Chapter

Visual Identification of Inconsistency in Pattern

Nwagwu Honour Chika, Ukekwe Emmanuel, Ugwoke Celestine, Ndoumbe Dora and George Okereke

Abstract

The visual identification of inconsistencies in patterns is an area in computing that has been understudied. While pattern visualisation exposes the relationships among identified regularities, it is still very important to identify inconsistencies (irregularities) in identified patterns. The significance of identifying inconsistencies for example in the growth pattern of children of a particular age will enhance early intervention such as dietary modifications for stunted children. It is described in this chapter, the need to have a system that identifies inconsistencies in identified pattern of a dataset. Also, techniques that enable the visual identification of inconsistencies in patterns such as fault tolerance and colour coding are described. Two approaches are presented in this chapter for visualising inconsistencies in patterns namely; visualising inconsistencies in objects with many attribute values and visual comparison of an investigated dataset with a case control dataset. These approaches are associated with tools which were developed by the authors of this chapter: Firstly, ConTra which allows its users to mine and analyse the contradictions in attribute values whose data does not abide by the mutual exclusion rule of the dataset. Secondly, Datax which mines missing data; enables the visualisation of the missingness and the identification of the associated patterns. Finally, WellGrowth which explores Children's growth dataset by comparing an investigated dataset (data obtained from a Primary Health Centre) with a case control dataset (data from the website of World Health Organisation). Instances of inconsistencies as discovered in the explored datasets are discussed.

Keywords: missing data, contradictory, inconsistent, pattern, ConTra, visualisation, bad data

1. Introduction

It is often said that data is the lifeline of research. Due to the importance of data, several research areas such as machine learning, data science, data mining, data analytics and big data has been devoted to the full study and understanding of data. The use of data driven marketing (DDM) as an effective tool in determination of a strategic part of business management is proposed in [1] while the development of data-driven planning for management decision making is advocated in [2]. Also, there is a need for data driven research through open data source [3]. Also, it is noted in [4] that in order to effectively plan an experiment, there is need for preliminary data as a starting point. Even so, the need for valid data in research cannot

be overemphasised. In fact, invalid and inconsistent data could inadvertently impart negatively on results of a research. The authors in [5] pointed out the importance of data validation for systematic software development. Similarly, the authors in [6] explained the importance of health records for diagnosis and treatment purposes. In general the need for valid data is indeed a concern that cuts across every research area. The study of big data has been found to have great impact on scientific discoveries and value creation [7]. The study continues to gain recognition as the quest for tools and measures for validating data continues. Also, [8] explains that the presence of noise hampers the induction of Machine Learning models from data, and can also make the training time longer. Noisy data according to [9] cannot be avoided but rather dealt with. Data, whether structured, semi-structured, or unstructured must be scrutinised with utmost care. The rigour of validating data could be tasking and are usually left in the hands of data scientists.

Data scientists acquire datasets from different environments which in most cases could be noisy. A noisy dataset contains uncertain and inconsistent data that could arise from missing values, imprecise data, and contradictory values, among others. The work of a data scientist includes among others, to explore big dataset in order to find interesting patterns and build supporting arguments for decision making. Such interesting patterns are likely to exclude noise in the form of conflicting or missing data in the dataset which do not support the arguments presented by the analyst. Data which are inconsistent with decision supporting facts should also be analysed. An approach to this analysis is to visually explore the patterns of the decision support data and associated inconsistencies.

Visual analytics is defined in [10] as the science of analytical reasoning facilitated by interactive visual interfaces. Data visualisation is useful for data cleaning, exploring data structure, detecting outliers and unusual groups, identifying trends and clusters, spotting local patterns among others [11]. The visual platform and representations enables better understanding and facilitates analytic as well as deductive reasoning. On the same note, visual analysis of data is important in understanding data and has been found to yield fruitful results in research. According to [12], visual analysis of data enables grasping the multidi-mensional "information reality" from the perspective of users. Visual analytics entails more than a mere visualisation. In fact, it can rather be seen as an integral approach to decision-making, combining visualisation, human factors and data analysis [13]. Visual analytics from another perspective is a data representation approach that employs interactive visualisation to integrate human judgement into algorithmic data-analysis processes [14]. Thus, visual representation of data plays a vital role in data interpretation and analysis.

It is important to analyse interesting patterns and associated noise from big datasets so as to identify the hidden patterns and knowledge in them. Unfortunately, some data scientists advocate deleting or not including the noisy data instead of visually depicting the noise and reporting the analyses. Certainly, deleting inconsistent data from a noisy dataset will increase the incompleteness in the dataset thereby reducing the soundness of the information retrieved from the dataset. Consequently, the noise in a dataset should be tolerated and its tolerance will enable the avoidance of losing interesting information about the dataset. The analysis of incomplete biological data of an organism for example, enhances the understanding of the abnormalities in the organism. Incomplete biological data existing in datasets from laboratory investigations such as data about genes and proteins provides clues to genetic disorder.

The importance of identifying inconsistencies in pattern can also be evident in survey dataset. A survey on pattern of menstruation can reveal a pattern that ladies between the ages of 20 to 30 years old who have not seen their menstruation

for more than two months are pregnant. This pattern does not mean that all ladies of this same age bracket who have not seen their menstruations for more than two months from the survey data are pregnant. Obviously, there can exist ladies suffering from Hormonal aberrations for instance.

Also, respondents to survey questions may provide inaccurate responses, such as giving many consecutive items a response of "4" or repeating a pattern of "1, 2, 3, 4, 5…." as explained in [15]. Such purposefully deceptive or even contradictory responses are herein assessed as inconsistencies in patterns and should be portrayed visually as the wrong side of analysis. An example of inconsistency in a survey pattern involving giving many consecutive items a response of "4" is a pattern that shows responses that do not give many consecutive items a response of "4". It is therefore important to identify such inconsistencies in patterns of interest in order to properly provide analytical reports that expose the issues.

