Intrinsic Indicators for Numerical Data Quality

Milen S. Marev, Ernesto Compatangelo and Wamberto W. Vasconcelos
Department of Computing Science, University of Aberdeen, Aberdeen, AB24 3UE, U.K.

Keywords: Data Quality, Intrinsic Data Quality, Data Quality Indicators, Pre-processing, Numerical Data Quality, Numerical Data Quality.

Abstract: This paper focuses on data quality indicators conceived to measure the quality of numerical datasets. We have devised a set of three different indicators, namely Intrinsic Quality, Distance-based Quality Factor and Information Entropy. The results of quality measures based on these indicators can be used in further data processing, helping to support actual data quality improvements. We argue that the proposed indicators can adequately capture in a quantitative way the impact of different numerical data quality issues including (but not limited to) gaps, noise or outliers.

1 INTRODUCTION

The generation and processing of very large amounts of digitally recorded information from a variety of heterogeneous sources (sensorial or otherwise) is at the core of the 'big data society' emerging at the onset of the 21^{st} century. A relevant portion of this information, which is seamlessly produced and consumed to keep society going, consists of numerical datasets. These are generated and processed as part of the wider digitalised management of goods and services, performed using computational workflows that run from inception to delivery. Such workflows increasingly use a combination of artificial intelligence and other advanced software technologies to derive new results in a variety of monitoring and processing scenarios. For the results of workflows that produce and use numerical datasets to be meaningful, accurate and reliable (*i.e.*, to be of quality), data used in each workflow step input must comply with some context-dependent quality metrics. A number of frameworks and dimensions have been proposed to measure data quality (Batini et al., 2009; Li, 2012; Redman, 2008; Deming, 1991; Dobyns and Crawford-Mason, 1991; Juran, 1989; Group, 2013; Batini and Scannapieco, 2006; Cai and Zhu, 2015; Pipino et al., 2002; Swanson, 1987; Olson and Lucas Jr, 1982; De Amicis et al., 2006; Todoran et al., 2015; Hufford, 1996; Loshin, 2001; Marev et al., 2018); however, most of these frameworks focus on non-numerical types such as alphanumerical strings, free text or timestamps. Once focus is restricted to numerical types, uncertainty is explicitly taken into account, and the existing quality dimensions are fully analysed in depth, the resulting landscape only remains populated by very few useful notions. Hence, new concepts for the effective measurement (*i.e.*, for the *quantitative* evaluation) of numerical data quality must be introduced. Hence, in this paper we define a set of novel numerical data quality indicators specifically designed to address effective quality measurements.

Paper Structure and Content. Section 2 identifies the relevant aspects for the definition of a quantitative framework to measure numerical data quality and its changes. Section 3 discusses the normalised decimal format used to represent numerical data with uncertainty. Section 4 introduces the core concept of information entropy and describes entropy variation as the basis for measuring numerical data quality improvements. Section 5 uses the formulas as defined in the previous section and evaluates their efficiency with the use of simple sine wave dataset. Finally, Section 6 draws conclusions on the entropy-centred framework presented in this paper and outlines further research work in this area.

2 DATA QUALITY ASPECTS

The question arises as to how to define and measure quality in a numerical dataset characterised by a given degree of associated uncertainty; this issue is made more complicated by the fact that any numerical data quality figure is inherently context-dependent.

Marev, M., Compatangelo, E. and Vasconcelos, W.
Intrinsic Indicators for Numerical Data Quality.
DOI: 10.5220/0009411403410348
In *Proceedings of the 5th International Conference on Internet of Things, Big Data and Security (IoTBDS 2020)*, pages 341-348
ISBN: 978-989-758-426-8

Various frameworks have been proposed which address data quality definition and measurement for both numerical and non-numerical data, with emphasis on data types typically found in (No)SQL databases (Batini et al., 2009; Li, 2012). For quality measurement purposes, these frameworks have analysed the concept of data quality along a number of different dimensions, proposing a specific metric for each such dimension. Moreover, a framework has been recently proposed which explicitly addresses numerical data (Marev et al., 2018), focusing on eight data quality dimensions that are relevant to the numerical subdomain.

