensions.

Because of the relevance that land data plays in agriculture (e.g., for crop monitoring or planning for sustainable development), it is necessary to consider data quality issues in the development of agricultural land-use databases. According to [FAO 1997] the value of land-use databases is influenced by their accuracy, coverage, timeliness, and structure. The importance to maintain suitable geo-referenced data is also recognized.

Since agriculture applications rely heavily on geospatial data, one must consider geospatial metadata standards such as [ISO 19115 2003] and the [FGDC 1998], which have been developed aiming at the documentation and exchange of geospatial data among applications and institutions that use these kind of data. [ISO 19115 2003] defines a data quality class to evaluate the quality of a geospatial data set. Besides the description of data sources and processes, this class encompasses positional, thematic and temporal accuracy, completeness, and logical consistency. The FGDC metadata standard includes a data quality section allowing a general assessment of the quality of the data set. The main elements of this section are attribute accuracy, logical consistency report, completeness report, positional accuracy, lineage and cloud cover.

[Congalton and Green 2009] highlight the need to incorporate positional and thematic accuracy when the quality of geospatial data sets like maps are evaluated. Positional accuracy measures how closely a map fits its true reference location on the ground. Thematic accuracy measures whether the category labeled on a map at a particular time corresponds to the true category labeled on the ground at that time. According to [Goodchild and Li 2012] accuracy dimension is also an important attribute in the determination of quality of VGI. This approach is acquiring importance in all domains where non-curated data are used, including agriculture. Beyond accuracy, precision is also an important quality attribute that needs to be considered. [Chapman 2005] distinguishes statistical and numerical precision. The first one reflects the closeness to obtain the same outcomes by repeated observations and/or measurements. The last one reflects the number of significant digits with which data is recorded. It can lead to false precision values - e.g., when databases store and publish data with a higher precision than the actual value.

Completeness in the context of geospatial data encompasses temporal and spatial coverage [ISO 19115 2003, FGDC 1998]. Coverage reflects the spatial or temporal features for geospatial data. For instance, [Barbosa and Casanova 2011] use the spatial coverage dimension to determine whether a dataset covers (fully or partially) an area of interest.

Remote sensing is another major source of data for agriculture applications, in particular satellite or radar images. Image producers, such as NASA or INPE, directly or indirectly provide quality information together with images - e.g., dates (and thus timeliness), or coordinates (and thus spatial coverage). FGDC's cloud cover is an example of metadata field for images. Methodologies to measure quality of an image set combine manual and automatic processes (e.g., see [Moraes and Rocha 2011] concerning the cleaning of invalid pixels from a time series of satellite images, to analyze sugar cane yield). Information concerning the sensors aboard satellites is also used to derive quality information. Analogously, information concerning ground sensors is also taken into

account.

6. Summing up

We distinguish two groups of quality dimensions: qualitative and quantitative - see Table 4. We use the dimensions identified by [Wang and Strong 1996], since these authors are the most referenced in the literature.

Qualitative dimensions are those that need direct user interaction and their measurement is based on the experience and background of the measurer. This measurement can be supported by statistical or mathematical models [Pipino et al. 2002]. On the other hand, quantitative dimensions can be measured using a combination of computing techniques - e.g., machine learning, data mining - and mathematical and/or statistical models [Madnick et al. 2009]. For instance, simple ratios are obtained measuring the percentage of data items which meet with specific rules [Blake and Mangiameli 2011]. Parsing techniques consider how the information are structured in a database, in a document, etc [Naumann and Rolker 2000]. There are dimensions such as believability and accuracy that can be evaluated combining manual and automatic approaches. Choosing the best strategy for measuring the quality of data depends on the application domain and the dimensions of interest for that domain.

