TOWARDS A VISUAL APPROACH TO AGGREGATE DATA QUALITY MEASUREMENTS

(Research-in-Progress)

Tom Haegemans
KU Leuven, Belgium
tom.haegemans@kuleuven.be

Michael Reusens
KU Leuven, Belgium
michael.reusens@kuleuven.be

Bart Baesens
KU Leuven, Belgium
bart.baesens@kuleuven.be

Wilfried Lemahieu
KU Leuven, Belgium
wilfried.lemahieu@kuleuven.be

Monique Snoeck
KU Leuven, Belgium
monique.snoeck@kuleuven.be

Abstract Data quality measurements are often aggregated in such ways that only one small aspect of the measurements is conveyed (e.g. the average of the measurements). However, to gain insight into the root causes of data deficiencies such as errors or missing values, stakeholders typically need information about many other aspects like, for example, how the data deficiencies are distributed. As such, the goal of this research is to develop a new approach able to visually represent objective data quality measurements with respect to the data to which these measurements correspond. This is accomplished by generating an embedding based on a selection of attributes and projecting data quality measurements on this embedding using colourings. This study contributes to the work on data quality by providing a novel method to aggregate data quality measurements at the level of data records or data items in a visual way regardless of the type of scale on which the data quality measurements were made.

Keywords data quality, metrics, measurement, visualisation

1 INTRODUCTION

Many datasets contain data items that are of imperfect quality: they contain data deficiencies such as errors, missing values, or inconsistencies (see e.g. Haegemans, Snoeck, Lemahieu, Stumpe, & Goderis, 2016; Weiskopf & Weng, 2013; Espetvedt, Reksen, Rintakoski, & Osterås, 2013; DeHoratius & Raman, 2008; Thiru, Hassey, & Sullivan, 2003; Arts, De Keizer, & Scheffer, 2002; Goldhill & Sumner, 1998). Several aspects of these data deficiencies are measurable and are of great interest to data quality stakeholders. For

example, in the case of errors in data, stakeholders might be interested in the size or occurrence of errors (Haegemans, Snoeck, & Lemahieu, 2016). Likewise, when considering missing values, stakeholders might simply be concerned about whether the value of the data item is missing, or might find it useful to discriminate between the values that are missing because the value does not exist in real world or simply because the data was not recorded (van der Meyden, 1999). Similarly, in the case of data inconsistencies, one could be interested to what degree the data item matches a predefined regular expression (see e.g. Bronselaer, Nielandt, De Mol, & De Tré, 2016) or simply whether the data item is consistent or not.

One of the reasons that measurement information about specific aspects of data deficiencies like in the examples above can be of interest to data quality stakeholders is because it can be used to identify the root causes of such deficiencies (Wang, 1998, p. 64). Eliminating the root causes of data quality issues is the most efficient way to improve the quality of data. However, in order for data quality measurement information to effectively serve the purpose of root cause analysis, two properties are desirable. The first desirable property is that the aggregation of multiple individual data quality measurements of single data items should be presented with a minimal loss of detail. For example, simple statistics, such as the average size of errors, do not retain much measurement information and might raise questions like "Are there many large errors, or are there many small errors and only some very large errors?". The second desirable property is that the information should indicate whether the distribution of the data deficiencies occurs randomly or systematically. That is, when deficiencies are distributed in a systematic way, it is probable that they share a common cause and thus hint towards the direction of the root cause that led to the data deficiency. As we will demonstrate, when the information of data quality measurements possesses both properties, it can aid in decisions about which data needs improvement and where one should start looking for the root causes of poor quality data.

To the best of our knowledge, in the data quality literature, only a small number of attempts have been made to aggregate multiple individual data quality measurements of single data items or to present data quality measurement information so that the distribution of the deficiencies becomes apparent. Pipino, Lee, and Wang (2002) identified three functional forms, or statistics, by which data quality measurements can be aggregated. Subsequently, Fisher, Lauria, and Matheus (2009) proposed to use the Lempel-Ziv algorithm to indicate the randomness or complexity of the occurrence of errors with respect to their location (e.g. record) in a dataset. Yet, none of these attempts presents data quality measurement information in such a way that the individual deficiencies can be identified and linked to the distribution of the values of the data itself.

In the recommender systems literature, we have recently encountered a similar research problem (Reusens et al., 2017): certain users of a recommendation system were receiving erroneous (irrelevant) recommendations while other users were receiving high quality (relevant) recommendations and we did not know which characteristics of the users triggered this problem. To gain insight in this matter, we proposed the following approach: all users receiving recommendations were projected on a two dimensional embedding that was created by reducing the dimensionality of the features that provide information about the users of the recommendation service. As such, business users, like marketeers, can gain a better understanding about which users were being supported by a recommender system, and which users were not. The research problem encountered in the recommender systems literature is similar to this one in the sense that irrelevant recommendations are, in a way, defective data items and that the information about the users of a recommendation service is a collection of attributes related to the root cause of poor quality recommendations.

Consequently, in this paper we adapt the approach developed in the context of recommender systems (Reusens et al., 2017) to the context of data quality with the purpose of visualising the distribution of data deficiencies with respect to a collection of attributes. We validate the approach in a preliminary way by demonstrating how it could provide utility for practitioners in a real-world setting (Hevner, March, Park, & Ram, 2004, p. 86).