However, some of the numerical data quality dimensions proposed in (Marev et al., 2018) – namely, accessibility, currency, timeliness, and uniqueness – only address *extrinsic* data quality aspects. More specifically, access easiness and speed, newness, real-time loading and processing, and lack of duplicates (*i.e.*, the exemplifying instantiations of these dimensions) are not intrinsic properties of numerical data as such, but depend on some external conditions. These can be actually addressed by modifying 'the machinery' around data rather than data themselves. For instance, extrinsic data quality issues may delay workflows (because of the extra time needed to acquire and filter all data needed for computations) but have no impact on the quality of workflow results.

Conversely, the other four dimensions proposed in (Marev et al., 2018) (namely, accuracy, consistency, completeness, and precision) represent properties of numeric datasets that explicitly affect the quality of workflow results. In other words, they address *intrinsic* numerical data quality aspects. This is because the quality improvement of workflow results explicitly depends on the improvement of the workflow-consumed datasets along one or more of the accuracy, consistency, completeness, and precision dimensions, which are discussed in detail below.

We now introduce and describe the following features that set numerical data apart from other types:

Intrinsic Approximation. Numerical data are often the result of either physical measurements or model-based calculations. Hence, in theory at least, such results can take any value in a given subset of real numbers. In a very few cases, complex numbers (*i.e.*, real and imaginary value pairs represented as $z = x + iy$, where $i = \sqrt{-1}$) are used. However, they are not discussed in this paper as the real and the imaginary part would be separately treated using techniques developed for real numbers. Similarly, we do not discuss integer numbers, as they either represent extremely approximated values (in which case they can be treated as very rough real numbers subject to our framework) or they represent counters/identifiers of no interest in our context.

Having restricted our focus to real numbers, we note that there are two compelling reasons why numerical data values are never actually represented as real values but rather as rational values. The latter are defined as ratios between two integer values (with a non-zero denominator) and are either characterised by a finite number of digits or by an endless repetition of the very same finite sequence of digits.

The first reason why rational numbers are used to represent real numeric entities in any practical situation is that both measurements and model-based calculations are approximations of the measured/computed reality. This leads to a truncation in the number of digits used to represent a real number, which depends on the accuracy of a measurement or of a calculation in each specific context.

The second reason why rational numbers are used in place of real ones is that current (and likely, future) digital technologies have limited capacity to store and process real numbers. Pragmatically, although the accuracy of a number is constantly improving, we are unlikely to reach a short-term situation of endless capacity whatever the medium, which is what would be required to fully represent real values accurately with a mathematical precision.

Intrinsic Uncertainty. A fundamental characteristic of numerical data, which sets them apart from other data types, is that numbers generally have an intrinsic uncertainty associated to them. This is because numerical data typically represent the result of either approximate physical measurements or calculations based on truncations and finite-method approximations. Both such measurements and calculations associate an inherently unavoidable degree of uncertainty to their results. Uncertainty is an intrinsic property of all numerical datasets that are not just collections of integer counter values or identifiers. One of the contributions of this paper is the modelling of intrinsic uncertainty and how this can be used to measure data quality.

Numerical data uncertainty and its implications are often overlooked in numerical workflows. This may be due to uncertainty not being perceived to have a major impact on numerical information processing and on their results, which tend to focus on datasets as if they were uncertainty-free. However, this is a dangerous misconception, as uncertainty (which typically represents an estimate of the average indeterminacy associated with dataset values) is actually the basis to measure numerical data quality and thus to evaluate the effect of different kinds of data quality improvements. Uncertainty is not only unavoidable because

of the way most numerical datasets are generated; it is also needed as a basis for any quality considerations.

Characteristic Series Structure. A feature shared by many numerical datasets is their structure as *series* of value pairs, triples or quadruples. In this paper, we only focus on pairs for simplicity, as n-ples are treated similarly in the context we focus on.

A numerical series is a finite discretisation of some theoretical function $y = f(x)$, *i.e.*, $y_k = f(x_k)$, where $k = 1, 2, \ldots n$. Such discretisation is often arranged as a list of pairs (y_k, x_k) ordered by the value of x_k, where both numerical values in a pair have associated uncertainties that are generally independent from one another. If x_k represents time, then the list of pairs (y_k, x_k) is called a time series.

The abscissa x_k represents the physical 'independent' variable, while the ordinate y_k represents the 'dependent' variable. Numerical data series are customarily ordered by abscissa values, which are generally spaced evenly. This may make it easier to detect whether a dataset is lacking any elements within the interval of independent variable values the dataset is a record of. Ordinate values generally follow some underlying physical *pattern* that is representative of a physical phenomena or a discretised model.