Table 4. Classification of quality dimensions

Dimensions of quality	Qualitative	Quantitative	Type of approach	Example of approach
Believability	x		Manual	user feedback
		x	Automatic	mathematical models
Objectivity	x		Manual	user feedback
Reputation	x		Manual	user experience
Value-added	x		Manual	user feedback
Relevancy	x		Manual	questionnaires
Interpretability	x		Manual	user experience
Ease of understanding	x		Manual	user feedback
Concise representation	x		Manual	user feedback
Accuracy	x		Manual	crowd-sourcing
		x	Automatic	cleansing techniques
Timeliness		x	Automatic	mathematical models
Completeness	x		Manual	control matrices
		x	Automatic	parsing
Consistent representation		x	Automatic	parsing
Access security		x	Automatic	mathematical models
Accessibility		x	Automatic	mathematical models
Appropriate amount of data		x	Automatic	mathematical models

Table 5 shows the most common quality dimensions investigated by research reviewed in the previous sections. We observe that the most frequent quality dimensions studied in the literature are accuracy, timeliness and completeness, followed by consistency and relevancy. Beyond these dimensions, accessibility is also of interest to agriculture field. This set of dimensions can become the basis to evaluate the quality of data in agricultural applications.

As we have seen, agricultural applications cover a wide variety of data. How to measure and enhance the quality of these data becomes a critical factor. It is important to adopt strategies and rules that allow to maintain the quality of data starting from the collection, consolidation, and storage to the manipulation and presentation of data. Common errors that need to be tackled are related to missing data, duplicate data, outdated data, false precision, inconsistency between datums and projections, violation of an organization's business rules and government policies, among others.

Table 5. Main data quality dimensions studied for the related work

Quality Dimension (QD)	Papers that studied these QD
Believability	[Prat and Madnick 2008]
Reputation	[Xie and Burstein 2011]
Reliability/Trustworthiness	[Dai et al. 2008], [Bobrowski et al. 1999] and [U.S. Agency for International Development 2009]
Relevancy	[CountrySTAT 2012], [Kyeyago et al. 2010], [FAO 1997] and [Bobrowski et al. 1999]
Ease of understanding	[Kyeyago et al. 2010]
Accuracy	[Ballou et al. 1998], [FGDC 1998], [ISO 19115 2003], [Parssian 2006], [Blake and Mangiameli 2011], [Kyeyago et al. 2010], [FAO 1997], [Bobrowski et al. 1999] and [Congalton and Green 2009].
Timeliness	[Ballou et al. 1998], [Hartig and Zhao 2009], [U.S. Agency for International Development 2009], [Blake and Mangiameli 2011], [CountrySTAT 2012], [FAO 1997] and [Bobrowski et al. 1999]
Completeness	[FGDC 1998], [ISO 19115 2003], [Shankaranarayanan and Cai 2006], [Parssian 2006], [CountrySTAT 2012], [Kyeyago et al. 2010], [Bobrowski et al. 1999] and [Barbosa and Casanova 2011].
Consistency	[FGDC 1998], [ISO 19115 2003], [Madnick and Zhu 2006], [Blake and Mangiameli 2011] and [Bobrowski et al. 1999]
Accessibility	[CountrySTAT 2012] and [Kyeyago et al. 2010]

Table 6 summarizes the main quality dimensions considered in agriculture, according to our survey. The table shows the dimensions that predominate in the literature and the context where they can be applied. It also shows that some dimensions include other quality attributes to encompass different data types - e.g., completeness for geospatial context is described in terms of spatial and temporal coverage. We point out that most dimensions are common to any kind of application. However, like several other domains, agriculture studies require analysis from multiple spatial scales and include both natural factors (e.g., soil or rainfall) and human factors (e.g., soil management practices). Moreover, such studies need data of a variety of types and devices. One of the problems is that researchers (and often practitioners) concentrate on just a few aspects of the problem.

For instance, those who work on remote sensing aspects seldom consider ground-based sensors; those who perform crop analysis are mainly concerned with biochemical aspects. However, all these researchers store and publish their data. Correlating such data becomes a problem not only because of heterogeneity issues, but also because there is no unified concern with quality issues and the quality of data is seldom made explicit when data are published. This paper is a step towards trying to minimize this problem, by pointing out aspects that should be considered in the global view. As mentioned before, these issues are not unique to agriculture applications and can be found in, for instance, biodiversity or climate studies.