The importance of identifying and assessing inconsistent data is explained in works such as [15–17] but very few publications exist in the area of visually identifying inconsistencies in patterns of interest [18]. There is therefore a need to have a system that enables the visual analysis of inconsistencies in patterns of interest in a dataset. This is to provide data users with a holistic understanding of data of interest. It is stated in [18] "Of 612 data visualizations from 121 articles published online in February 2019 by a set of leading purveyors of data journalism, social science surveys, and economic estimates, 449 (73%) presented data intended for inference, but only 14 (3%) portrayed uncertainty visually, either by depicting explicit quantifications like intervals or conveying variance through raw data".

Consequently, the authors of this chapter emphasise the need for visualising inconsistencies in identified data patterns by explaining existing approaches and implementing novel approaches for visual analysis of inconsistencies in patterns. In Section 2, a detailed explanation on the concept of inconsistencies in pattern is given. In Section 3, two approaches for visualising inconsistencies in patterns are presented. The visual analyses of inconsistencies in objects with many attribute values and the visual comparison of an investigated dataset with a case control dataset is described. These approaches and their associated tools which were developed by the authors are discussed in the same section. The WellGrowth application is discussed in the same section. The WellGrowth app integrates the use of fault tolerance and colour coding to visualise inconsistent pattern while using data curated from Nsukka Medical Centre (NMC) and data from the website of World Health Organisation (WHO) as their control studies. A comparison of ConTra, Datax and WellGrowth Apps is presented in Section 4 while Section 5 is the conclusion and research focus for future work.

2. Inconsistencies in patterns

Any inconsistent data associated to a pattern reduces the quality of findings presented by the analyst about the pattern. An assessment of such inconsistent data can increase the trustworthiness of the findings from the analyst. There are everyday instances of inconsistent data in identified patterns which are likely to mar the patterns. Meade and Craig in [15] explain how inconsistent data from careless respondents of students' survey can be identified among data patterns common among respondents of the survey. Patterns derived from survey data can be associated to contradictory or incomplete responses. Also, patterns discovered in biological investigations can be associated to inconsistent and incomplete data. A gene expression dataset whose columns includes gene name, tissue name, expression and experiment ID can contain inconsistent data in an identified pattern where many experiments are performed for a particular gene in a particular tissue. An expression can be detected,

not detected, or not available. If one of the interesting patterns is that a gene "xxx" is "detected" in experiments on tissue "yyy" of an organism at a particular developmental stage, then inconsistency of the pattern from the dataset will exist where there are data that shows that the gene "xxx" is "not detected" in other experiments that investigates the tissue "yyy" of the same organism at the same developmental stage. Also, uncertainty about the presence of the gene "xxx" can exist in the dataset where the information about the presence of the gene in the experiment about the tissue "yyy" at the same developmental stage is missing. Such missing information can be denoted by "unavailable" or empty space, among others. Inconsistent data relating to gene expressions in tissues of different developmental stages are reported in [17, 19]. Finally, a Radiologist chest x-ray report can be used to detect aortic unfolding which is mostly associated with systemic hypertension. However, there are instances of aortic unfolding which are not associated with systemic hypertension. There are also, some instances of aortic unfolding which it is not known if they are associated with systemic hypertension. These instances are inconsistent in a pattern involving systemic hypertension as a cause of aortic unfolding.

2.1 Visual analysis of inconsistencies in patterns of dataset

Inconsistent data which are associated to patterns in a large dataset can be difficult to visualise. This is because they are not explicitly indicated in the dataset as inconsistent. For example, missing data can exist as "unavailable", "forthcoming", "-", "not existing", or even empty spaces. Contradictions on the other hand, differ from one dataset to another, depending on the semantic definition of the data in the dataset. Interestingly, there are dedicated Applications such as CUBIST [19], ConTra [20], and R Package VIM [21] which enables the visualisation of the amount or pattern of contradiction and missingness in a noisy dataset. Inconsistent data whose pattern involves mutually exclusive type of contradictions is depicted by ConTra. Nwagwu explains in [20] how the contradictory attribute values in the gene "TSPAN6" of the tissue "Pancreas" is detected by ConTra and visualised in a pie chart. ConTra applies colour coding on charts to enable the visualisation of inconsistencies in a large dataset. Also, ConTra enables the visualisation of the pattern of distribution of contradictions across the dataset. It is further discussed in Section 3.11.

R Package VIM is a good analytical tool that focuses on visual presentations and analysis of missingness. It is used in plotting the aggregates of missingness in variables of a Barplots. It also shows missing data in a matrix plot, Histogram, Spline plot, Parallel coordinate plots and in Maps [21]. It uses Barplot to show the number and distributions of missing values for a sub-sample of the EU-SILC data from Statistics. Notwithstanding VIM's comprehensive collection of visualisation methods for exploring missing data, its environment requires extensive training in R skills in order to access its visualisation methods. Also, the VIM package does not enable the analysis of other types of inconsistencies such as contradictions in a dataset apart from the missingness.

There are other tools which enables the visualisation of inconsistencies as explained in [19, 22, 23]. A graphical tool is proposed in [22] that highlight inconsistent instances in the network such as the highlights of direct comparisons that strongly drive other treatment effect estimates and hot spots of network inconsistency. It also proposed a clustering approach that automatically groups comparisons for highlighting hot spots. CUBISTs [19] is an example of an application that applies colour coding and fault tolerance in traditional visualisation tools such as pie or bar chart to enable easy visual analysis of inconsistencies. Even so, these applications are not holistic in exploring inconsistencies in patterns and most of them are designed for particular domain of data analysis.

The analysis of inconsistencies in patterns of a dataset can be enhanced by adapting computational techniques such as fault tolerance and colour coding in traditional visualisation tools such as graphs to enhance the visualisation of inconsistencies in patterns. Fault tolerance necessitates the introduction of softness (statistical defined tolerance) in retrieving the inconsistencies in a dataset. Colour coding necessitates identifying the different ranges of inconsistencies with different colours. Section 3.0 presents how these computational techniques are used in computing inconsistencies in pattern as integrated in the approaches presented in this chapter.