In time series, the uncertainty associated with each pair of values representing the element is often the same for all x_k on the one side and for all y_k on the other, although these two uncertainty values are generally different from each other. In case uncertainty values are the same for all x_k (and, separately, for all y_k) they do not need to be recorded beside each pair of values, but can be specified separately elsewhere for the whole dataset. This approach, which avoids the overloading of a dataset with the unnecessary repetition of identical information, may however result in the role and in the impact of the associated uncertainty values being overlooked.

3 DATA UNCERTAINTY

The representation of numerical data with uncertainty has a long history that encompasses over four hundred years of experimental physics. In this leading branch of science, the result of an observable variable measurement is represented as a pair that specifies both the measured value and its associated measurement error, where the latter characterises the uncertainty associated with the measurement process. Although different numeric representations are currently used in science, technology, engineering, and mathematics, the decimal representation (see below) is by far the most widely used to record and display numerical data.

3.1 The Normalised Decimal Format

In order to make our discussion precise, we introduce a general and flexible format for our numeric data. The decimal representation is a class of different variants conveying the same information in different formats. For instance, let us consider a rational number R: this could be either expressed in natural decimal format as $R = 0.00000234567$ or, more compactly, in exponential format as $R = 0.0234567 \times 10^{-4}$. The latter leaves some degree of freedom as to how to 'distribute' non-zero numerical digits between base and exponent. In our example,

$$R = 0.0234567 \times 10^{-4} = 0.234567 \times 10^{-5} = 2.34567 \times 10^{-6}$$

One useful, standardised way to represent 'real' numerical data in exponential format is the *normalised decimal format* (NDF), where the base is always a number smaller than one but its first fractional digit is always non-zero. The exponent is set accordingly. Using the EBNF (Extended Backus-Naur Form) notation, a (rational) number representing the result of some measurement or digital calculation can be thus represented in a 'normalised' form as

$\langle NormalisedNumber \rangle ::= \langle NDF \rangle \times \langle Power \rangle$

where

$\langle NDF \rangle ::= [\ \langle Sign \rangle\ ]\ 0.\langle NonZeroDigit \rangle \langle Digit \rangle$

$\langle Power \rangle ::= 10\ ^{[\ \langle Sign \rangle\ ]\ \langle Digit \rangle\ \{\ \langle Digit \rangle\ \}}$

$\langle Sign \rangle ::= + \mid -$

$\langle NonZeroDigit \rangle ::= 1 \mid 2 \mid 3 \mid 4 \mid 5 \mid 6 \mid 7 \mid 8 \mid 9$

$\langle Digit \rangle ::= 0 \mid \langle NonZeroDigit \rangle$

Each numeric value resulting from a measurement is characterised by an uncertainty, even if this is sometimes omitted or specified elsewhere. This depends on a number of factors (*e.g.*, limited accuracy of the measurement process and/or limited precision of the measuring instrument, environmental noise). Hence, the $\langle NDF \rangle$ of numeric values that represent measurements is often expressed in the form

$\langle NDF \rangle ::= (\ \langle Baseline \rangle \pm \langle SemiUncertainty \rangle\ )$

where

$\langle Baseline \rangle ::= [\ \langle Sign \rangle\ ]\ 0.\langle NonZeroDigit \rangle\ \{\ \langle digit \rangle\ \}$

$\langle SemiUncertainty \rangle ::= 0.\{0\}\langle NonZeroDigit \rangle$

3.2 Baseline, Decimal Significance, and Uncertainty Amplitude

A numeric data item with its associated uncertainty, once represented using a normalised decimal format (*e.g.*, $N = 0.12379 \times 10^{-4} \pm 0.0004 \times 10^{-4}$) is characterised by different elements, such as

- the *baseline* (0.12379 in this case),
- the number of decimal digits in its decimal part (5 for baseline 0.12379, as in this case),
- the uncertainty half amplitude expressed as a fractional decimal part (0.0004 in this case),
- the order of magnitude (10^{-4} in this case).