2 TERMINOLOGY: DATA QUALITY DIMENSIONS, MEASURES AND DEFICIENCIES

Data quality is a multidimensional concept (Zmud, 1978; Wang, Reddy, & Kon, 1995; Ballou & Tayi, 1999; Moges, Dejaeger, Lemahieu, & Baesens, 2013). Examples of data quality dimensions are data accuracy, data completeness and data consistency (Wang & Strong, 1996). Quality dimensions are often categorised according to the ability to measure them (Fenton & Pfleeger, 1996, p. 74). Internal data quality dimensions are dimensions that can be measured by looking at the data (and possibly the real world values of the data), without considering the data's use. Measurements of internal dimensions can later be compared to the requirements of a task to make sure the level of these dimensions is fit for use. An example of an internal dimension is the correctness dimension: the correctness of data can be evaluated by only looking at the data and the real world counterparts, and can be presented as the error rate (Pipino, Wang, Kopcso, & Rybolt, 2005). The error rate can later be evaluated in the context of a specific use, for example, by comparing it against the highest acceptable error rate of a specific task. External data quality dimensions cannot be measured without considering the context or the task for which the data is used. For example, the relevance of data should always be evaluated with respect to a specific use. In this work, we target measurements or data deficiencies that correspond to internal data quality dimensions.

A key use of data quality dimensions, both internal and external, is to ease communication between stakeholders about certain data quality issues, i.e. they "function as a common set of terms" (Wand & Wang, 1996, p. 95). However, in practice, many organisations do not adopt a precise and common definition of data quality dimensions, which can lead to misunderstandings, and even wrong decisions (see e.g. Fenton & Pfleeger, 1996, p. 106). For example, when an organisation does not adopt a common definition for data consistency, it might not be clear for a stakeholder whether other stakeholders are communicating about the consistency between different attributes of a tuple, whether the value of the attribute adheres to a specific format (see e.g. Bronselaer et al., 2016) or whether the value of the attribute is different on another storage location (e.g. as a result of the CAP theorem (Brewer, 2001; Gilbert & Lynch, 2002)). As such, while these dimensions do enable stakeholders to communicate about data quality issues on a high level, they are rather unsuccessful in letting people communicate about a very precise aspect of data quality.

One way to accomplish a precise articulation of internal data quality dimensions is to encourage communication in terms of the exact measurement operations used to measure a dimension or a certain aspect of a dimension. Such measurement operations almost always directly correspond to a certain aspect of a certain data deficiency. For example, the size (aspect) of errors (deficiency) can be measured by the absolute difference between a data item and its true value (measurement operation) and the occurrence of an error in a data item can be measured by a Boolean expression that is 1 in case the data item is correct and 0 in case the data item is incorrect. Because this paper is concerned about the communication of precise aspects of internal data quality dimensions, we purposely do not use the term dimension but rather talk about aspects of data deficiencies such as the size of errors.

Two concepts of data deficiencies need to be clarified in order to clearly delineate the limitations of the approach that will be proposed in the next section: *(1)* the granularity of the data object of which the measurements can be represented and *(2)* on which type of scale the measurements can be made.

First, data deficiencies can be defined on different levels of the data hierarchy (Redman, 1996, p. 230; Even & Shankaranarayanan, 2007, p. 78). The lowest level of the data hierarchy is the data item, which is the most fine-grained data element. A data item contains a value for a certain attribute of a specific entity. The second-lowest level is the tuple. A tuple is a set of data items that contain the values for a set of attributes for a certain entity. Our approach allows to represent aspects of data deficiencies that are defined at the level

of a tuple or at the level of a data item.

Second, measurements of the aspects of a deficiency are presented on a certain scale (Krantz, Luce, Suppes, & Tversky, 1971). Common types of such scales are the nominal-, ordinal-, interval-, ratio- and absolute scale (Stevens, 1946). For example, the Boolean expression to indicate whether a data item is correct or not is, intuitively, expressed on an ordinal scale: 0 means that the data item is incorrect and 1 means that it is correct and, in this case, correct is greater/better than incorrect and so is its numerical value. Specifying which scale types can be represented by the approach is important because scale types dictate which kind of transformations are permissible such as to aggregate the measurements. For example, in case of nominal scales, there are almost no possible ways to aggregate the measurements. As we will demonstrate, our approach allows to represent aspects of data deficiencies regardless of their scale type.

3 DIMENSIONALITY REDUCTION AND EMBEDDINGS

In what follows we explain the concept of dimensionality reduction[1] and how two possible results of dimensionality reduction, a two or three dimensional embedding, can be interpreted by using a small (fictitious) dataset (Table 1).

Dimensionality reduction is the process of translating the number of dimensions[2] of an input structure to an output structure with fewer dimensions. The output structure is called an embedding. The core idea of dimensionality reduction is to retain the pairwise distances between elements in the multidimensional structure in the lower dimensional structure (embedding) as well as possible. If the embedding contains two or three dimensions, its visualisation can be easily interpreted by humans, as the remaining dimensions can be interpreted as x,y (or x,y,z in three dimensions) Cartesian coordinates in a grid. There are many algorithms that can be used to generate embeddings, such as principal component analysis (Jolliffe, 2002), Kohonen maps (Kohonen, 1995; Azcarraga, Hsieh, Pan, & Setiono, 2005) and t-sne (van der Maaten & Hinton, 2008; van der Maaten, 2014). Because these are all dimensionality reduction techniques, they all discard a part of the information contained within the original data. However, different dimensionality reduction algorithms can focus on maintaining different pieces of the original information. The reason behind why certain features are preferred (or not) to distinguish between observations in the resulting two dimensional embedding is linked to the optimisational nature of many embedding-techniques: find the best mapping from the original feature space to the two (three) dimensional coordinate space so that distances in 2D (3D) are as similar as possible to distances in the original feature space.

