Analyzing Missing Data in Metric Spaces[1]

Safia Brinis, Agma Juci Machado Traina, Caetano Traina Jr

Database and Images Group - GBdI
Institute of Mathematics and Computer Science - ICMC
University of São Paulo - USP
São Carlos, SP, Brazil
{sbrinis, agma, caetano}@icmc.usp.br

Abstract. Similarity search in multimedia databases has challenged researchers for the last two decades, whose studies resulted in several achievements. However, searching in incomplete databases, i.e., databases with missing attribute values, has been less studied so far. In this article, we present a set of experimental analyzes that evaluate the impact of missing data on the query performance in metric spaces. The results show that missing data cause severe skew in the metric space with only 2% of missing values and drastically affect the performance of the metric indexing techniques. Interestingly, our analyzes, confirmed by the presented experiments, show that data missing not at random are more prone of skew and raise the conditions of distance concentration phenomenon where the distances between pairs of elements in the space become homogeneous. Thus, this study provides an understanding of the issues involved with metric spaces when indexing incomplete databases and gives ground for research that supports the development of advanced metric access methods with handling of missing attribute values.

Categories and Subject Descriptors: H.1.4 [**Data Management Systems**]: Incomplete Data

Keywords: Distance Concentration, Data Distribution, Missing attribute values, Similarity Search

1. INTRODUCTION

Advanced database applications often deal with imperfect data, i.e., erroneous or incomplete. Such applications usually involve large databases of complex and high-dimensional data objects, therefore, there is always a possibility of having erroneous or missing data.

The way data are missing depends mainly on the application domain, and there are a number of reasons why data may not be observed. For example, when filling a survey, some questions may be intentionally or unintentionally skipped by the participants, or when collecting medical tests, patients may fail to appear for testing and then cause lack of information in their records. Another example concerns time series databases. There are a variety of application domains where time series analysis is particularly useful, such as signal processing, pattern recognition, statistics, mathematical finance, weather forecasting, earthquake prediction, among others. However, the transient interruptions in sensor readings, such as those caused by incorrect hardware design, improper calibration, or by low battery levels, may lead to voids in the time series.

There are two forms of missingness: missing records (i.e., objects) and missing attribute values (i.e., some but not all attribute values from an existing object). While missing records may only be problematic in restricted situations (e.g. smaller size of the sample provides lower statistical power of the data), missing attribute values can severely affect the quality of data and the performance of tasks that operate on such data. In the present work, we investigate the case of missing attribute values.

For any type of data, missingness can be *legitimate* or *illegitimate*. Legitimate missingness is an absence of data where there is no suitable value for the missing data; illegitimate missingness refers to the absence of data despite being able to get the value in principle. Legitimate missingness is

[1]This work has been supported by CAPES under Grant No. DS-7843261/D, CNPq and FAPESP.

very common in surveys. For example, when filling out a survey that asks whether the participant is employed, and if so, what is the value of his/her salary. If the participant is not employed, then, it is legitimate to skip the value of the salary. Illegitimate missingness also exists in all types of data, nevertheless, it is becoming more prevalent in modern databases because of the exotic nature of the data, such as images, time series, and videos. It happens, for instance, when an equipment stops recording data unexpectedly. Therefore, the diversity of the equipments used to collect data, such as X-ray generators, sensors, video recorders and finger print scanners, and their susceptibility to malfunctions and failures often lead to illegitimate missing data. Illegitimate missingness is one of the most pursued problems by researches, because it has the potential of *bias* or *skew*.