Uncertainty Set. Following the above grammar, a numeric data element in NDF-compliant form can be more concisely represented as

$$N = \{V \pm u\} \times 10^{\langle Exp \rangle} \tag{1}$$

where V is the numeric value of the element, u is the (numeric) uncertainty associated with such value and $\langle Exp \rangle$ is the exponent of the power of ten used to represent the 'normalised' magnitude of both V and u. Moreover, $\{V \pm u\}$ is explicitly used to indicate that the numeric data element N is actually the set of all possible rational values in the discrete interval $(V-u, V+u)$, where the *distance* between any two elements of this set is given by the unit value of the least significant digit in u. The finite set of all possible V values in the interval $(V-u, V+u)$ defines the *uncertainty set* of N. In other words, N can take any u-dependent value in the above interval. For instance, if $N = \{0.12379 \pm 0.00004\} \times 10^{-4}$, the unit value of the least significant digit in u is 0.00001 and $N = \{V \pm u\} = \{0.12375, 0.12376, 0.12377, 0.12378, 0.12379, 0.12380, 0.12381, 0.12382, 0.12383\} \times 10^{-4}$.

The order of magnitude of both the baseline and the uncertainty tends not to play a big part in the numerical data quality metrics introduced and discussed later in this paper if the numerical elements considered are all expressed using the normalised decimal format. In fact, these quality metrics are either based on ratios between V and u or on the fractional decimal part of u, which means that the order of magnitude of NDF-compliant numerical data can be ignored altogether. Here and in the following we will thus use examples where a numeric element is represented in the simpler form $N = \{V \pm u\}$, thus avoiding having to show the order of magnitude unless strictly necessary in specific contexts.

Extension to Datasets. Our discussion so far has focused on datasets where each element is implicitly composed of one value with associated uncertainty. We need to consider what changes are necessary if a dataset is a series of pairs of numeric values rather than a sequence of single numbers. Extending formula 1, it is possible to represent each paired element in the series as

$$\begin{aligned} C_k &= (C_x^k, C_y^k), \\ &\quad \text{where} \\ C_x^k &= \{V_k^x \pm u_k^x\} \times 10^{\langle Exp_k^x \rangle}, \\ C_y^k &= \{V_k^y \pm u_k^y\} \times 10^{\langle Exp_k^y \rangle} \end{aligned} \tag{2}$$

and where the meaning of symbols in this formula is clear. Note that while $\langle Exp_k^y \rangle$ is likely to be the same for all elements (such that $\forall k : \langle Exp_k^y \rangle = \langle Exp^y \rangle$) and, independently, $\langle Exp^x \rangle$ – where for any two values k, k', $\langle Exp_k^y \rangle = Exp_{k'}^y$ – is likely to have the same value for all x_k (such that $\forall k : \langle Exp_k^x \rangle = \langle Exp^x \rangle$), in general $\langle Exp^y \rangle \neq \langle Exp^x \rangle$.

Dealing with a numerical series in terms of evaluating and measuring data quality means that the considerations introduced in the following section will have to be separately applied to the independent and to the dependent variable in the series.

4 DATA QUALITY INDICATORS

The big challenge of all methods and frameworks introduced to evaluate numerical data quality is the identification of suitable *quality indicators* that can be used along each (intrinsic) numerical data quality dimension as the basis for a corresponding set of metrics. Such indicators would make it possible to compute 'quality values' that can be used to position a given dataset element (as well as the whole dataset) along a numerical data quality scale. The quality level associated to any such value would necessarily be entirely context-dependent. This is because the very same numerical element or dataset could be fit for purpose (and thus deemed of 'suitable' quality) in a specific computational scenario, while it could not be so (and thus be deemed of 'unsuitable' quality) in another, different scenario.

For the purpose of identifying which parameters should be used as indicators to compute numerical data quality, let us consider a numerical dataset $\mathbf{S} = \{N_1, N_2, \ldots, N_k\}$ composed of N elements, where the *k-th* element N_k is such that $\forall k = 1, 2, ..., N$: $N_k = \{V_k \pm u_k\}$. We focus on this straightforward dataset structure for sake of simplicity. In any case, it is implicit that the set $\mathbf{S} = \{N_k\}$ corresponds to the sampling of some variable V at regular intervals along a discretised, implicit x axis, so that in reality $V_k = \Phi(x_k)$.