3. Our approaches

Two approaches are presented for visualising inconsistencies in patterns in this section namely; visualising inconsistencies in objects with many attribute values and Visual comparison of an investigated dataset with a case control dataset. These approaches and their associated tools which were developed by the authors are discussed in this section.

3.1 Visualising inconsistencies in objects with many attribute values pattern

A dataset contains data about real world objects. These data contains objects which are associated to attributes and the attributes can be associated to single or many values. Real world objects 'G' such as house, book, car, and television are associated with different attributes 'M' which may have many values 'V'. A book (object) for example, can have colour (attribute) which can be black, white or brown (values). It can be established that particular object ($g \in G$) is associated with an attribute ($m \in M$) which contains many values. For example, a name (object) has marital status (attribute) such as married or single (values). Contradictory data can exist in a dataset when there is conflicting information such that an object ($g \in G$) that is associated with an attribute ($m \in M$), contains contradictory values such that m is associated with A and ¬A. An experiment (object) for example, can be associated with outcome (attribute) such as neutral, high, or low (values). A student (object) took a course (attribute) whose values can be absent, pass or fail. Some of the many valued attribute are likely to be mutually exclusive and should conform to mutual exclusion rule. The mutual exclusion rule can simply be stated that real world objects whose attribute values are mutually exclusive (meaning more than one attribute values cannot be associated with the object at once) are contradictory. Also, any attribute which do not contain the expected values is said to contain missing data.

Two open source tools are presented in this chapter to explain how to visualise inconsistencies in objects with many attribute values pattern namely ConTra and Datax. ConTra is discussed in an earlier publication [20] by some of the authors of this chapter and it is also discussed herein. Datax is another tool for highlighting inconsistency in patterns through mining and depicting missing data is presented in Section 3.12.

3.2 ConTra

ConTra[1] is an open source App developed by some of the authors of this chapter and it is used for mining contradictory data from attributes with many values

[1] https://github.com/ncjoes/contra

pattern where the contradictory data are objects associated with mutually exclusive attribute values. It enables its users to query attributes of particular objects whose attribute values are mutually exclusive and display the percentage of the values that contradicts the mutual exclusive rules and the percentage of the values that abide by the rule in a pie chart. Algorithm 1 displays the pseudocode for mining objects whose attribute values are contradictory and those whose attribute values are consistent.

ConTra was used to analyse objects in a Comma Separated Values (CSV) dataset containing over a million rows and six columns. The dataset 'Normal Tissue' dataset is deposited in [24] and it contains expression profiles for proteins in human tissues. It consists of the following columns: 'Gene', 'Gene name', 'Tissue', 'Cell type', 'Level', and Reliability. It has a size of 79.5 MB. Normal Tissue dataset reports experiments on tissues and identified gene expression levels such as low, medium, and high. It also indicates the annotated cell type ("Cell type") and the gene reliability. There can be multiple records for the same gene from different investigations (experiments) on the same tissues in Normal Tissue dataset.

Algorithm 1: ConTra's Algorithm for mining contradictory and consistent data as evident in [11]

1. Given a set of records in CSV format
2. Let G = Set of Objects from a selected column
3. Let M = Set of Attributes (titles of every column excluding the Object column)
4. Let O(a,b) = empty list where a = contradictory object index and b = contradictory attribute values
5. Let C(c,d) = empty list where c = consistent object index and d = consistent attribute values
6. For each Object 'g' in the set of objects 'G' which are associated to a set of mutually exclusive attributes 'M'
   i. If 'M' contains more than one mutually exclusive value then
      Store 'g' and also store each of the contradictory values in the list O(a,b)
   ii. Else
      Store 'g' in set of consistent objects and also store each of the consistent values in the list C(c,d)

End

7. Print contradictory objects O(a, b) and consistent objects C(c, d)

3.2.1 Evaluation of ConTra

ConTra was used to analyse the Normal Tissue dataset. Any experiment on a tissue in Normal Tissue dataset which indicates that its identified gene expresses more than one level of expression such as not detected, medium, high or low is inconsistent. As identified through the use of ConTra and discussed in [20], contradictions exist in two of the records (9.09%) of the gene 'TSPAN6' expression levels in the tissue 'Pancreas' in Normal Tissue dataset. This is depicted graphically in **Figure 1** as adopted from [20]. Evidently from **Figure 1**, it will be wrong to state that the pattern of expression of the gene 'TSPAN6' in the tissue 'Pancreas' of Normal Tissue dataset is of a particular level. This is because there are cases of contradictory expression in the associated data (TSPAN6 expression levels). Consequently, a holistic analysis of the expression levels of TSPAN6 on Pancreas in the Normal Tissue dataset should depict the existence of the contradictory data as shown in **Figure 1**.

ConTra provides a platform for visualising such inconsistencies in datasets whose objects exhibit a many attribute value pattern and are associated with mutual exclusive attribute values.

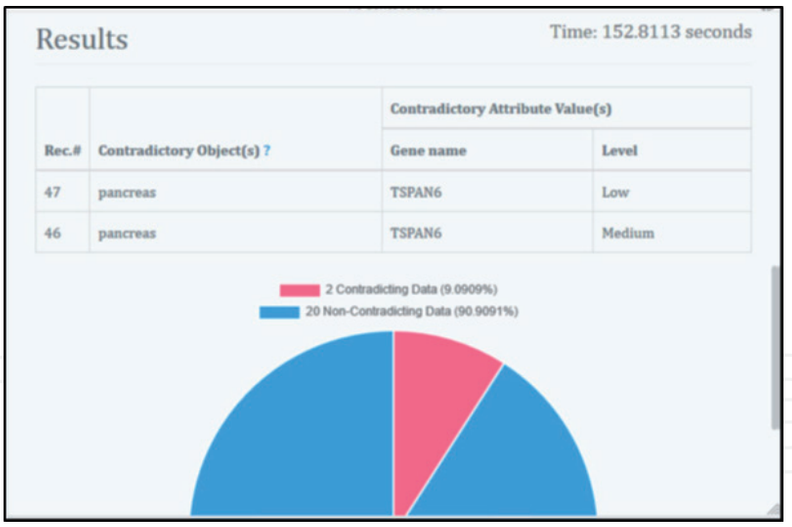

Figure 1.
Result of the analysis of the normal tissue dataset by ConTra's multiple attribute values approach.

