Quality Assessment of Big Data with GIS

Abstract

Keywords: Information Society, Data Quality Assessment, Big Data, GIS, Fitness for use, Data Reference Quality

The increase of available data and more users with different needs influence the approach regarding data quality. In this paper, we propose an integration of Big Data with data quality assessment. This Big Data Quality Assessment Model (BDQAM) is illustrated by two cases of Infofolio where GIS functionalities play an important role. We conclude that although the approach is promising, more research is needed regarding the data quality dimensions of Big Data and the relationship between GIS and Big Data.

1 Introduction

The information society develops rapidly. More and more data are collected via network connected sensors. New applications are introduced [KNAW, 2016], quite often in a disruptive way. This development, called 'Big Data' refers not only interpretation of large data sets, but also smart combining and fusion of different data sources [Klous et al, 2016].

To use Big Data in a responsible and accessible manner, aspects as transparency, governance, privacy and datafication should be discussed.

Datafication refers to the collective tools, technologies and processes used to transform an organization into a data-driven enterprise. Datafication pays attention to data quality assessment and assurance. Big Data consist of a large quantity of (semi-)structured and non-structured data from different sources that can be combined in many applications. The usage in an uncontrolled manner makes it difficult to assure the quality. Discrepancy is possible between the original data sources and the desired 'fitness for use' quality in working processes and applications of the end-users.

The 'fitness for use' principle will be more difficult to assess and assure if many different data sources are used. Another challenge is to describe an objective 'fitness for use' data quality level for the end-users, considering that the end-users of the combined data sources are not the same users as those of the original data sources.
So not only the original owner of the data is responsible, but also all the stakeholders involved in upgrading the data towards an objective level of data quality for the end-users. This objective level of data quality is defined as the data reference quality.

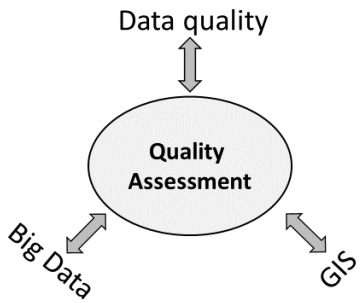
This paper describes a new model how the process of data quality assessment can be organized from the start with different data sources to the data reference quality for the end-users. GIS functionalities will be used in this process of data quality assessment and give us special functionalities by working out the data quality assessment tools.

First, this paper discusses the theoretical background of the terms 'data quality', 'big data' and 'GIS'. Based on the characteristics and possibilities of these terms and other studies the new model 'Big Data Quality Assessment Model' [Jellema, De Bakker, 2016] is introduced.
Finally, this new model will be applied to the daily data process of Infofolio [Infofolio, 2017]. Making use of the possibilities of Big Data and multiple linear regression analysis [Jellema et al, 2015], the data from over more than 50 different data sources are collected, analysed and made usable, in more than 150 data attributes on address-level of almost 9 million buildings, for organisations in the insurance-, risk- and safety-branches.

2 Theoretical background

For this paper the term quality assessment is related to the terms 'data quality', 'Big Data' and 'Geographical Information Systems: GIS'. In this chapter, each term will be described, including its relation with the process of data quality assessment. In figure 1 the terms are visualised.

Figure 1. Terms related to Quality Assessment in this paper



2.1 Data Quality Dimensions

Since the 1950s researchers have begun to study quality issues, especially the quality of products and a series of definitions. Later, with the rapid development of information technology, research turned to study of the data quality. Many universities and institutions have undertaken research into the study of data quality [Cai et al, 2015].
For example, the Total Data Quality Management group of MIT University has done in-depth research in data quality. They defined 'data quality' as 'fitness for use' and proposed that data quality judgement depends on data consumers [Wang et al, 1996].

As result of the research into data quality there have been also many studies into data quality dimensions and data quality frame-works. In 2013 a DAMA UK Working Group defined the following six best practice definitions as generic data quality dimensions:

1. Completeness: The proportion of stored data against the potential of '100% complete'.
2. Uniqueness: No object will be recorded more than once based upon how that object is identified.
3. Timeliness: The degree to which data represent reality from the required point in time.
4. Validity: Data are valid if it conforms to the syntax (format, type, range) of its definition.
5. Accuracy: The degree to which data correctly describes the 'real world' object or event being described.
6. Consistency: The absence of difference, when comparing two or more representations of an object against the definition.