Currently, t-sne seems to provide the best visualisations for most datasets, as it focuses on maintaining distances between similar observations and less on observations that are dissimilar. This strategy leads to a grouping of similar observations causing groups of similar observations in the data to become visually apparent (van der Maaten, 2017).

Visualisations of two or three dimensional embeddings are based on the concept of neighbourhood: items or entities that are related to each other are depicted close together. This is illustrated by the example in Table 1 and Figure 1. Table 1 contains values for five attributes and measurements of three aspects of data deficiencies of a fictional home loan dataset. Suppose that four of these attributes are of interest for root cause analysis. Then, these four attributes can be used to generate an embedding of which the result is

[1] Note that dimensionality reduction is a machine learning technique and has nothing to do with dimensions in the sense of data quality.

[2] In this context, the 'dimensions' of an input structure could, for example, be seen as the number of attributes (or features) of an entitiy.

shown in Figure 1. From this figure, data quality stakeholders can see that records 1 and 2, and 3 and 4 are close together. Indeed, all four of these records were entered by Bart, and the home loans in records 1 and 2 were used to buy a house while the home loans in record 3 and 4 were used to buy land. Stakeholders can also see that home loans 5 and 6 are different from these four home loans: they were entered by Laura (and not by Bart) and are used to buy a garage. Yet, as already mentioned, because the embedding in Figure 1 has fewer dimensions than the input data, some information was lost. For example, while home loans 6 and 7 are visually distant, they were both used to buy a garage.

ID	Data Producer	Purpose of Loan	Duration	Borrowed Amount (BA)	Error Size BA	Is Error BA	Missing BA
1	Bart	Purchasing a house	240 months	150,000$	30$	1	Not missing
2	Bart	Purchasing a house	220 months	105,000$	40$	1	Not missing
3	Bart	Purchasing land	230 months	110,000$	500$	1	Not missing
4	Bart	Purchasing land	205 months	115,000$	600$	1	Not missing
5	Laura	Purchasing an apartment	100 months	NULL	/	/	Not registered
6	Laura	Purchasing a garage	90 months	NULL	/	/	Non-existent
7	John	Purchasing a garage	85 months	50,000 $	0$	0	Not missing

Table 1: An example of a dataset containing information about home loans and data quality measurements of the *Borrowed Amount* attribute.

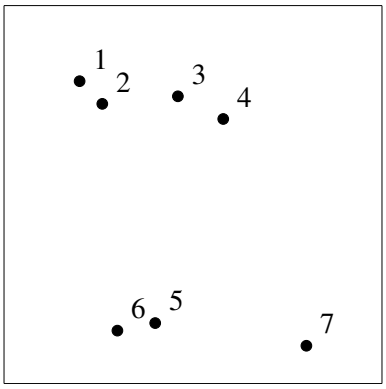
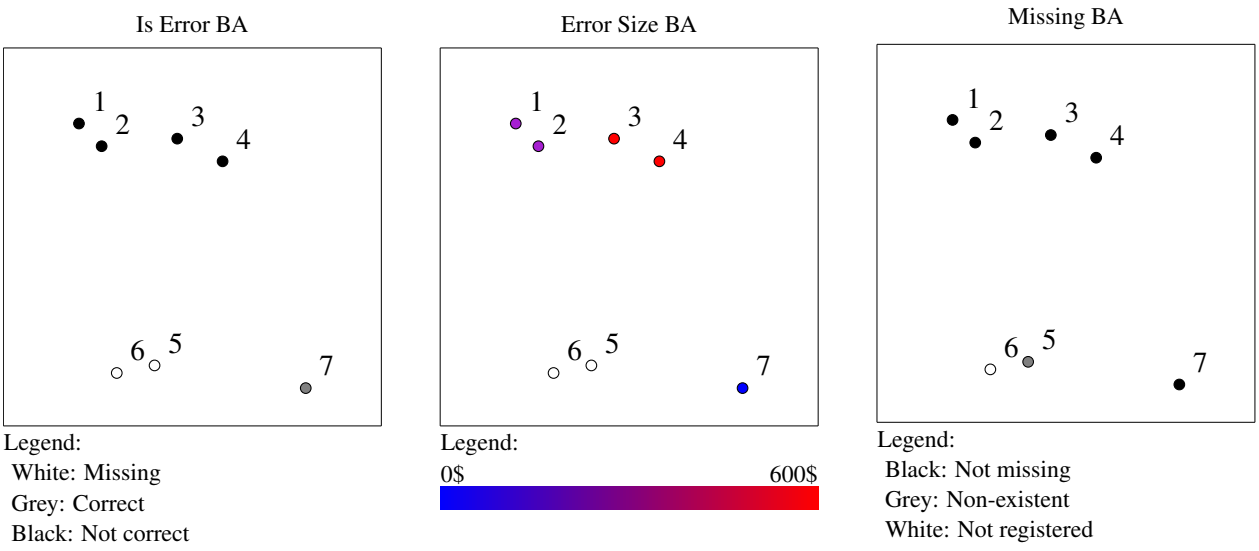


Figure 1: An example of an embedding based on the data in the columns *Data Producer*, *Purpose of Loan* and *Duration* of Table 1. A good embedding ensures that related records are depicted close together. For example, the embedding shows that records 1 and 2, and 3 and 4 are related because they were entered by Bart. Likewise, records 1 and 2 are grouped together because these loans were used to purchase a house.