However, if (as in most cases) the distance between any two values of the sampling abscissa is a constant for a given dataset, *i.e.*, $\forall k : \; x_k - x_{k-1} = c$, we can avoid explicitly considering abscissa values x_k

and we can thus refer to element values in an abridged way as V_k directly rather than as $V(x_k)$. Considering a dataset as a set of elements $N_k = \{V_k(x_k) \pm u_k\}$ would add extra complexity due to the handling of two distinct variables; there is no need to do so explicitly except for those datasets where sampling occurs at irregular intervals.

4.1 Intrinsic Quality $\mathcal{Q}$

Once a numerical element and its uncertainty are expressed in NDF as a set $\{\langle Baseline \rangle \pm \langle Uncertainty \rangle\}$, it is possible to use the baseline-to-uncertainty ratio to define the *intrinsic quality* $\mathcal{Q}$ of that element. In other words, if $N_k = \{V_k \pm u_k\}$, then its intrinsic quality can be mathematically expressed as

$$\mathcal{Q}(N_k) = log_{10}\left(\frac{V_k^2}{u_k \cdot |V_k - \widetilde{V}_k|}\right) \tag{3}$$

This formula captures the dependencies on uncertainty and on the distance between an element value and the corresponding value on the best fit regression curve for the dataset (whatever this curve may be). Notably, the intrinsic quality of an element:

- Is inversely proportional to uncertainty. A numerical data element with no uncertainty (*i.e.*, an infinitely precise element) is characterised by an infinite intrinsic quality. However, intrinsic data quality in our context is only defined for numerical dataset elements with a non-null uncertainty.
- Is inversely proportional to the distance between the actual element value and the value of the corresponding element laying on the best fit regression curve.
- Can be approximated using the simpler formula $\mathcal{Q}(N_k) = log_{10}|V_k/u_k| \quad \text{if} \quad |V_k - \widetilde{V}_k| \gg 1$

The intrinsic quality $\mathcal{Q}(\mathbf{S})$ associated with a set $\mathbf{S} = \{N_1, N_2, \ldots, N_M\}$ of M numerical values, each characterised by its own uncertainty and expressed in NDF, is defined as the sum of the individual quality values of each element N_k in the set, *i.e.*

$$\mathcal{Q}(\mathbf{S}) = \sum_{k=1}^{M} \mathcal{Q}(N_k) \tag{4}$$

The intrinsic quality of a numerical dataset element with associated uncertainty is an indicator that can be used as the basis of accuracy metrics. One could be led to conclude that, as intrinsic quality is not only just based on both dataset element value and uncertainty but on their ratio, this is by far the best possible option in the circumstances. However, such conclusion is not completely true, as other indicators are needed to perform a more comprehensive dataset quality assessment from different perspectives.

4.2 Indicators based on Value Only: Distance-based Quality Factor

So far we have defined numerical data quality indicators based on value-uncertainty pairs. However, it is possible to introduce a value-only indicator, namely the *distance-based quality factor* $\mathcal{Q}_\mathcal{D}$, which is proportional to the distance between a dataset element and the corresponding point on the best fit regression curve for this dataset. For sake of simplicity, let us consider a dataset $\mathbf{S} = \{N_k\}$ of M elements, where $\forall k : N_k = \{V_k \pm u_k\}$ and 1 $\widetilde{V}_k = \varphi(\mathbf{S})$ is the corresponding best fit regression curve. In formulae, the distance-based quality factor is thus defined as

$$\mathcal{Q}_\mathcal{D}(N_k) = log_{10}\left(\frac{V_k^2}{|\widetilde{V}_k \cdot (V_k - \widetilde{V}_k)|}\right) \tag{5}$$

Note that $\mathcal{Q}_\mathcal{D}$:

- Is very similar to the intrinsic quality $\mathcal{Q}$. However, differently from $\mathcal{Q}$, $\mathcal{Q}_\mathcal{D}$ does not include any reference to uncertainty u.
- Is inversely proportional to the distance between the actual element value, namely V_k and the value of the corresponding element laying on the best fit regression curve, namely $\widetilde{V}_k$.
- Can be approximated using the simpler formula $\mathcal{Q}_\mathcal{D}(N_k) = log_{10}|V_k/\widetilde{V}_k| \quad \text{if} \quad |V_k - \widetilde{V}_k| \gg 1$

The distance-based quality factor associated with a set $\mathbf{S} = \{N_1, N_2, \ldots, N_M\}$ of M numerical values is defined as the sum of the individual $\mathcal{Q}_\mathcal{D}(N_k)$ values associated with each element in $\mathbf{S}$, *i.e.*

$$\mathcal{Q}_\mathcal{D}(\mathbf{S}) = \sum_{k=1}^{M} \mathcal{Q}_\mathcal{D}(N_k) \tag{6}$$