Table 6. Main data quality dimensions in agriculture applications

Quality dimensions	Context	Example of kinds of data
Accuracy:	Relational databases, statistical information and data files	table, tuple, attribute, query, yield information, production of crops, growth rate, XML files, spreadsheets documents, etc.
Positional and Thematic accuracy	Geospatial datasets	geographic coordinates, VGI, satellite images, maps, aerial photography, etc.
Completeness:	Relational databases, statistical information and data files	schema, column, attribute, population census, land data, rates of harvested areas, farm production, CVS text files, spreadsheets, etc.
Spatial and Temporal coverage	Geospatial datasets	cartographic materials, geographic coordinates, etc.
Timeliness	Information Manufacturing systems	age and shelf life of products, delivery time of products, etc.
	(Geographic) Information/Web systems and statistical information	access, creation or delivery time of data items, age of a data item, sensor data streams, population census, harvest dates, etc.
Consistency	(Geospatial) Databases	tables, data, maps, time series, reports and charts, etc.
Relevancy	Information systems, databases and statistical information	text and spreadsheets documents, census, historical weather datasets, trade information, etc.

References

Babu, S. and Widom, J. (2001). Continuous queries over data streams. *SIGMOD Rec.*, 30(3):109–120.

Ballou, D., Wang, R., Pazer, H., and Tayi, G. K. (1998). Modeling Information Manufacturing Systems to Determine Information Product Quality. *Manage. Sci.*, 44:462–484.

Barbosa, I. and Casanova, M. A. (2011). Trust Indicator for Decisions Based on Geospatial Data. In *Proc. XII Brazilian Symposium on GeoInformatics*, pages 49–60.

Blake, R. and Mangiameli, P. (2011). The Effects and Interactions of Data Quality and Problem Complexity on Classification. *J. Data and Information Quality*, 2:8:1–8:28.

Bobrowski, M., Marré, M., and Yankelevich, D. (1999). A Homogeneous Framework to Measure Data Quality. In *Proc. IQ*, pages 115–124. MIT.

CEPEA (2012). Center of Advanced Studies in Applied Economics. http://cepea.esalq.usp.br/pib/. Accessed in June 2012.

Chapman, A. D. (2005). Principles of Data Quality. *Global Biodiversity Information Facility, Copenhagen*.

Chrisman, N. R. (1984). The Role of Quality Information in the Long-term Functioning of a Geographic Information System. *Cartographica*, 21(2/3):79–87.

Congalton, R. G. and Green, K. (2009). *Assessing the accuracy of remotely sensed data: principles and practices*. Number 13. CRC Press, Boca Raton, FL, 2 edition.

CountrySTAT (2012). Food and Agriculture Organization of the United Nations. www.fao.org/countrystat. Accessed on March 2012.

Dai, C., Lin, D., Bertino, E., and Kantarcioglu, M. (2008). An Approach to Evaluate Data Trustworthiness Based on Data Provenance. In *Proc. of the 5th VLDB Workshop on Secure Data Management*, pages 82–98, Berlin, Heidelberg. Springer-Verlag.

eFarms (2008). http://proj.lis.ic.unicamp.br/efarms/. Accessed in June 2012.

FAO (1997). *Land Quality Indicators and Their Use in Sustainable Agriculture and Rural Development*. FAO Land and Water Bulletin. Accessed in January 2012.

FAO (2012). Food and Agriculture Organization of the United Nations. http://www.fao.org/. Accessed on March 2012.

FGDC (1998). Content Standard for Digital Geospatial Metadata FGDC-STD-001-1998. Technical report, US Geological Survey.

Goodchild, M. F. and Li, L. (2012). Assuring the quality of volunteered geographic information. *Spatial Statistics*, 1:110–120.