One of the main aspects of missingness is the degree of randomness in which missing data occur. Rubin [1976] developed a framework of inference about missing data and defined two main classes of missingness (also known as *mechanisms of missingness*): *Missing At Random* (MAR) and *Not Missing At Random* (NMAR). With MAR mechanism, the likelihood of missing data may depend on observed values but does not depend on missing values. With NMAR mechanism, the likelihood of missing data depends on missing values. Schafer and Graham [2002] provided a formal definition for MAR and NMAR mechanisms, following the probability theory to distinguish between the two mechanisms of missingness. The definitions are presented in Section 3.
For example, to estimate the average age for some population in a survey, where 90% of men reported their age whereas only 50% of women did. At an initial stage, it would appear that missingness depends on the attribute gender, because women may be more reluctant to report their age, regardless of the value. In this case, missingness falls in the MAR category and missing data would not affect the resulting average age. On the other hand, women are probably older than men. If so, it would appear that older people are less likely to report their age, and thus, missingness would depend on the values of the attribute age where missingness occurred. This time missingness falls in the NMAR category and the average age will be probably lowered (or underestimated) since high values of age are missing. Note that the opposite is valid if the small values instead of the high values of the attribute age were missing, that is, the average age will be increased (or overestimated). Therefore, any parameter estimation of the attribute age will yield bias (i.e., underestimate or overestimate) in the results. The latter example shows how NMAR data can cause bias in the parameter estimation, therefore, determining the mechanism of missingness is very important, because it can influence the choice of an appropriate treatment for missing data. In the rest of the article, we use only the term skew for the sake of uniformity.

Important operations over high-dimensional data include the *Similarity Search*. It involves finding objects that are similar to a given query object based on some similarity measure. The searching process is usually performed with range query or nearest neighbor query on an index structure that describes the distribution of the objects in the space. There are two families of access methods that provide indexing support for similarity search: *Multidimensional Access Methods* and *Metric Access Methods* (MAMs). Multidimensional access methods, such as R-tree and kd-tree, employ the attribute values of the data objects to build the index structure, by dividing the space recursively along the dimensions (i.e., attributes) into hyper-regions and projecting the objects into the proper hyper-region, based on their corresponding values for each dimension. Metric access methods do not support a direct concept of projection, that is, the objects are not projected along the different dimensions. Instead, a distance function is provided and the space is divided into regions using a set of chosen objects, called *representatives*, and their distances to the rest of the objects in the space, building a *metric space*.

There are a variety of proposals for multidimensional and metric access methods [Böhm et al. 2001] [Hjaltason and Samet 2003]. However, nearly all of them are designed to operate on complete databases. When applied to incomplete databases, objects with missing attribute values are discarded from the index structure because the indexing techniques fail to index such objects. This effect can considerably reduce the size of the database and, consequently, affect the query performance by retrieving non-relevant objects (or false hits) and missing the relevant ones (or false dismissals).

The most popular approach used to make the existing indexing techniques operational on incomplete databases is to replace the missing values with indicators that are not in the domain of the attributes. For example, if the domain of an attribute are the positive integers, then, a value of -1 can be used to denote the missing values. With this scheme and for a multidimensional indexing technique, all the objects with missing values along a dimension are projected into the same value (i.e., the value of the indicator) within the hyper-region, causing skew in the indexed space. Metric access methods are even more susceptible to skew despite the fact that there is no direct concept of projection. The reason is that metric indexes are built on the pairwise distances between the objects, and the indicators of missing values will be certainly involved in the estimation of the distances. But since the indicators are invalid values of the attributes, this can cause distortion in the distances between the objects of the metric space.

The goal of this article is to provide a framework to study the issues of indexing high-dimensional databases with missing attribute values and to evaluate the impact of the different mechanisms of missingness on the metric access methods. The main contributions include the following:

(1) Stating the important concepts related to missing data theory and providing a taxonomy of different mechanisms of missingness.
(2) Discussing the problems involved when indexing incomplete databases for multidimensional and metric access methods.
(3) Formalizing the problem of missing data in metric spaces and demonstrating theoretically that missing data not at random can cause skew in the metric space.
(4) Conducting a set of experiments to evaluate the performance of the metric access methods when applied to incomplete databases, and also, to examine the effect of each mechanism of missingness on the metric spaces, in order to reinforce our theoretical basis.