The DAMA UK Working Group suggested that these dimensions and definitions should be adopted by data quality practitioners as the standard method for assessing and describing the quality of data.

These six generic data quality dimensions, related to the 'fitness for use' principle, are useful for each data source because the end-user of every data source is well known.

For validating the data quality outcomes of the different data sources these generic data quality dimensions can be used by defining the objective data reference quality. The data reference quality describes the optimal data quality of every data attribute, considering the possibilities of the original data sources.
The outputs of different data quality checks, related to the data reference quality, may be required to determine how well every data attribute supports the needs of the end-user, in accordance with the 'fitness for use' principle for each end-user.

2.2 Characteristics of Big Data

The disruptive development of Big Data generates an enormous increase of data and presents new features. Big Data can be defined by the following characteristics [Katal, 2013]:

- Variety: Data being produced is not of single category as it not only includes the traditional data but also the semi structured data from various resources. All this data is totally different consisting of raw, structured, semi structured and even unstructured data.
- Volume: The Big word in Big Data itself defines the volume. At present the data existing is in petabytes and is supposed to increase to zettabytes in nearby future.
- Velocity: Velocity in Big Data is a concept which deals with the speed of the data coming from various sources. This characteristic is not being limited to the speed of incoming data but also speed at which the data flows.
- Variability: Variability considers the inconsistencies of the data flow.
- Complexity: It is quite an undertaking to link, match, cleanse and transform data across systems coming from various sources. It is also necessary to connect and correlate relationships, hierarchies and multiple data linkages.
- Value: The user can run queries on the Big Data for his own applications and can deduct results from the filtered data obtained.

All the characteristics of Big Data have a greater or lesser effect on the area of data quality. A consequence is that the data reference quality can not only be described by the generic data quality dimensions (§ 2.1), but also needs attention to the characteristics of Big Data. We call this the Big Data Quality Dimensions.

2.3 GIS functionalities

There are many ways to classify the analytic functions of a GIS. In this paper the classification in four classes is used from the study Principles of Geographic Information Systems [By de, R.A. et al., 2001].

1. Retrieval, classification, and measurement functions:
   These GIS functions allow exploring the data without making fundamental changes, and therefore they are often used at the beginning of data analysis. Measurement functions include computing distances between objects and the computation of area size of 2D, 3D objects or volume size. Counting, to understand frequency of objects, is also included. Spatial queries retrieve objects selectively, using quality dimensions

defined conditions. Classification means the controlled (re)assignment of a data attribute value in a dataset.

2. Overlay functions: This group of GIS functions forms the core computational activity of many GIS applications and Business Intelligence tools. Data layers are combined and new information (or non-information?) is derived, usually by creating objects in a new layer.
3. Neighbourhood functions: Whereas overlays combine objects at the same location, neighbourhood functions evaluate the characteristics of an area surrounding a location of an object. This allows looking at buffer zones around objects.
4. Connectivity functions: These functions evaluate how objects are connected. This is useful in applications dealing with networks of connected objects. This allows looking at the connection of objects.