3.3 Datax

Datax[2] is an open source application that mines missing data and associated patterns from a Comma Separated Values (CSV) Dataset. It is designed to enable the visualisation of the missing data in attribute values of a dataset by generating charts which depicts the incompleteness and any associated pattern. It has the following features:

- Ability to load and store CSV datasets for further visualisation
- Ability to display the statistics of incomplete data in a stored dataset
- Ability to visualise through matrix or bar plot, amount and distribution (patterns) of missingness in a selected dataset

The user of Datax can select attribute(s) or column(s) of interest from a dataset to visualise the missingness in them. Bar charts are programmed to use white lines to dynamically indicate the missingness in a dataset. Other important parameters measured in Datax include the number of columns in the investigated dataset and the percentage of missingness in each column.

Datax was used to mine incomplete data in an Amazon open source dataset[3]. The Amazon dataset has a size of 365.82 MB. It contains a list of over 1,500 consumer reviews of Amazon products such as the Kindle, and Fire TV Stick as provided by Datafiniti's Product Database[4]. It has a total of 27 columns which includes basic product information such as rating, review text, and more for each product. It also has a total of 1598 rows.

3.3.1 Evaluation of Datax

Datax was used to analyse the Amazon product review dataset as provided by Datafiniti's Product Database. **Figure 2** depicts the evaluation pane and shows a sneak peek into the first five rows of the investigated dataset while the right side

[2] https://github.com/marioJoker/Datax

[3] https://data.world/datafiniti/consumer-reviews-of-amazon-products

[4] https://datafiniti.co/products/product-data/

displays the statistical analysis of the dataset. The statistical analysis displayed includes: the names of the investigated columns, the total number of missing values per column of the dataset, and the percentage of the missing data per column (**Figure 3**).

As evident in **Figure 4**, the amount of missingness is indicated by the heights of the bars. Bars with equal height indicate joint missingness in investigated attributes. A Datax Bar plot reveals the amount of missing data and commonalities of such instances across the dataset. It can be observed in **Figure 4** that the columns V, W and Y have the same amount of missingness. Also the column headings U and X have the same amount of missing instances. Obviously, any analysis that includes columns whose pattern indicates significant amount of missingness should acknowledge such missingness in its reports. Columns that do not have missing data are also revealed in **Figure 4**. For example, columns A, B, C, D, F, G, J, M, N,

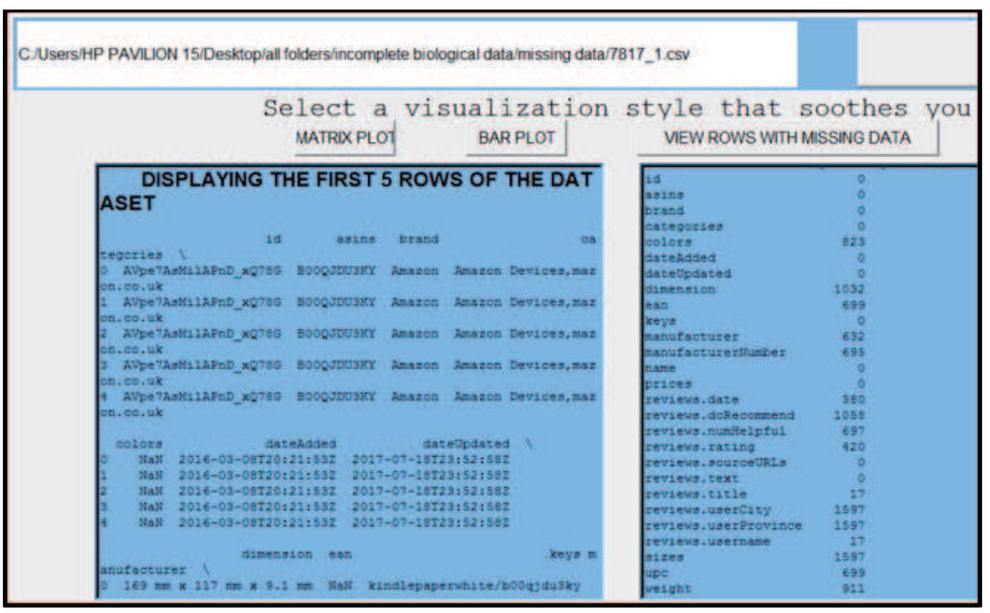

Figure 2.
Datax evaluation pane depicting the number of missing data per investigated column.