The experimental results show that data missing at random affect the effectiveness of the query performance, but data not missing at random are more related to the efficiency of the query performance. The reason is that, regarding MAR data, some objects in the clusters where missingness occurred are likely to move randomly to other clusters, changing completely the neighborhood and, consequently, causing false hits and false dismissals in query responses. On the other hand, regarding NMAR data, the objects of the clusters where missingness occurred move closer as the pairwise distances become smaller, because NMAR missingness either underestimates or overestimates the distances among objects of the cluster, depending on the process that causes missing data. For example, in our case study, the distances among objects are lowered due to distance shrinking, which caused skew in the metric space. However, skew favors the node overlap in the metric indexes and, thus, worsens the efficiency of the query performance.

The remainder of this article is organized as follows: Section 2 discusses the related work. Section 3 presents our theory of missing data in metric spaces. Section 4 discusses the experimental results and finally, Section 5 presents our conclusions.

2. RELATED WORK

Metric access methods are known to be more suitable to index sets of complex and high-dimensional data, such as images and time series. Such data usually require preprocessing to extract relevant features in the form of feature vectors, which are used in the place of the original data to perform similarity queries. In some cases, missing attribute values occur in the feature vectors during the feature extraction process. For example, when extracting the gray-level histogram from a set of satellite images, the brighter values of the histogram may be missing due to sensor saturation with very high energy levels. Such data cannot be indexed in a metric space because metric spaces are built on pairwise distances among the objects, and the distance between two objects with missing values is

unknown. Therefore, a special treatment for missing data is required to enable the indexing process. In other situations, missing attribute values occur in the original data during the collecting process but fade out in the feature vectors during the feature extraction procedure. For example, when recording time series, some objects may lack values on certain attributes, but after they are submitted to a transform, like Fourier and Wavelets, the missing values fade out in the resulting transforms but turn into energy loss for the whole time series. This fact enables metric access methods to index objects with missing values. Nevertheless, any potential of skew held by the missingness in the original data remains in the feature vectors and, consequently, affects the query performance of the metric access methods.

One way to address the problem of missing values is to remove the objects with missing values and operate only on the complete objects. Another way consists of replacing the missing data with valid values in the domain. The latter, called *imputation*, is popularly used in statistics and resulted in several methods, such as Mean Substitution and Multiple Imputation [Allison 2001] [Graham 2009]. However, missing data not at random are difficult to predict as they tend to produce skewed results. If we take the example that estimates the average age of the population when missingness is not at random, we can notice that without additional knowledge about the fact that older people are reluctant to report their age, which is generally very difficult to obtain, it would be difficult to distinguish between the MAR and the NMAR cases. Therefore, a simple method for data prediction, like mean substitution (that is, replacing the missing values by the mean of the observed values), will produce a large gap between the real values and the predicted ones.
On the other hand, when data are missing legitimately, any data prediction would be inappropriate. In addition, some applications do not require data prediction, but a special treatment is necessary to allow the applications to operate on the available data properly.

The basic approach to extend multidimensional access methods to support queries over incomplete datasets is to replace the missing values with an indicator to denote them, then index the resulting dataset with a conventional technique, like the R-tree. The problem with this approach is that all the objects with missing values on a given dimension will be projected into the same value of the hyper-region. And if the proportion of missing values is large, it will result in a highly skewed data space. In addition, in order to perform a query that involves k attributes, it is necessary to search for all the objects that match with the query object on their observed values, including objects with missing values. For example, given a query that involves 3 attributes $(1, 2, 3)$, then all the following objects $(1, 2, 3), (?, 2, 3), (1, ?, 3), (1, 2, ?), (?, ?, 3), (?, 2, ?), (1, ?, ?), (?, ?, ?)$ are matching objects with the query. This is achieved by processing a number of sub-queries that include the query object with all the combinations of missing and observed values in the attributes. Formally, the number of sub-queries is given by 2^k. This strategy is very simple and easy for implementation. Its main drawback is that the search space grows exponentially with the growing dimension of the data, causing poor query performance, as illustrated in the experiments by Canahuate et al. [2006].

To address the problem of skew caused by data projection into a single indicator value, Ooi et al. [1998] defined a function that replaces the missing values with distinguished indicator values and randomly scatters the objects within the dimensions where missing data occurred. The benefit of this approach lies in its ability to reduce the skew in the projected space. However, it suffers from a poor query performance since it still requires performing an exponential number of sub-queries to find all the matching objects to the query.