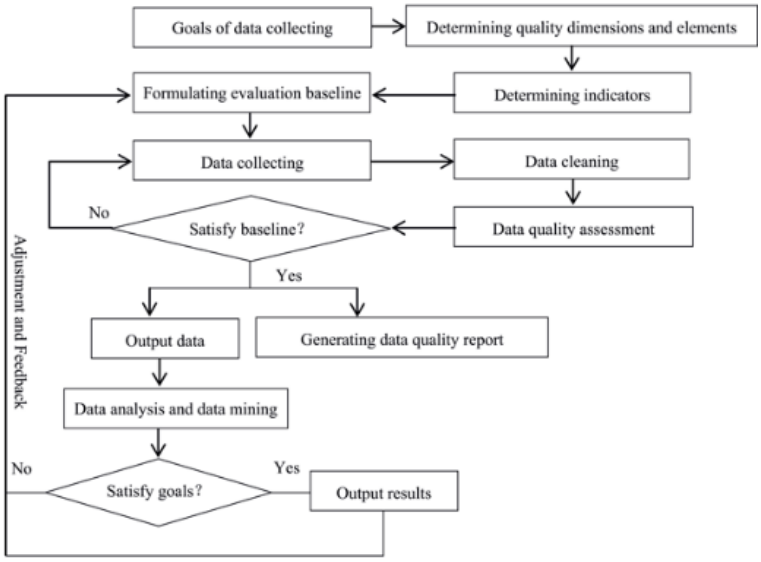
All the GIS functionalities can be applied in the all the different steps of the process of data quality assessment. Chapter 4 of this paper gives some examples of GIS-applications related to the quality assessment.

3 Big Data Quality AssessmentModel (BDQAM)

There are a lot of quality assessment models using different data quality dimensions. Most of these tools are developed and suitable for the traditional structured data. Nowadays, the challenge is to develop a data quality assessment model which is applicable for Big Data.

Figure 2 shows a data quality assessment process for Big Data with a dynamic feedback mechanism based on the characteristics of Big Data [Cai et al, 2015].

Figure 2. Quality assessment process for Big Data



Source: Cai et al, 2015

Per the purpose of the process, the quality assessment process for Big Data is mainly focused on collecting, cleaning and analysing the different datasets. Strictly speaking, data analysis and data mining do not belong to the scope of Big Data quality assessment, but they play an important role in the dynamic adjustment and feedback of data quality assessment.

The visualised quality assessment process for Big Data starts from the perspective of the end-users by with defining the Goals of data collecting and determining the quality dimensions and elements [Cai et al, 2015].

When we want to use more different data sources and the end-users are not the same users of the original data source, we advise not to use directly the 'fitness for use' principle but defining the data reference quality based on the Big Data quality dimensions.
When it is possible to develop a connection between the quality assessment process for Big Data and the data reference quality, we suppose there will be a good foundation for further study of the data quality assessment process for Big Data.

A first quality assessment process for Big Data based on this assumption is shown in Figure 3.

This new 'Big Data Quality Assessment Model' is a further development of Figure 2. with the following aspects:

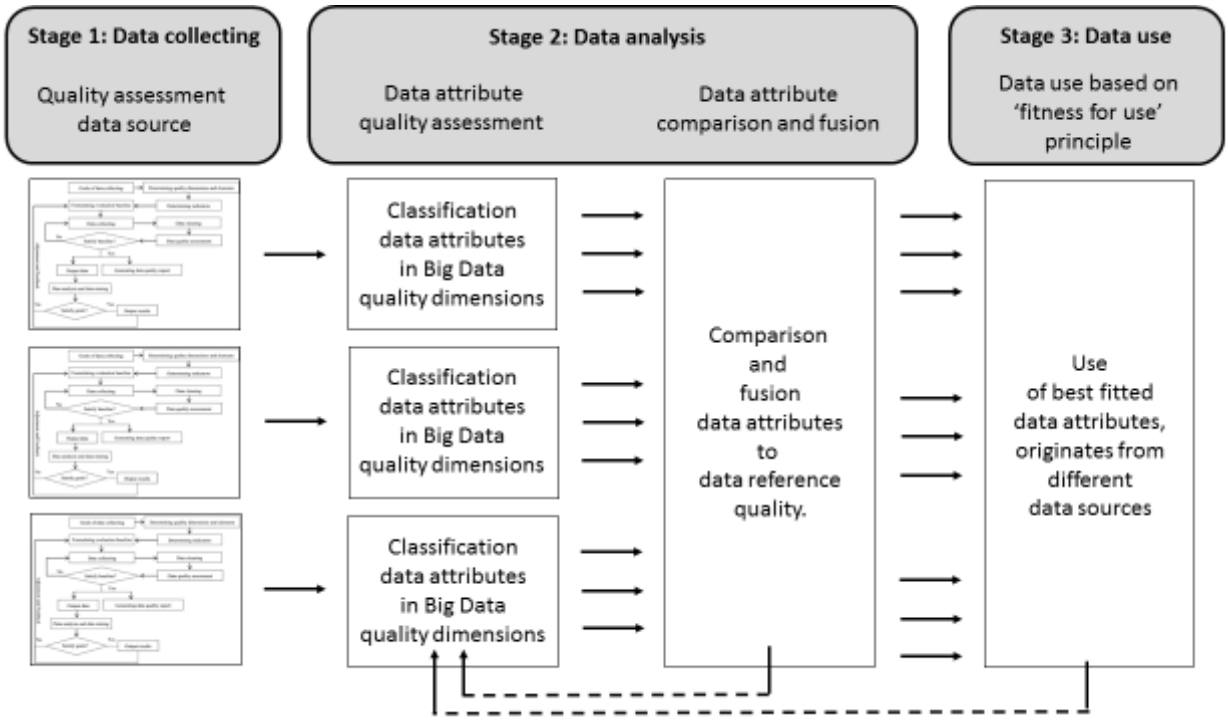
- The model divided the Big Data processes in a three-stage procedure: data collecting, data analysis and data use [WRR, 2016]. The process from figure 2 takes place