4.3 Indicators based on Uncertainty Only: Information Entropy $\mathcal{E}$

To identify a different kind of numerical data quality indicator, let us consider uncertainty only. We noted in Section 3.2 that the uncertainty set associated to a numeric element is expressed using a single significant digit (*e.g.*., $E = \{0.9783 \pm 0.005\}$ rather than $E = \{0.9783 \pm 0.0027\}$). Once represented in NDF (*e.g.*, $u = \{\pm 0.003\}$ in this case), uncertainty u specifies both the accuracy magnitude of a numeric element value and the amplitude of its *inaccuracy* (see below). For instance, in the above example, the NDF-normalised order of magnitude for accuracy is 10^{-3}, which represents the uncertainty unit. Once u is taken into account, the actual value of E can be any of the numbers in the set defined by its uncertainty

amplitude, *i.e.*, $E \in \{0.9780, \ldots, 0.9783, \ldots, 0.9786\}$. This is denoted as *the uncertainty set* associated to the numerical element E.

For each numerical element with uncertainty expressed as $E = \{V \pm u\}$, we can define the measure μ of the corresponding *uncertainty set* as the number of elements in this set. This measure, which explicitly depends on the uncertainty amplitude, can be easily calculated from the single significant (*i.e.*, non-zero) digit of the corresponding inaccuracy amplitude, *i.e.* $\mu = \langle NonZeroDigit \rangle \times 2 + 1$ For example, if $\{0.9783 \pm 0.003\}$, then $\mu_E = 3 \times 2 + 1 = 7$.

We can now define the notion of *information entropy* $\mathcal{E}$ associated with a number N as

$$\mathcal{E}(N) = \log_{10}(\mu_N) \tag{7}$$

For instance, let us consider a number $N_1 = 0.34 \pm 0.05$. In this case, $\mu_{N_1} = 5 \times 2 + 1 = 11$ and the associated entropy is thus $\mathcal{E}(N_1) = log_{10}(11) = 1.0414$.

Although information entropy is a property defined on individual data with associated uncertainty, it can be extended to entire datasets. The information entropy associated with a set of M numerical values $\mathbf{S} = \{N_1, N_2, \ldots, N_M\}$ – each characterised by its own uncertainty and expressed in NDF – is defined as the sum of the individual information entropy values associated with each Nk in the set, *i.e.*

$$\mathcal{E}(\mathbf{S}) = \sum_{k=1}^{M} \mathcal{E}(N_k) \tag{8}$$

If a number N is *infinitely accurate* (*i.e.*, if $u = 0$), then its uncertainty set is a singleton set, which has one element only. Hence, the associated information entropy is $\mathcal{E}(N) = log_{10}(1) = 0$.

Information entropy was first introduced in (Shannon, 1948) in the context of communications based on the transmission of analog electric signals. In our context, however, information entropy is considered as a quantitative measure of numerical data quality. More specifically, it is a measure of data accuracy, consistency, completeness, and precision, namely the four intrinsic numerical data quality dimensions described in Section 2 of this paper.

5 EVALUATION

This section focuses on the evaluation of the metrics (the data quality indicators) introduced and discussed in the previous section. To guarantee consistency and to ensure that the assessment of the proposed metrics is unbiased, we used a perfect sinusoidal wave as the input in our computational workflow. This type of input was chosen as it models many different physical phenomena such as the propagation of elastic (seismic) vibrations, sound and light.

The sinusoidal dataset used in the simulation experiments described in this paper has a rate of 44100 and a frequency of 44100 Hz with volume at 100%. Figure 3 shows a time representation of the generated wave. Figures 1 and 2 show the experiment runs. Each experiment is designed to run 100 times, with each run consisting of 100 iterations, with a grand total of 10000 runs for each experiment. To test the results of the data quality indicators introduced in this paper, we devised 3 different test cases:

- Introduce 100 | 1000 | 10000 gaps
- Introduce 100 | 1000 | 10000 outliers
- Introduce 100 | 1000 | 10000 gaps and outliers