Still based on the indicators of missing data, Canahuate et al. [2006] proposed an extended variation of Bitmaps and VA-Files, capable to operate on incomplete datasets. When Bitmaps are employed to index incomplete datasets, each attribute acquires an extra bitmap to denote missing values of the data objects on that attribute. Alternatively, the VA-Files employ bit-strings to encode the attribute values and an extra string of 0's to denote missing values. Query processing is modeled on a basis that enables to decide if a data object with missing values belongs to the query response or not. The

benefit of these techniques is that each dimension is indexed independently and searched separately without needing to transform the query to an exponential number of sub-queries. However, Bitmaps and VA-Files are not appropriate for large and complex datasets, and are useful only for a small class of datasets where the domain of the attributes is relatively small. In fact, the size of a bitmap index grows quickly with the growing domain of the attributes, as each possible value of an attribute requires a proper bitmap.

Aggarwal and Parthasarathy [2001] used the conceptual reconstruction concept that explores the correlation aspect of the attributes to reconstruct incomplete datasets, then the resulting complete datasets can be safely used for data mining purposes. The technique is very useful for an efficient data prediction, but only when the data are missing at random. Cheng et al. [2014] investigated the problem of missing data in similarity queries where the dimensions that have missing values are unknown. The authors developed a probabilistic framework to model the uncertainty caused by the dimension incompleteness. Yet, this approach requires imputation of the missing values and it does not tackle the case where data are not missing at random.

In the present study, we conduct a set of experiments to evaluate the impact of missing data on metric spaces. Without any attempt to predict the missing values, we consider data where missingness occurred in the original data in order to index the objects with missing values and then we analyze the effect of MAR and NMAR data on the query performance for the metric access methods. To the best of our knowledge, there is no attempt to discuss and address the problem of missing data in metric access methods. Thus, this article is the first to address this problem and to present solutions for two main goals: improve the understanding of the issues involved with metric spaces when indexing databases with missing values, and provide a theoretical basis to support the development of an effective and efficient solution for missing data that can be adapted to extend the range of metric access methods.

3. A THEORY OF MISSING DATA IN METRIC SPACES

In this section, we provide a formal definition of missing data and we describe the theory of missingness in metric spaces. Without loss of generality, we employ the Euclidean distance as a distance function.

Let $\mathbb{S}$ be a dataset of cardinality n and dimensionality m, generated by a random variable $X = (x_1, x_2, \ldots, x_m)$, with a probability density function f. The dataset $\mathbb{S}$ is incomplete if there exist some attributes that have missing values. We denote the original data by $X_{org} = (X_{obs}, X_{miss})$, where X_{obs} are the fully observed attributes and X_{miss} are the attributes with missing values.

Let R be an indicator variable that identifies missing values. R is defined with a set of random variables with a join probability distribution g. In statistical literature, the distribution of R corresponds to the mechanism of missingness, discussed in the literature.

Definition 1. **(Missing At Random - MAR)** *Missing data are said to be MAR if the distribution of missingness depends on X_{obs} but does not depend on X_{miss}:*

$$g(R|X_{obs}, X_{miss}) = g(R|X_{obs}) \tag{1}$$

Definition 2. **(Not Missing At Random - NMAR)** *Missing data are said to be NMAR if the distribution of missingness depends on X_{miss}:*

$$g(R|X_{obs}, X_{miss}) = g(R|X_{miss}) \tag{2}$$

Definition 3. **(Bias of the Parameter Estimation)** *Bias, also called skew, of a parameter estimation α, is the gap between the expected estimation of α and its real estimation. It occurs when α is underestimated or overestimated with a certain value ϵ.*

Definition 4. **(Metric Space)** *A Metric Space is defined as a pair* $(\mathbb{S}, d)$*, where* $\mathbb{S}$ *denotes a domain of valid data objects and* $d : \mathbb{S} \times \mathbb{S} \rightarrow \mathbb{R}^+$ *is a metric that complies to the following metric axioms,* $\forall x, y, z \in \mathbb{S}$*. Reflexivity:* $d(x, x) = 0$*; Non-negativity:* $d(x, y) > 0$ *if* $x \neq y$*; Symmetry:* $d(x, y) = d(y, x)$*; and Triangle inequality:* $d(x, y) + d(y, z) \geq d(x, z)$*.*

As far as known by the authors, no model exists for missing data represented in metric spaces. Thus, we propose the following lemma.