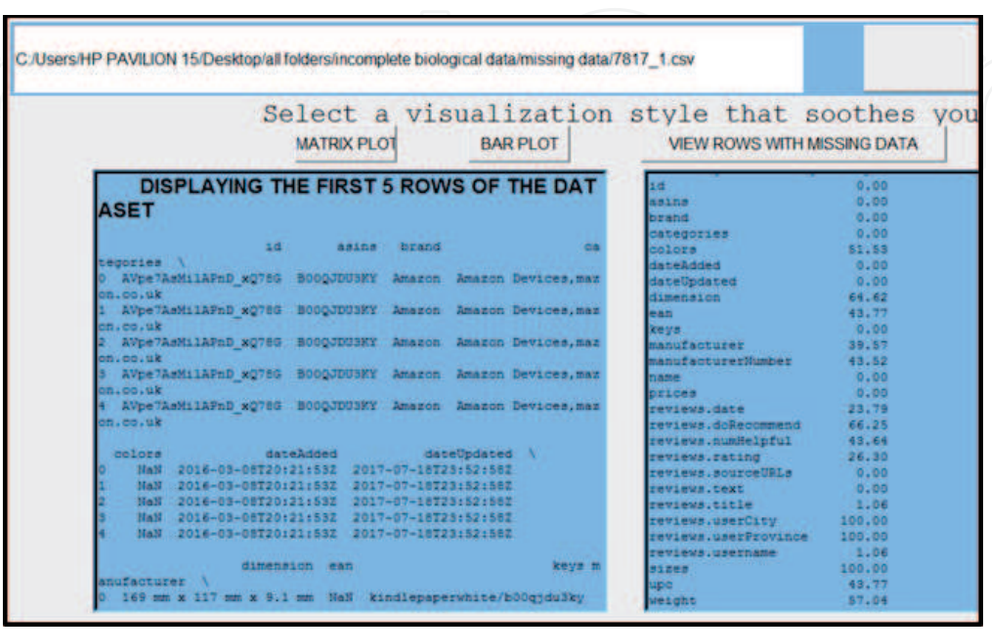

Figure 3.
Datax evaluation pane depicting the percentage of Missingness per investigated column.

S, and T do not have any missingness associated to them. Any assertion made by a data analyst about any column should first be evaluated for the relevance of missing data. Datax's Bar plot do not however, show the distribution of missingness among its investigated columns. This is explored by the matrix plot as depicted in **Figure 5**.

A Matrix plot of missing data as evident in **Figure 5** reveals the amount and distribution of missingness in the dataset. White colour is used for missing values while black colour is used for the available data values. It can be observed from **Figure 5** that the columns V, W and Y have 1597 missing data in common. The column headings U and X have equal amount of missing instances implying that each

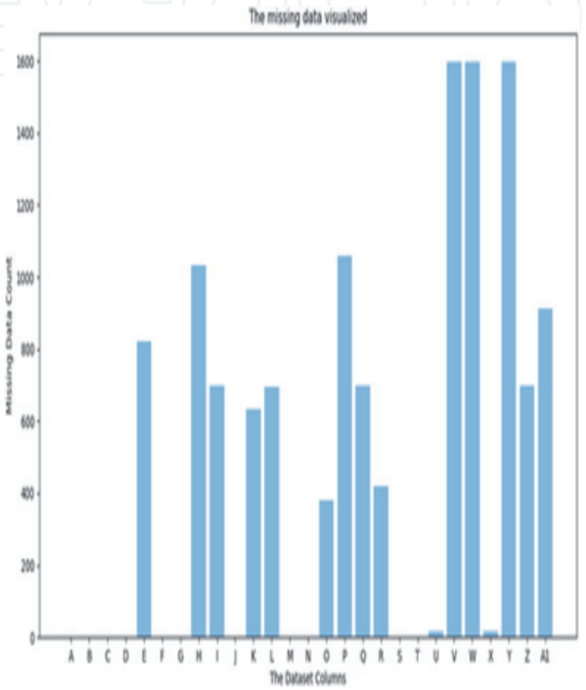

Figure 4.
Datax Bar plot depicting the amount of Missingness per investigated column.

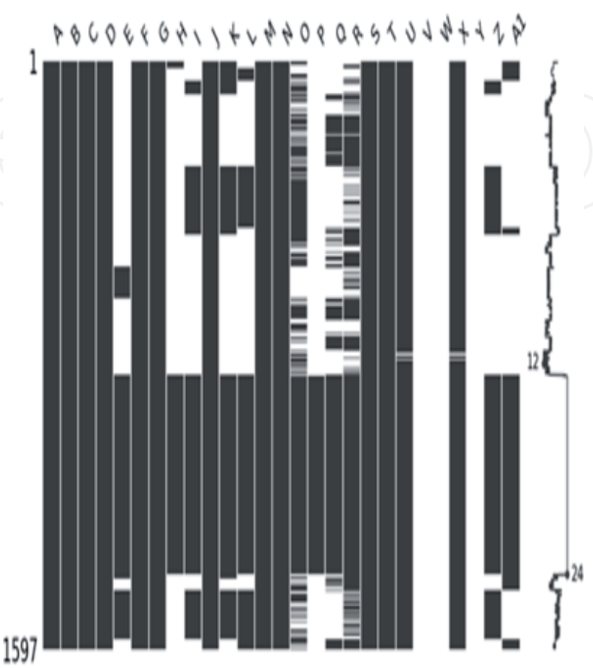

Figure 5.
DataxMatrix plot depicting the amount and distribution of Missingness and available data per column.

reviewer that did not fill data in U, did not also fill data in X. The same observation holds for columns I and Z which have same distribution of missingness. The data analyst should make efforts to understand the relationships among the columns with joint and same distribution of missingness to present a robust report about the missingness in any discovered pattern.

Datax has also been used to evaluate cell phone reviews on the amazon online shopping store. The dataset is also deposited along Datax open source code[5]. It contains 11 columns and 1,048,576 records. Datax was evaluated by a team of software developers in University of Nigeria, Nsukka and they described its efficiency in mining missing data and visualisation of associated patterns as excellent. Even so, it does not visualise the different forms of missing data. It specifically mines empty cells without noting representations such as " -", "not existing", "not available", among others as missing data. The authors hope to integrate this ability in the next update of the application.

3.4 Visual comparison of an investigated dataset with a case control dataset

The visualisation of inconsistent data can be achieved through direct comparison of an investigating dataset with a case control dataset. Investigations that involve a comparison of an investigating dataset with a standard dataset are scenarios in which this approach can be used. This section of this chapter describes how WellGrowth app is used to enable the visual comparison of an investigated dataset with a case control dataset. It also describes the datasets investigated and how WellGrowth App was used in the investigation of the datasets.