Currently, visualisations based on embeddings are used in all kinds of domains, such as optical character recognition (van der Maaten & Hinton, 2008), and marketing analysis (Seret, Verbraken, Versailles, & Baesens, 2012). A common way to use these visualisations is to project a certain variable of interest on these embeddings using colours so that its occurrence and distribution becomes apparent.

To the best of our knowledge, the dimensionality reduction technique has not yet been used to gain insight in the distribution or presence of data deficiencies.

Therefore, we will demonstrate, using the same example as above, that this technique also has potential to provide insight in data quality concerns. Projecting measurements of aspects of data deficiencies onto the embedding is a matter of using the right colouring. In case of measurements presented on a nominal or ordinal scale, one can pick one colour per category and paint the records that contain this defect with the corresponding colour. In case of interval, ratio or absolute measurements, one can use a gradation of two colours and paint the records according to the degree to which they possess the deficiency. In our example,



(a) Shows how the occurrence of errors in the Borrowed Amount attribute is distributed.

(b) Shows how the size of errors in the Borrowed Amount attribute is distributed.

(c) Shows how the type of missing values in the Borrowed Amount attribute are distributed.

Figure 2: The embeddings in this figure show how three aspects of data deficiencies in the Borrowed Amount attribute are distributed with respect to four attributes of the original dataset: *Data Producer*, *Purpose of Loan*, *Duration* and *Borrowed Amount*.

we projected the three different types of data quality measurements of Table 1 onto the embeddings in Figure 2. For each type of data quality measurements, we created a copy of the embedding.

From the embeddings in Figure 2 two types of information can be derived. On the one hand, data quality stakeholders obtain an idea of the presence of data deficiencies. For example, one can conclude that there are more errors than missing values (see Figures 2a and 2c) and that the errors in records 1, 2 are small and those in 3, 4 and 7 are relatively large (see Figure 2b). On the other hand, one can also deduce information about the distribution of data deficiencies. For instance, one can see that records 1, 2, 3 and 4 contain errors and are close to each other and therefore require further investigation. When these records are consulted, it becomes apparent that all four records were entered by the same data producer. This information hints to the direction that the data producer himself is the root cause of these errors.

4 Proposed Approach

In this section, we describe a methodology which can be followed to apply the concepts presented in Section 3 in practice. The resulting graphs should allow data quality stakeholders to obtain an overview of the presence of deficiencies and whether these deficiencies are distributed randomly or systematically with respect to the values of the chosen attributes.

4.1 Step 1: Select Attributes

The first step is to select the collection of attributes that will be used to create the embedding. It is key that some attributes in this collection contain information that is able to hint in the direction of a certain root

cause. For example, in the case of manually acquired data, an attribute that can be included is an identifier of the person who entered the data (i.e. the data producer). If the final visualisation later on shows that data deficiencies are clustered together, the root cause of the data deficiency might have something to do with the motivation or ability of certain data producers.

However, selecting the attributes that could point to the root cause of data quality issues is not an easy task because this knowledge might not be available. Therefore, in an initial stage, we advise to include as many as possible attributes to generate the embedding.

4.2 Step 2: Generate the Embedding

In the next step, the embedding can be created by using a dimensionality reduction technique such as t-sne or principal component analysis. It is important that the reduction algorithm results in an embedding in which the distances between elements largely correspond to the distances between the records in the dataset.

Generating a perfect embedding will be close to impossible because when a multidimensional space is reduced to a space with fewer dimensions, some information will inevitably be lost. Although exact guidelines on creating embeddings are not within the scope of this preliminary research paper, we advise to tune the parameters of the embedding algorithm and to generate embeddings until the result is satisfactory. Trial and error until an interpretable embedding is generated is currently the accepted approach (van der Maaten, 2017).

4.3 Step 3: Define Data Quality Measure(s)

In the third step, precise data quality measures that measure a certain aspect of a data deficiency should be defined. It is important that the definition of a data deficiency is made at the level of a single data item or at the level of a record. For example, a measure that can be presented might be the absolute difference between the value of the tuple for a specific attribute and its true value. Another measure that can be presented on the embedding can be the number of attributes in a tuple that have an erroneous value. To ensure the validity of the final visualisation, the definition of the data quality measures should be communicated in great detail to the stakeholders.

4.4 Step 4: Evaluate the Embedding

The goal of this step is to ensure that the attributes having a high correlation with the data quality measurements appear as clusters in the embedding. In other words, this step aims to make sure that it is possible to inspect the distribution of the data deficiencies with respect to the values of the attributes that were selected to generate the embedding in Step 1. Because, if the attributes that correlate highly with the data quality measurements are not dominant in the embedding, a data quality stakeholder might wrongly conclude from the visualisation that the data deficiencies are randomly distributed instead of systematically. This evaluation should be executed for each measure defined in Step 3.