in the first stage procedure data collecting. Without interference from any other dataset, the quality assessment process is executed for every individual dataset.

- In the second stage procedure data analysis, the outcome of every dataset is classified per data attribute in the Big Data quality dimensions and comparison is possible per comparable data attribute of the different datasets. After the quality assessment on the level of data attributes, the outcome will be related to the data reference quality of each data attribute.
- In the third stage procedure data use, combination and fusion of the best fitted data attributes, originates from different datasets, lead to information for the end-user based on the 'fitness for use' principle.

The new model makes in all stage procedures use of the dynamic feedback mechanism based on the characteristics of Big Data. The well-known 'quantity over quality' principle is ignored because in this model the quality of data is more important than the quantity of data.

Figure 3. Big Data Quality Assessment Model (BQDAM)



Source: Jellema, De Bakker, 2016

4 Infofolio-applications of BDQAM

By introducing, investigating, learning and working with the 'Big Data Quality Assessment Model', Infofolio follows the Theory of Change approach [Weiss, 1995]. This indicates that the next two cases are used for further development of the model. The first case is focussed on the data attribute 'volume of a building'. The second case gives a clear view on how the GIS functionalities are used.

4.1 Case: Volume of a building [Bakker, M. de et al, 2015]

The volume of a building is quite often measured with different methods. Eight different data sources with different coverage and quality can be used. One source 'pointclouds', derived from aerial photos with a resolution of 7.5 cm for this attribute of a building is discussed in more detail. The BAG (National Dutch Base Registration Addresses and Buildings, 2017) is used for the footprint of a building with the attribute residential.

In stage 1 quality of the individual datasets is assessed. Aerial photos were good enough to deliver a Digital Elevation Model (DEM) (10 cm grid) with inverse distance interpolation (example of neighbourhood GI functionalities). Quality of the DEM was assessed with the measured points and indicated a 95% similarity. The accuracy of the footprint of the BAG was different in each municipality, but mostly in the range of 5 cm geometric accuracy. Volume of the building was calculated by vacuuming the building from the outside by comparing the points inside the footprint with the level of the surface.

In stage 2 we compared the calculated volume with the volume as registered in the other eight data sources e.g. Valuation of Immovable Property Act [WOZ]. Although the definition of the volume is different (often it excludes internal floors and walls, so it is net volume instead of gross volume) the regression analysis between the pointcloud volume and seven other data sources was 94%. In table 1 we show the classified results for 2 dimensions.

Table 1. Overview results Quality assessment

Stage:	**1**	**1**	**3**
Data quality dimension	**Pointcloud**	**WOZ**	**All sources combined**
Actuality	Low (1 year old)	High (2 months)	Medium, but for some uses good enough
Completeness	High (every building → 100%)	Low (sometimes only 98%)	High, give insight in missing values

Source: Jellema, De Bakker, 2016

Overall the reference quality was visualized as indicated in figure 4 (from red: low to green: high quality).