3.5 Case control method

The case–control studies approach was used in comparing two datasets where one of the datasets is the case control while the other is the investigated dataset. World Health Organisation[6] (WHO) is the case control dataset and the dataset generated from Nsukka Medical Centre (NMC) is the investigated dataset. WHO data is gotten from children's empirical data which includes the length/height and weight of children at different stages of their growth for a sex matched reference. The weight and length of the children's data from WHO child growth standards for 0–12 months were used in investigating the NMC data. The average (50th percentile) score of the different children's weights and lengths in each month was used in the case control studies. This dataset is stored in WellGrowth app open source (see Section 3.22) for further analysis. The researchers collected the data (length for age and weight for age percentiles for girls and boys) directly from WHO web site (https://www.who.int/toolkits/child-growth-standards/standards/length-height-for-age).

NMC data are the weight and length of the children's growth data from 0 to 12 months for a sex matched references which are collected from the medical center Nsukka. The average (50th percentile) score of the different children's weights and lengths in each month as curated from NMC was used in the case control studies as the investigated dataset. This dataset is stored in WellGrowth app open source (see Section 3.22) for further analysis. The data collected from the NMC were not classified. So, the researcher classified them into different files according to the sex and the growth parameter per months from 0 to 12 months. The data were collected from NMC from May to august of 2020. **Table 1** presents a record of the number

[5] https://github.com/marioJoker/Datax/tree/master/amazon-cell-phones-reviews

[6] https://www.who.int/toolkits/child-growth-standards/standards/length-height-for-age

Sex	Height/Length	Weight
Girl	451	451
Boy	497	497

Table 1.
Number of children data collected according to the sex for the length and weight.

Age/Sex	Girls		Boys	
	Length	Weight	Length	Weight
0	54.75	3.95	52.5	4.3
1	57.5	5.449	58.75	5.85
2	61.5	6.75	63.75	7.1
3	62.75	6.5	65	7.05
4	63.5	6.7	66	7.6
5	64	6.8	66.5	7.1
6	65	7.4	66.75	8.1
7	65.78	8	67.2	9.5
8	66.25	7.5	67.7	9.5
9	67.25	8.4	70	8.9
10	67.8	8.8	72	9.5
11	68	8.5	72.45	9
12	68.5	8.4	72.9	9.5

Table 2.
The 50th percentile of data collected from NMC.

child's data collected by the researchers in NMC while **Table 2** presents the 50th percentile of the data collected from NMC.

3.6 WellGrowth

WellGrowth[7] app enables the visualisation of inconsistent data through direct comparison of an investigating dataset with a case control dataset. This is achieved through the visual evaluation of inconsistencies in children's growth pattern using the dataset from World Health Organisation[8] (WHO) as the case control dataset and dataset generated from Nsukka Medical Centre (NMC). These datasets are stored in the WellGrowth App for further evaluation by the App users.

WellGrowth adopts the average (50th percentile) score of WHO's children growth data for each month from 0 to 12 months in building WHO's growth curve. WHO's children growth data are gotten from children's empirical data such as height and weight at different stages of their growth for a sex matched reference. WellGrowth also integrates children growth data collected from the NMC. The average (50th percentile) score of the different children's weights and lengths in each month as curated from NMC are used to build the NMC/local growth curve. Finally, the individual growth curve is generated from inputs of a child's monthly weight/length as keyed into the WellGrowth input form by the WellGrowth App

[7] https://github.com/dora-png/growth-of-child

[8] https://www.who.int/toolkits/child-growth-standards/standards

user. A visual comparison of growth patterns from WHO to the growth patterns from NMC is used to enable the analysis of inconsistencies in children's growth data from Nsukka Medical Centre. Also, input of a child's growth data is used by WellGrowth app to enable a visual comparison of growth patterns of a Child with WHO's growth curve.

The authors designed WellGrowth for plotting growth pattern graphs from WHO, NMC and user's input data. WellGrowth App adopts colour coding and fault tolerance to enable easier visualisation of inconsistencies in their investigated datasets. For example, the average growth data whose value is less than 2 units from WHO's data value are yellow; the average growth data whose values are greater or equal to 2 and less than 4 are filled with red; while those values are greater or equal to 4 are filled with a yellow colour. Further details of WellGrowth implementations are expected in another publication by the authors.

3.7 Evaluation of WellGrowth app

Figure 6 presents individual, WHO and local (NMC) growth graphs showing the growth pattern (weight) of children whose ages are 0 to 12 months. **Figure 7** present a print view of Individual, WHO and local (NMC) WellGrowth's graphs. The average growth data whose values are less than 2 units from WHO's data value are indicated by yellow line while the average growth data whose values are greater or equal to 2 are filled with red line. **Figure 7a** shows the weight and **Figure 7b** shows the length of children whose ages are 0 to 12 months. It is evident from **Figure 7a** that there is no instance of inconsistency given the tolerance level of less than 2 units from WHO's a data value. The individual