The experiments with synthetic sinusoidal wave data that are described in this paper are intentionally designed for simplicity. However, more advanced experiments with real datasets (*e.g.*, using Distributed Temperature Sensing data in Oil & gas production wells) are currently being performed. These will be introduced and discussed in future publications. Each of our current experiments provides a variety of indicators: Average (AVG), Standard Deviation (STD) and the three indicators introduced in section 4. Both AVG and STD are introduced as a tool to verify and validate the effectiveness and the correctness of our algorithms that artificially introduce defects (namely, gaps and outliers) in the 'clean' dataset. Such algorithms are not discussed in this paper. In our first experiment we only introduced gaps in the dataset. As a result, both AVG and STD are gradually increasing in parallel with the increase in the number of gaps. The difference in results is due to the fact that gaps are randomly generated, replacing any value in the dataset with one characterised by an almost infinite uncertainty. The results show that two of the indicators introduced in Section 4, namely the intrinsic quality Q and the distance-based quality factor Q_D both decrease in value following the introduction of more data gaps. This is to be expected, as the effect of gaps can be either represented by infinite uncertainty or by an out-of-scale value if uncertainty is not used as a parameter in the considered numerical data quality indicator. The Information Entropy metric value in all 3 experiments increases if each gap is considered as a number with an infinite uncertainty, so that it cannot be precisely localised any more. It should be noted that the theoretical treatment of a gap as a dataset point whose uncertainty becomes infinite or, alternatively, as a dataset point whose value becomes out of bound, is a theoretical artifice to keep the number of dataset elements unchanged during the experiments.

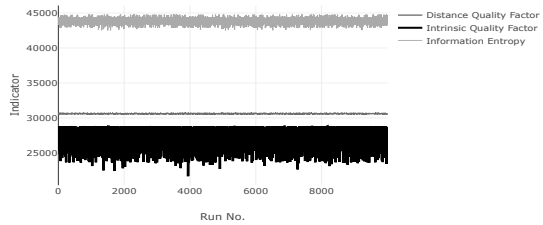

Figure 1: 1000 gaps Metrics.

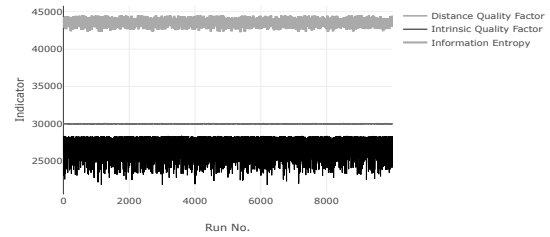

Figure 2: 1000 Gaps & Outliers Metrics.

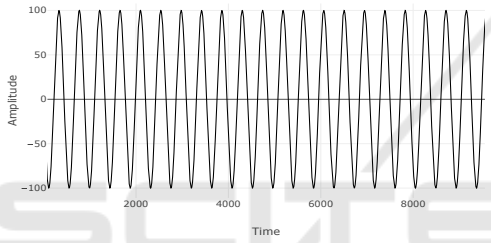

Figure 3: Sinewave.

This is because any methodologically sound comparison should be done in like-for-like conditions. In our case, this means that once a single element gap is introduced the corresponding dataset should not lose an element and thus become smaller. Simply, the element is retained for both count and indicator purposes but its parameters (uncertainty or value) are changed to reflect the practical effect of generating a gap.

In the second experiment we introduced only outliers in the dataset. An outlier is defined as an element whose value is at least three times bigger than that of its immediate left and right neighbours in the ordered series of dataset elements. The outlier is thus represented as a number that is $3 \times \sigma$ where σ is the base case standard deviation. This definition of outlier is purely arbitrary but it is rooted in the empirical analysis of production data in a variety of industries, where time series are assessed for temporal stability and random sensorial malfunctionings lasting fractions of a second happen frequently in extreme environmental conditions. For instance, thermal sensors located deep in hydrocarbon production wells, where temperature can easily reach 100C, sometimes 'go crazy' and provide a single measured value that is at least three or four times higher. Here too the value and the position of outliers are generated on a random basis. As in our previous experiment, the defect is accurately indicated by all of the numerical data quality indicators. In this experiment we observed that in all 3 test cases information entropy only changed a little. On the other hand the intrinsic quality factor and the distance quality factor were indicating much lower values. This is consistent with the conceptual treatment of an outlier as an element whose uncertainty is widely increased and whose value differs substantially from that of the corresponding element on a best fit regression curve. A big difference between actual outlier value and corresponding best fit regression point value lowers both Q and Q_D. A substantial uncertainty increase (which is conceptually necessary so that the element still remains compatible with the original trend despite its increased distance from the regression curve) increases information entropy on the basis of an increase in the order of magnitude, which is addressed by the logarithmic information entropy formula.