One way to evaluate the embedding with respect to a measure would be to first calculate the correlation of selected attributes with respect to the measured aspect of the data deficiency. Next, the correlation of the attributes could be compared to the degree to which the attributes dominate the embedding.

If the attributes that are dominant in the embedding do not correspond to the attributes that are correlated with the data quality measurements, there are basically two options. On the one hand, one could return to

the first step of the methodology and build a different embedding so that the dominant attributes correspond to the attributes that correlate well with the data quality measurements. This could be a valid approach because building an embedding could be a process of trial and error. Yet, when there are multiple data quality measures of which some fail the evaluation whilst others do not, one might find it difficult to create an embedding on which all the measures are well represented. For example, it might be that measures about the completeness of the data are correlated with the dominant attributes, whilst measures about data consistency are not correlated with the dominant attributes. In this case, it might be hard to create an embedding where measures of both data quality dimensions are correlated with the dominant attributes in the embedding. On the other hand, one could choose to simply not display the measurements for that specific data deficiency aspect on the embedding. This solution is advised when many other data quality measures are well represented.

4.5 Step 5: Project the Measurements on the Embedding

In the last step, the different data quality measures can be projected each on a separate copy of the embedding by using colours. This will enable data quality stakeholders to easily interpret the presence of the data deficiency in the data and help them search for the root causes of these deficiencies.

For a more close inspection, the visualisations can be accompanied by information about how the attributes are spread out on the embedding. One option to present this information is to display other copies of the embedding where, on each copy, the colours represent the values of another attribute.

5 Application

In this section we demonstrate how the proposed approach can be used in practice by applying it on a real world dataset. The dataset on which the approach was applied contains home loan data (812 tuples) of a large Belgian financial institution including data quality measurements.

In Step 1, four attributes of the dataset were selected that could potentially indicate were to look for the root causes of data deficiencies (see Table 2). For example, if the embedding would show that deficiencies occur more often when the true- or registered value of the asset is low, it could indicate that the people who enter this data are less motivated to correctly enter small values compared to large values. This might be because these people could be more aware of the importance to correctly enter large values compared to small values.

Attribute name	Description
Registered Value of Asset	The value of the asset that was registered in the database
True Value of Asset	The true value of the asset: this value could be found in the official documents that accompanied the purchase of the asset (e.g. the deed).
Has Movables	This attribute equals 1 if the transaction of purchasing the asset comprised movables and is 0 otherwise.
Nr. of Registrations	The total number of home loans in the entire home loan database that were entered by the person who entered the home loan. This attribute is a proxy for the experience an employee has.

Table 2: The attributes that serve as input for creating the embedding.

Figure 3: The embedding based on the attributes described in Table 2, generated by the t-sne algorithm.

In Step 2, we created an embedding of the four selected attributes using the t-sne algorithm (van der Maaten & Hinton, 2008; van der Maaten, 2014). The result of this step is shown in Figure 3. Figure 4 shows which attributes are dominant in the embedding.

In Step 3, data quality measures were defined. One of the data quality measures that was of interest to the business stakeholders was selected to be projected on the embedding: the occurrence of errors in the data. This data quality measure is 1 if the registered value of the collateral is not equal to the true value of the collateral and 0 otherwise.

In Step 4, we evaluated if the embedding is suited to show whether the errors occur randomly or systematically. We did this by visually inspecting the embedding using Figure 4 and checking whether the dominant attributes in the embedding correspond to the important attributes as indicated by the correlation of the attributes with the *Is Error* measure. This method confirmed that the embedding had a good fit with the data quality measure because the *Has Movables* appears to be the most important attribute and, in the embedding, the home loans which contained movables are clearly shown in a distinct cluster (see Figure 4a).

Attribute	Correlation with *Is Error*
Has Movables	0.43327624
Registered Value of Asset	0.11246633
True Value of Asset	0.07275483
Nr of Registrations	0.04091791

Table 3: The Pearson correlation for each of the attributes in the dataset with *Is Error*.