Hartig, O. and Zhao, J. (2009). Using web data provenance for quality assessment. In *Proc. of the Workshop on Semantic Web and Provenance Management at ISWC*.

International Monetary Fund (2003). Data Quality Assessment Framework. http://dsbb.imf.org/. Accessed on January 2012.

ISO 19115 (2003). Geographic information – Metadata. http://www.iso.org/iso/. Accessed on January 2012.

Kyeyago, F. O., Zake, E. M., and Mayinza, S. (2010). In the Construction of an International Agricultural Data Quality Assessment Framework (ADQAF). In *The 5th Int. Conf. on Agricultural Statistics (ICAS V)m*.

Lee, Y. W., Strong, D. M., Kahn, B. K., and Wang, R. Y. (2002). AIMQ: a methodology for information quality assessment. *Information & Management*, 40(2):133–146.

Lunetta, R. S. and Lyon, J. G. (2004). *Remote Sensing and GIS Accuracy Assessment*. CRC Press.

Madnick, S. and Zhu, H. (2006). Improving data quality through effective use of data semantics. *Data Knowl. Eng.*, 59:460–475.

Madnick, S. E., Wang, R. Y., Lee, Y. W., and Zhu, H. (2009). Overview and Framework for Data and Information Quality Research. *J. Data and Information Quality*, 1:2:1–2:22.

Medeiros, C. B. and de Alencar, A. C. (1999). Data Quality and Interoperability in GIS. In *Proc. of GeoInfo*. In portuguese.

Moraes, R. A. and Rocha, J. V. (2011). Imagens de coeficiente de qualidade (Quality) e de confiabilidade (Reliability) para seleção de pixels em imagens de NDVI do sensor MODIS para monitoramento da cana-de-açúcar no estado de São Paulo. In *Proc. of Brazilian Remote Sensing Symposium*.

Naumann, F. (2001). From Databases to Information Systems - Information Quality Makes the Difference. In *Proc. IQ*.

Naumann, F. and Rolker, C. (2000). Assessment Methods for Information Quality Criteria. In *IQ*, pages 148–162. MIT.

Parssian, A. (2006). Managerial decision support with knowledge of accuracy and completeness of the relational aggregate functions. *Decis. Support Syst.*, 42:1494–1502.

Pierce, E. M. (2004). Assessing data quality with control matrices. *Commun. ACM*, 47:82–86.

Pipino, L. L., Lee, Y. W., and Wang, R. Y. (2002). Data Quality Assessment. *Commun. ACM*, 45:211–218.

Prat, N. and Madnick, S. (2008). Measuring Data Believability: A Provenance Approach. In *Proc. of the 41st Hawaii Int. Conf. on System Sciences*, page 393.

Redman, T. C. (2001). *Data quality : The Field Guide*. Digital Pr. [u.a.].

Scholten, H. and Ten Cate, A. J. U. (1999). Quality assessment of the simulation modeling process. *Comput. Electron. Agric.*, 22(2-3):199–208.

Shankaranarayanan, G. and Cai, Y. (2006). Supporting data quality management in decision-making. *Decis. Support Syst.*, 42:302–317.

U.S. Agency for International Development (2009). TIPS 12: Data Quality Standards. http://www.usaid.gov/policy/evalweb/documents/TIPS-DataQualityStandards.pdf. Accessed in January 2012.

Wang, R. Y. and Strong, D. M. (1996). Beyond accuracy : What data quality means to data consumers. *Journal of Management Information Systems*, 12(4):5–34.

Widom, J. (2005). Trio: A System for Integrated Management of Data, Accuracy, and Lineage. In *Proc. of the 2nd Biennial Conf. on Innovative Data Systems Research (CIDR)*.

Xie, J. and Burstein, F. (2011). Using machine learning to support resource quality assessment: an adaptive attribute-based approach for health information portals. In *Proc. of the 16th Int. Conf. on Database Systems for Advanced Applications*.