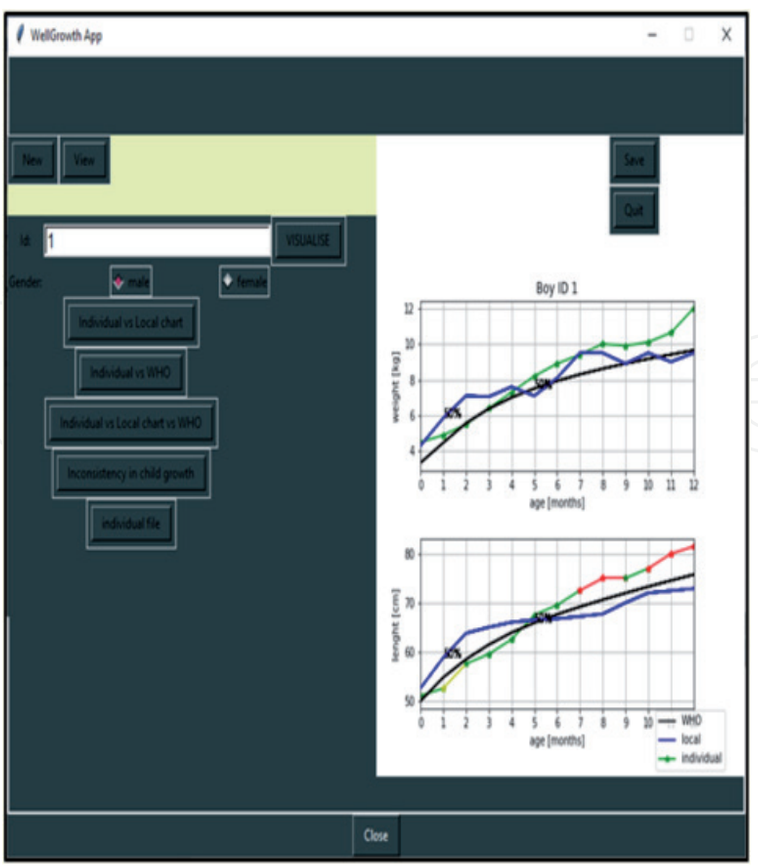

Figure 6.
WellGrowth app visualisation of growth patterns of individual vs. WHO vs. local (NMC) age graph after 12 months (weight in first graph vs. length in second graph).

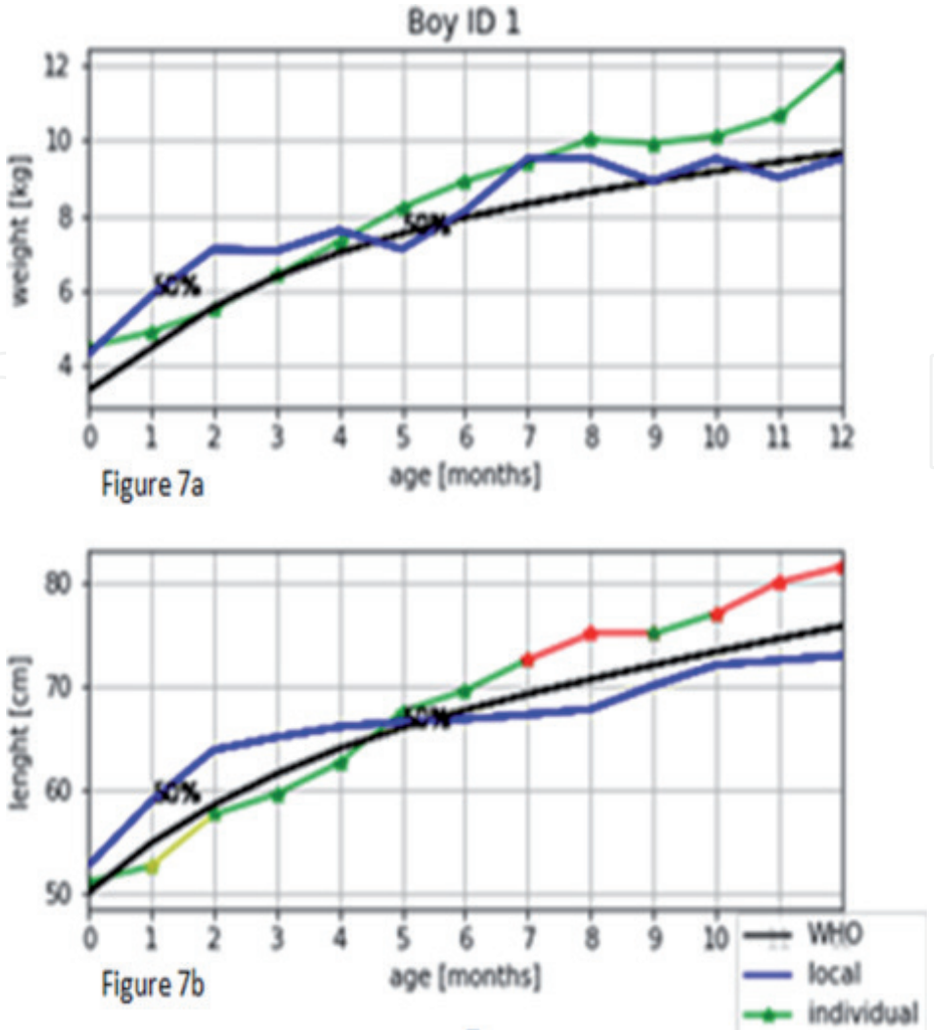

Figure 7.
*A print view of WellGrowth App's individual, WHO and local (NMC) growth graphs showing the weight (**Figure 7a**) and length (**Figure 7b**) growth patterns of children whose ages are 0 to 12 months.*

graph in **Figure 7b** is consistent with the WHO's growth pattern but there are issues of inconsistency in the 1st and 2nd month as indicated by the yellow line. Inconsistencies is depicted in **Figure 7b** individual line graph where the average growth data values are greater or equal to 2 and are indicated with red coloured line. The individual graph is consistent with the growth pattern of the WHO in the 5th to 7th month and 9th to 10th month but there are issues of inconsistencies in the 7th and 9th month and 10th to 12th month as indicated by the red line (see **Figure 7b**).

4. Comparison of ConTra, Datax and WellGrowth apps

Visual identification of inconsistencies in established patterns is achievable through data mining and analysis tools such as ConTra, Datax and WellGrowth apps. Each of these tools has its area of applicability depending on the kind of inconsistency explored. Datax for example, is most appropriately used for visualising patterns of missingness in CSV datasets unlike ConTra or WellGrowth that are used for mining and visualising contradictory data in patterns. **Table 3** presents a summary of the appropriateness of each of the tools in visualising inconsistencies in established patterns.

	ConTra	Datax	WellGrowth
Pattern of missingness	×	✓	×
Amount of missingness	×	✓	×
Amount of contradiction	✓	×	✓
Pattern of contradictory values	✓	×	✓
Colour coding	✓	×	✓
Fault tolerance	×	×	✓

Table 3.
Comparison of ConTra, Datax and WellGrowth apps.