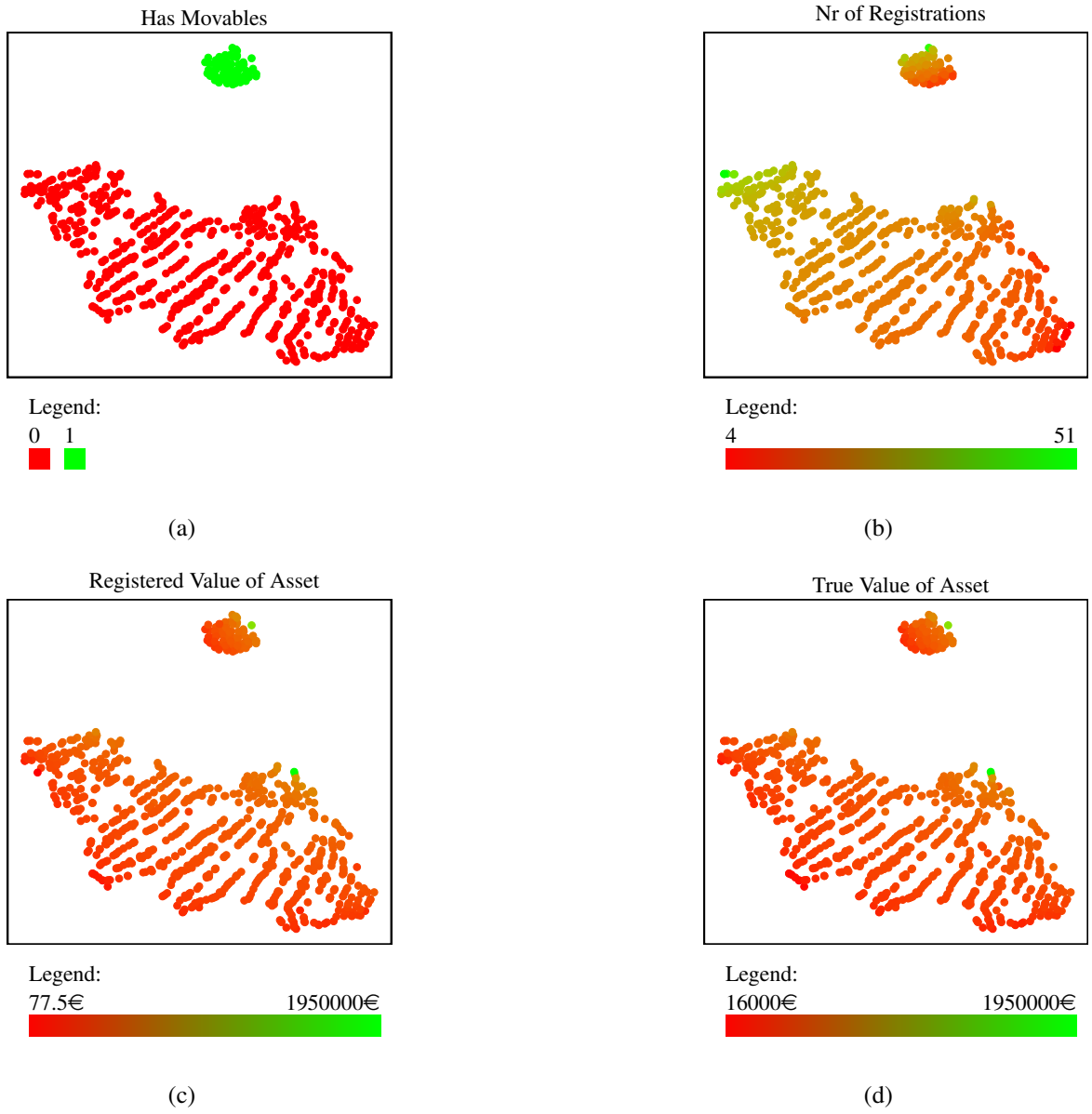


Figure 4: The embeddings in this figure show which attributes of Table 2 were favoured by the t-sne algorithm. The *Has Movables* attribute shows the most clear separation in the visualisation.

In Step 5, the data quality measurements were projected on the embedding which resulted in Figure 5. The embedding was shown to three data quality stakeholders. Each stakeholder was explained how to interpret the embedding and was interviewed separately in an informal way. The three data quality stakeholders were able to derive several useful insights from the visualisation. First, they could get a feel to which degree the home loans contained erroneous values for the asset that was used as collateral. Second, they could see that many errors were located in the *Has Movables* cluster. When looking closer to these home loans, it became apparent that many of them were in error because the data producer had to deduct the value of the movables from the price written on the deed. A possible solution for this root cause is to provide clear guidelines to the people who enter the data on how to enter home loans when the transaction of the good comprises movables. Third, the random distribution of the other errors could be interpreted by the stakeholders as that

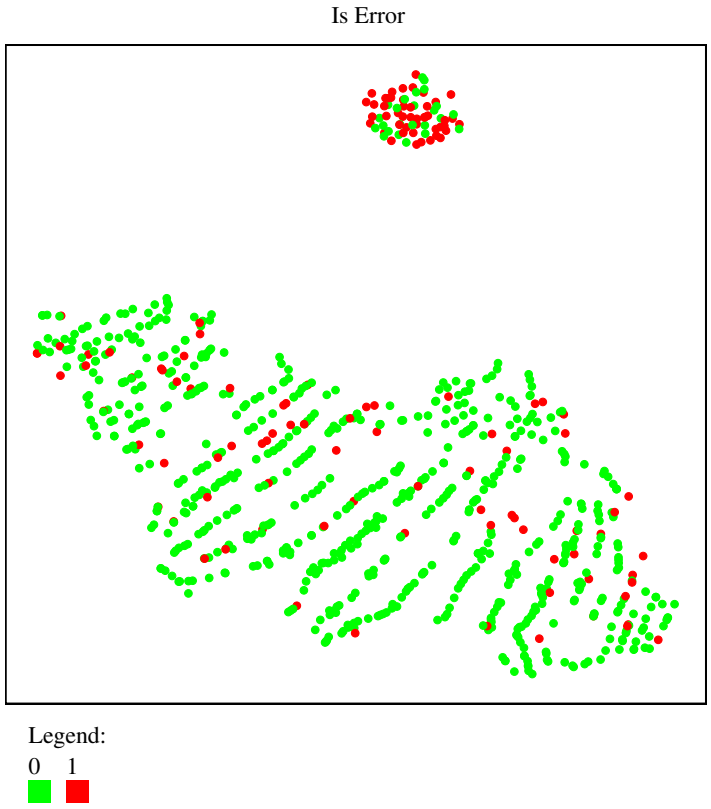


Figure 5: The embedding containing the projection of the data quality measurements.

there might be a cause which is not captured or explained by any of the included attributes. Because the data is manually acquired, this part of the errors might be explained by, for example, the intention of the data producers to enter the data correctly (Murphy, 2009; Haegemans, Snoeck, Lemahieu, Stumpe, & Goderis, 2016).