The third experiment combines the first two, creating 100, 1000 or 10000 combined gaps and outliers respectively in the dataset. In terms of AVG and STD, the introduction of these defects follows the same pattern described in the previous two paragraphs above. The results of this experiment are quite conclusive as both the distance quality factor and the intrinsic quality factors show a significant decrease in indicator values with the introduction of more defects. The value of the information entropy indicator, on the other side, increases with the increasing number of defects introduced. The quantitative difference (and the spread of values per run) depends on the structure of the indicator. Information entropy is based on the size of the element E_k uncertainty set; this increases substantially in case of gaps or outliers - by an large following the progression 10 – 100 – 1000 if $\mu(E_k)$ increases more than one order of magnitude. This explains the higher $\mathcal{E}$ values with respect to Q or Q_D. The distance-based quality factor only has one parameter (the distance between an element and the corresponding one on a best fit regression curve), so it is characterised by less noise than Q, which has two paramenters (distance and uncertainty) and thus two degrees of freedom that allow a higher noise level in the various experiment rounds.

6 CONCLUSIONS

Our results indicate that numerical data quality can successfully be measured with the indicators intro-

duced and described in this paper. All indicators were successfully showing the effect of gaps and outliers in reducing data quality. Information entropy $\mathcal{E}$ shows how data quality worsens with each experiment that results in more noise. Intrinsic data quality Q and distance-based data quality Q_D show how distance from the best fit curve impacts on overall quality, whether uncertainty is explicitly considered or not. The proposed data quality indicators can be considered as the initial data quality assessment step for a numerical dataset, serving as a precondition to any subsequent data manipulation. By knowing what is wrong with the evaluated dataset, appropriate cleaning and improvement techniques can be applied, or in case of low and non-improvable quality indicators, the dataset can be deemed as unreliable.

ACKNOWLEDGEMENTS

This research is funded by EPSRC Doctoral Training Partnership 2016-2017 University of Aberdeen with award number: EP/N509814/1

REFERENCES

Batini, C. et al. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, 41(3):1–52.

Batini, C. and Scannapieco, M. (2006). *Data Quality: Concepts, Methodologies and Techniques (Data-Centric Systems and Applications)*. Springer.

Cai, L. and Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. *Data Science Journal*, 14(0):2.

De Amicis, F., Barone, D., and Batini, C. (2006). An analytical framework to analyze dependencies among data quality dimensions.

Deming, W. E. (1991). Out of the crisis, 1986. *Massachusetts Institute of Technology Center for Advanced Engineering Study iii*.

Dobyns, L. and Crawford-Mason, C. (1991). Quality or else: The revolution in world business. *Regional Business*, 1:157–162.

Group, D. U. W. (2013). The six primary dimensions for data quality assessment. *DAMA UK*.

Hufford, D. (1996). Data warehouse quality. *DM Review, January*.

Juran, J. (1989). Juran on leadership for quality: An executive handbook.

Li, L. (2012). *Data quality and data cleaning in database applications*. PhD thesis, Napier University, Edinburgh.

Loshin, D. (2001). Dimensions of data quality. *Enterprise Knowledge Management*, page 101–124.

Marev, M., Compatangelo, E., and Vasconcelos, W. W. (2018). Towards a context-dependent numerical data quality evaluation framework. Technical report, Computing Sci. Dept., University of Aberdeen.

Olson, M. H. and Lucas Jr, H. C. (1982). The impact of office automation on the organization: some implications for research and practice. *Communications of the ACM*, 25(11):838–847.

Pipino, L. L., Lee, Y. W., and Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM*, 45(4):211.

Redman, T. C. (2008). *Data Driven: Profiting from Your Most Important Business Asset*. Harvard B. R. P.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3).

Swanson, E. B. (1987). Information channel disposition and use. *Decision Sciences*, 18(1):131–145.

Todoran, I.-G., Lecornu, L., Khenchaf, A., and Caillec, J.-M. L. (2015). A methodology to evaluate important dimensions of information quality in systems. *JDIQ*.