Lemma 1. *For a metric data associated with the Euclidean distance, missingness exhibits the same phenomenon described in the statistical literature. That is, the MAR mechanism does not introduce skew in the metric space for a certain amount of missing data, whereas the NMAR mechanism causes skew in the metric space.*

Under this assumption, we search for evidence of skew in the distribution of data in the metric space. For this purpose, we analyze the distance parameter among the objects, because it determines the data distribution in the metric space.

To provide evidence that Lemma 1 holds in real data, we show that NMAR mechanism has a direct effect of skew on the metric space. Thereafter, we report interesting practical evaluations of missing data responsible for skew in the experimental study to support our theoretical assumption. Suppose we have a metric space $(\mathbb{S}, d_E)$ and an incomplete dataset $S \subset \mathbb{S}$ associated to the Euclidean metric d_E. We denote the existing dataset by X_{org}, the observed attributes by X_{obs}, and the attributes with missing values by X_{miss}. Because a metric space is organized around representative objects and their distances to the rest of the objects in the space, analysis of the pairwise distances is fundamental to reveal important features about the data distribution in the space. So, let p be a representative object taken randomly from the metric space. The following equation represents the average of the squared distances between p and the rest of the objects in the space:

$$\frac{1}{n}\sum_{i=0}^{n-1} d_E^2(\vec{p}, \vec{x_i}) = \frac{1}{n}\sum_{i=0}^{n-1}\sum_{j=0}^{m-1} (p[j] - x_i[j])^2 \tag{3}$$

where $n > 0$ and $m > 0$, $n, m \in \mathbb{Z}$ are the cardinality and dimensionality of the dataset, respectively. Note that we take the squared distances to remove the root in the Euclidean distance in order to facilitate further developments without affecting the results. After developing the right hand side of Equation 3 we obtain:

$$\frac{1}{n}\sum_{i=0}^{n-1} d_E^2(\vec{p}, \vec{x_i}) = \sum_{j=0}^{m-1} (p[j])^2 - \frac{2}{n}\sum_{i=0}^{n-1}\sum_{j=0}^{m-1} (p[j] \times x_i[j]) + \underbrace{\sum_{j=0}^{m-1} (\frac{1}{n}\sum_{i=0}^{n-1} (x_i[j])^2)} \tag{4}$$

The under-bracketed part of Equation 4 represents the summed averages of the attributes with squared values. In other terms, it is the sum of parameters α that estimate the average of every attribute (actually, squared attribute) in (X_{obs}, X_{miss}), where $\alpha = \frac{1}{n}\sum_{i=0}^{n-1} x_i^2$. If we assume that NMAR data occurred in at least one of the attributes of X_{miss}, then the corresponding parameter α will be skewed (i.e., underestimated or overestimated) and, subsequently, the overall equation that involves α will also be skewed.

The importance of this result lies in the fact that Equation 4 depends on a parameter α that can be subject to skew because it involves components of X_{miss}. Rubin [1976] claims that MAR mechanism does not yield to skew in parameter estimations and that NMAR mechanism does. So, if we assume that X_{miss} are MAR then α is not skewed and consequently the overall outcome of Equation 4 will not be skewed. Contrarily, if X_{miss} are assumed to be NMAR then α is skewed and the results of Equation 4 will be skewed. However, skew in Equation 4 implies skew in the distance distribution of

the data in the metric space. Notice that Equation 3 makes no assumption over the distance function d_E, so it is valid for any metric space. An important observation is that although the MAR mechanism does not cause skew in the metric space, it does not mean that we can obtain the same results as if there were no missing data.