6 Discussion

As demonstrated by the example and the application to the real-life dataset, the proposed approach to visualise data quality deficiencies has two main benefits. First, the approach enables data quality stakeholders to assess the degree to which several data deficiencies are present in the dataset, which is thanks to the use of colourings, independent of the scale of the data quality measurements. Second, the approach allows data quality stakeholders to judge whether the data deficiencies are distributed randomly or systematically with regards to the chosen set of attributes and can therefore hint towards the root cause of the data deficiencies.

Yet, it is clear that the idea presented in this paper requires further investigation before the approach can be readily applied in practice or automated in the form of a data quality tool or data quality dashboard. The key issue that needs to be solved is the formulation of clear guidelines for the evaluation step (Step 4).

The problem with using embeddings to visualise the spread of some kind of label (in this case the data quality measurements) is that, essentially, there needs to be a good fit between three types of information: the selected attributes used to generate the embedding, the embedding itself, and the label. The fit between the

attributes and the embedding is ensured by the dimensionality reduction algorithm. However, dimensionality reduction techniques do not guarantee the fit between the embedding and the label. There are essentially three options to deal with this issue.

The first and most straightforward option is simply to not use embeddings to present data quality measurement information (or any other kind of label). This way, the only information that needs to be interpreted is that of the correlation between the attributes and the measurements. This information can be easily obtained by calculating the correlation matrix or by using classification techniques such as decision trees, linear- or logistic regression. While this approach would provide information about the potential causes that led to the data deficiencies, it would not be able to convey information about the degree to which the deficiencies are present in the data. In addition, business stakeholders might find it easier to interpret several copies of a two or three dimensional embedding, where each copy serves as a surface to project a different label on, than to interpret, for example, several completely distinct decisions trees.

The second option would be to make a dimensionality reduction technique aware of the values of the labels that will be projected on the embedding which will lead to a distinct embedding for each label. Yet, one of the reasons that embeddings can be easily interpreted is that each copy of the same embedding can be used as a surface to project a different label on. Thus, if each label would be projected on a distinct embedding and not a copy of the same embedding, the exact purpose of using embeddings might be partially defeated.

The third, and probably the most valid option is the one we included as Step 4 of the approach. We believe that it is best to first assess which attributes are dominant in the embedding, next, determine which attributes correlate well with the label and finally evaluate whether the dominant attributes are largely the same as the well-correlating attributes. While the correspondence between the attributes and the data quality measurements can be assessed by measures such as information gain, correlation or coefficients of a classification technique, assessing which attributes are dominant in an embedding is not straightforward. Selecting an appropriate metric to assess this dominance requires further investigation.

7 CONCLUSION

In this paper, we set out to investigate how the distribution of data deficiencies can be visualised with respect to a collection of data attributes while at the same time providing insight about the degree to which a data deficiency is present in a dataset. To this end, we adapted the approach of Reusens et al. (2017), which was proposed in the context of recommender systems to gain insight in which users receive relevant recommendations, and which users do not. The core idea of the proposed approach is to project data quality measurements on low dimensional embeddings of several attributes of the dataset using colours. We provided a preliminary validation of this approach by demonstrating how the resulting visualisation could provide utility to data quality stakeholders by enabling them discover root causes of deficiencies in home loan data of a large Belgian financial institution. In future work, we aim at fine tuning this approach so that it can be partially automated and implemented in, for example, data quality dashboards.

8 REFERENCES

Arts, D. G. T., De Keizer, N. F., & Scheffer, G.-J. (2002). Defining and Improving Data Quality in Medical Registries: A Literature Review, Case Study, and Generic Framework. *Journal of the American Medical Informatics Association*, *9*(6), 600–11.

Azcarraga, A. P., Hsieh, M., Pan, S. L., & Setiono, R. (2005). Extracting salient dimensions for automatic SOM labeling. *IEEE Trans. Systems, Man, and Cybernetics, Part C*, *35*(4), 595–600.

Ballou, D. P., & Tayi, G. K. (1999). Enhancing Data Quality in Data Warehouse Environments. *Communications of the ACM*, *42*(1), 73–78.

Brewer, E. A. (2001). Lessons from Giant-Scale Services. *Internet Computing, IEEE*, *5*(4), 46 – 55.

Bronselaer, A., Nielandt, J., De Mol, R., & De Tré, G. (2016). Ordinal Assessment of Data Consistency Based on Regular Expressions. *Communications in Computer and Information Science*, 317–328.

DeHoratius, N., & Raman, A. (2008). Inventory Record Inaccuracy: An Empirical Analysis. *Management Science*, *54*(4), 627 – 641.