Figure 4. Data reference quality of volume

Source: Bakker, M. de et al, 2015

It seems that in this case the comparison and fusion of the 8 different data sources together deliver a higher 'fitness for use' than the data of the data source pointcloud.

4.2 Case: GIS functions and data reference quality

Infofolio describes the data reference quality with 8 Big Data Quality Dimensions (§ 2.2). Nowadays, Infofolio makes use of GIS functionalities to analyse 4 of 8 Big Data Quality Dimensions of the defined data reference quality.

Each of these four dimensions will be described including an example of the GIS application.

- *Completeness*: GIS functions are used in stage 1 and 2 to give insight in the proportion of the available data against the potential completeness of the data attribute. The potential completeness is not always the total of buildings in The Netherlands. For example, not every building has the destination of a monument. In stage 3 the GIS functions are used to give the end-users insight in the completeness of every data attribute related to working area of the end-user.
- *Timeliness*: GIS functions are applied to analyse the difference between the stored data (stage 3) and the new incoming data (stage 2). Measurement, retrieval, classification and connectivity functions are used to recognize and analyse all new, modified and no longer existing data attributes.
- *Accuracy*: The first example (§ 4.1) describes how GIS functions are used in stage 1 and 2 for this dimension accuracy. Measurement, neighbourhood and connectivity functions are for example also applied by the data attributes: surface, type of building, type of building, date of construction, location-based risk data. The overlay-function is used to derive new risk and environmental information on districts-level.
- *Consistency*: Different GIS functions are used to control and analyse the consistency between different data attributes. For example, there is a (statistical) consistency between surface, volume, destination and number of rooms in the building.

5 Conclusion

The proposed Big Data Quality Assessment Model delivers as shown by the cases a better insight in the process of data quality assessment of Big Data. The approach with different stages including the definition of data reference quality and 'fitness for use' makes the whole process more transparent for all stakeholders involved.

The integration of GIS delivers in all stages of the model several methods in order to assess the quality of data in relationship with the usage of the Big Data and information for the end-users.

In such a way, Big Data presents the usage of the relevant data with a quality label.

6 Acknowledgement

The authors like to thank Anne Muller for her review of our English.

7 References

BAG, (2017) https://www.kadaster.nl/bag

Bakker, M. de, Voets D., Jellema, M., Bozelie, W., (2015) *Innovatie in kwaliteit door combinatie van puntenwolken en vastgoed-informatie*, Geo-Info 2015-02, pp 18 -20 The Netherlands.

By de, R.A. et al (2001) *Principles of Geographic Information Systems*, An introductory textbook, ITC, The Netherlands.

Cai, L., Zhu, Y. (2015) *The Challenges of Data Quality and Data Quality Assessment in the Big Data Era*, Data Science Journal 14.

DAMA UK Working Group *'Data Quality Dimensions'*, (2013) Defining Data Quality Dimensions, final paper.

Infofolio (2017) www.infofolio.nl

Jellema, M., Autar, A. (2015) *Hermes-model: Universal Model to Estimate The Rebuilding Costs of Houses*, FIG Working Week 2015, Sofia Bulgaria.

Katal, A., Wazid, M., Goudar, R. (2013) *Big Data: Issues, Challenges, Tools and Good Practices*, Procedures of the 6th International Conference on Contemporary Computing, Noida India.

Klous, S., Wielaard, N. (2016) *Wij zijn Big Data*, Uitgeverij Business Contact, Amsterdam/Antwerpen.

KNAW (2016) *Nationale Wetenschapsagenda Route 9: Toegankelijke en verantwoorde waarde creatie uit Big Data*.

Wang, R.Y., Strong, D.M. (1996) *Beyond Accuracy: What Data Quality Means to Data Consumers*, Journal of Management Information Systems 12(4), pp 5-33.

Weiss, C. (1995) *Nothing as practical as good theory*.

WOZ, (2017), https://www.kadaster.nl/woz-waarde

Wetenschappelijke Raad voor het Regeringsbeleid (WRR) (2016) *Big Data in een vrije en veilige samenleving*, WRR-rapport nr. 95.