Six yardsticks were used in comparing the appropriateness of the explored tools and they include: pattern of missingness, amount of missingness, amount of contradiction, pattern of contradictory values, colour coding, and fault tolerance. ConTra and WellGrowth for example, does not mine missingness nor explore the pattern of missingness in a dataset. They do not measure the amount of missingness, unlike Datax that is designed to evaluate both the pattern and amount of missingness using Matrix Plot and bar charts respectively. It is evident from our discussions in this chapter, that ConTra and WellGrowth apps are used to explore inconsistencies notably contradictory data in established patterns of interest. In doing this, WellGrowth apps adopt colour coding and fault tolerance while Datax only adopts colour coding. **Table 3** depicts these discussed yardsticks for comparing ConTra, Datax, and WellGrowth apps.

5. Conclusion and research focus for future work

This chapter has focused on the discussion of identifying inconsistencies associated with patterns. Even so, it has restricted its discussions to instances of contradictory data, deviations from standard data and missing values. Real life examples and open source datasets were used to illustrate our proposed approaches. The researchers anticipate that this interesting but understudied area of computing should be explored further by computer scientist to avoid instances of misinformation by our data analysts. Novel approaches for visual analysis of inconsistencies should be proposed. Also better means of diagrammatically visualising inconsistencies in pattern should be initiated.

DOI: http://dx.doi.org/10.5772/intechopen.95506

Author details

Nwagwu Honour Chika[1*], Ukekwe Emmanuel[1], Ugwoke Celestine[2], Ndoumbe Dora[1] and George Okereke[1]

1 University of Nigeria, Nsukka, Enugu State, Nigeria

2 Nice diagnostic clinic Enugu State, Nigeria

*Address all correspondence to: honour.nwagwu@unn.edu.ng

References

[1] Veynberg RR, Timofeev A, Popov AA, Bortsova DE. Data driven marketing as a new approach to business development and sales methods. Espacios. 2018;39(12):3.

[2] Iyengar R, Mahal AR, Felicia UN, Aliyu B, Karim A. Federal policy to local level decision-making: Data driven education planning in Nigeria. International Education Journal: Comparative Perspectives. 2015;14(3):76-93.

[3] Krotoski AK. Data-driven research: open data opportunities for growing knowledge, and ethical issues that arise. Insights: the UKSG journal. 2012 Mar 7;25(1):28-32.

[4] Ezer D, Whitaker K. Point of View: Data science for the scientific life cycle. eLife. 2019 Mar 6;8:e43979.

[5] Patil MV, Yogi AN. Importance of data collection and validation for systematic software development process. Int'l Journal of Computer Science & Inf. Technology. 2011;3(2).

[6] Marinič M. The importance of health records. Health. 2015 May 5;7(05):617.

[7] Wang L, Alexander CA. Machine learning in big data. International Journal of Mathematical, Engineering and Management Sciences. 2016 Apr;1(2):52-61.

[8] Garćia LP, de Carvalho AC, Lorena AC. Noisy data set identification. InInternational Conference on Hybrid Artificial Intelligence Systems 2013 Sep 11 (pp. 629-638). Springer, Berlin, Heidelberg.

[9] Rao PS. Study and Analysis of Noise Effect on Big Data Analytics.

[10] Thomas JJ, Cook KA. A visual analytics agenda. IEEE computer graphics and applications. 2006 Jan 10;26(1):10-3.

[11] Unwin A. Why is data visualization important? What is important in data visualization?. 2.1. 2020 Jan 31;2(1).

[12] Cisek S, Krakowska M. Qualitative analysis of visual data in information behavior research. Zagadnienia Informacji Naukowej-Studia Informacyjne. 2019 May 6;57(1 (113)):7-25.

[13] Keim D, Andrienko G, Fekete JD, Görg C, Kohlhammer J, Melançon G. Visual analytics: Definition, process, and challenges. InInformation visualization 2008 (pp. 154-175). Springer, Berlin, Heidelberg.

[14] Cui W. Visual analytics: a comprehensive overview. IEEE Access. 2019 Jun 19;7:81555-73.

[15] Meade AW, Craig SB. Identifying careless responses in survey data. Psychological methods. 2012 Sep;17(3):437.

[16] Lin J, Keogh E, Lonardi S. Visualizing and discovering non-trivial patterns in large time series databases. Information visualization. 2005 Jun;4(2):61-82.

[17] Nwagwu HC, Orphanides C. Visual analysis of a large and noisy dataset. International Journal of Conceptual Structures and Smart Applications (IJCSSA). 2015 Jul 1;3(2):12-24.

[18] Hullman J. Why Authors Don't Visualize Uncertainty. IEEE transactions on visualization and computer graphics. 2019 Aug 19;26(1):130-9.

[19] Melo C, Aufaure MA, Orphanides C, Andrews S, McLeod K, Burger A. A conceptual approach to gene expression analysis enhanced by visual analytics.

InProceedings of the 28th Annual ACM Symposium on Applied Computing 2013 Mar 18 (pp. 1314-1319).

[20] Nwagwu HC, Okereke G, Nwobodo C. Mining and visualising contradictory data. Journal of Big Data. 2017 Dec;4(1):1-1.

[21] Templ M, Filzmoser P. Visualization of missing values using the R-package VIM. Reserach report cs-2008-1, Department of Statistics and Probability Therory, Vienna University of Technology. 2008 May 1.

[22] Krahn U, Binder H, König J. A graphical tool for locating inconsistency in network meta-analyses. BMC medical research methodology. 2013 Dec 1;13(1):35.

[23] White IR, Barrett JK, Jackson D, Higgins JP. Consistency and inconsistency in network meta-analysis: model estimation using multivariate meta-regression. Research synthesis methods. 2012 Jun;3(2):111-25.

[24] The Human Protein Atlas. http://www.proteinatlas.org/about/download Accessed 4 May 2020