Espetvedt, M. N., Reksen, O., Rintakoski, S., & Osterås, O. (2013). Data Quality in the Norwegian Dairy Herd Recording System: Agreement Between the National Database and Disease Recording on Farm. *Journal of Dairy Science*, *96*(4), 2271–82.

Even, A., & Shankaranarayanan, G. (2007). Utility-Driven Assessment of Data Quality. *ACM SIGMIS Database*, *38*(2), 75.

Fenton, N. E., & Pfleeger, S. L. (1996). *Software Metrics: A Rigorous and Practical Approach* (2nd ed.). Thomson Publishing.

Fisher, C. W., Lauria, E. J. M., & Matheus, C. C. (2009). An Accuracy Metric: Percentages, Randomness, and Probabilities. *Journal of Data and Information Quality*, *1*(3), 16:1 – 16:21.

Gilbert, S., & Lynch, N. (2002). Brewer's Conjecture and the Feasibility of Consistent, Available, Partition-tolerant Web Services. *SIGACT News*, *33*(2), 51–59.

Goldhill, D. R., & Sumner, A. (1998). APACHE II, Data Accuracy and Outcome Prediction. *Anaesthesia*, *53*(10), 937–943.

Haegemans, T., Snoeck, M., & Lemahieu, W. (2016). Towards a Precise Definition of Data Accuracy and a Justification for its Measure. In *International conference on information quality* (pp. 16:1 – 16:13). Ciudad Real, Spain.

Haegemans, T., Snoeck, M., Lemahieu, W., Stumpe, F., & Goderis, A. (2016). Towards a Theoretical Framework to Explain Root Causes of Errors in Manually Acquired Data. In *International conference on information quality* (pp. 15:1 – 15:10). Ciudad Real, Spain.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information Systems Research. *MIS Quaterly*, *28*(1), 75–105.

Jolliffe, I. (2002). *Principal component analysis*. Wiley Online Library.

Kohonen, T. (1995). *Self-organizing maps, volume 30 of springer series in information sciences.* Springer, Berlin, Heidelberg.

Krantz, D. H., Luce, D. R., Suppes, P., & Tversky, A. (1971). *Foundations of Measurement: Additive and Polynomial Representations* (Vol. 1). Academic Press.

Moges, H.-T., Dejaeger, K., Lemahieu, W., & Baesens, B. (2013, jan). A Multidimensional Analysis of Data Quality for Credit Risk Management: New Insights and Challenges. *Information & Management*, *50*(1), 43–58.

Murphy, G. D. (2009). Improving the Quality of Manually Acquired Data: Applying the Theory of Planned

Behaviour to Data Quality. *Reliability Engineering & System Safety*, *94*(12), 1881–1886.

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data Quality Assessment. *Communications of the ACM*, *45*(4), 211 – 218.

Pipino, L. L., Wang, R. Y., Kopcso, D., & Rybolt, W. (2005). Developing Measurement Scales for Data-Quality Dimensions. In R. Y. Wang, E. M. Pierce, S. E. Madnick, & C. W. Fisher (Eds.), *Information quality* (pp. 37 – 51). Armonk, NY: M.E. Sharpe.

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

Reusens, M., Haegemans, T., Lemahieu, W., Baesens, B., Snoeck, M., & Sels, L. (2017). *Understanding Recommendation Quality Using Embeddings* (Tech. Rep.). KU Leuven.

Seret, A., Verbraken, T., Versailles, S., & Baesens, B. (2012). A new som-based method for profile generation: Theory and an application in direct marketing. *European Journal of Operational Research*, *220*(1), 199–209.

Stevens, S. S. (1946). On the Theory of Scales of Measurement. *Science*, *103*(2684), 677–680.

Thiru, K., Hassey, A., & Sullivan, F. (2003). Systematic Review of Scope and Quality of Electronic Patient Record Data in Primary Care. *BMJ*, *326*(7398), 1070.

van der Maaten, L. (2014). Accelerating t-sne using tree-based algorithms. *Journal of Machine Learning Research*, *15*(1), 3221–3245.

van der Maaten, L. (2017). *t-SNE*. `https://lvdmaaten.github.io/tsne/`. ([Online; accessed 9-May-2017])

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, *9*(Nov), 2579–2605.

van der Meyden, R. (1999). Logical Approaches to Incomplete Information: A Survey. In J. Chomicki & G. Saake (Eds.), *Logics for databases and information systems* (Vol. 37, p. 143). Kluwer Academic Publishers.

Wand, Y., & Wang, R. Y. (1996, nov). Anchoring Data Quality Dimensions in Ontological Foundations. *Communications of the ACM*, *39*(11), 86–95.

Wang, R. Y. (1998). A Product Perspective on Total Data Quality Management. *Communications of the ACM*, *41*(2), 58 – 65.

Wang, R. Y., Reddy, M. P., & Kon, H. B. (1995). Toward Quality Data: An Attribute-Based Approach. *Decision Support Systems*, *13*(3), 349–372.

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

Weiskopf, N. G., & Weng, C. (2013). Methods and Dimensions of Electronic Health Record Data Quality Assessment: Enabling Reuse for Clinical Research. *Journal of the American Medical Informatics Association*, *20*, 144–151.

Zmud, R. W. (1978). An Empirical Investigation of the Dimensionality of the Concept of Information. *Decision Sciences*, *9*(2), 187 – 195.