4. PERFORMANCE STUDY

In this section we perform a controlled set of experiments to evaluate the effect of missingness on the distance distribution of time series datasets, evaluating both MAR and NMAR mechanisms. The experiments are conducted to answer the following questions:

(1) Does missingness have the effect of skew on metric spaces as described in the statistics for vector spaces?
(2) If yes, what are the forms of skew in metric spaces?
(3) Do the distinct mechanisms of missingness affect the performance of a MAM differently?
(4) How sensitive are the MAMs to missing data in terms of effectiveness and efficiency?

4.1 Experimental Set-up

4.1.1 *Data Description.* To conduct the experiments we use two datasets, one real and one synthetic. The real dataset consists of 500,000 NDVI[1] time series extracted from satellite images corresponding to the state of São Paulo. The time series possess 108 measurements monthly recorded, ranging from April 2001 to March 2010.
The synthetic dataset contains 10,000 time series of weather forecast obtained from the AgroDataMine Server of the GBdI group[2]. The data represent a simulation of three climate variables corresponding to pressure, minimum temperature, and maximum temperature, in a range of latitudes/longitudes covering the area of Brazil and standing from January 2012 to July 2015. This gives a total of 128 measurements for each time series.

4.1.2 *Data Preparation.* The datasets described in Section 4.1.1 do not have missing values. We generate the missing data in a controlled way to produce data with MAR mechanism and data with NMAR mechanism, each data with different amounts of missing values. We employ two procedures to produce datasets with missing attribute values based on the available data.

Procedure for MAR data

Given a complete m-dimensional dataset S, we generated the incomplete datasets by withdrawing some values from a set of specific attributes [Ooi et al. 1998]. We allow up to k attributes to have missing values, where $k < m$. Therefore, any data element in S can have at least 0 and at most k missing values. Equation 5 shows the proportion of data objects $F(i)$ that will have i missing values, $i = 0, 1, \ldots, k$. $F(i)$ is defined as follows:

$$F(i) = \left(\frac{k-i+1}{k+1}\right)^2 - \left(\frac{k-i}{k+1}\right)^2 \tag{5}$$

Table I shows an example of the distribution of missing data when $k = 5$; where k is the number of attributes allowed to have missing values. Notice that $F(i)$ is inversely proportional to i, that is, a larger number of data objects will have a smaller number of missing attribute values and vice versa. Moreover, $F(i)$ should satisfy the property $\sum_{i=0}^{k} F(i) = 1$. After the proportions of missing data

[1]Normalized Difference Vegetation Index (NDVI) indicates the soil vegetative vigor represented in the pixels of the images and it is strongly correlated with biomass.
[2]*http : //www.gbdi.icmc.usp.br/*

Table I. Distribution of missing data for $k = 5$.

Number of missing values i	Proportion of data $F(i)$
0	11/36
1	9/36
2	7/36
3	5/36
4	3/36
5	1/36

Table II. Incomplete datasets with the number of attributes having missing values and the percentage of missing data.

Name		N° attributes	% missing data
MAR	ndvi-2%	7	2.025
	ndvi-5%	17	5.1
	ndvi-10%	33	10.03
	ndvi-15%	49	14.97
	ndvi-20%	65	19.91
	ndvi-25%	82	25.15
	WeathFor-2%	8	1.97
	WeathFor-5%	20	5.08
	WeathFor-10%	39	10
	WeathFor-15%	58	14.91
	WeathFor-20%	78	20.05
	WeathFor-25%	97	24.90

Name		% missing data
NMAR	ndvi-20%	19.77
	WeathFor-2%	2.55
	WeathFor-5%	4.96
	WeathFor-10%	10.23
	WeathFor-12%	12.15
	WeathFor-16%	15.78
	WeathFor-18%	18

are established, we proceed to the selection of the k attributes that will have missing values. This step depends on the application and it is very important because it determines the distribution of missing data that will secure the MAR mechanism. Useful aspects that can be employed to choose the attributes that will have missing values for time series data are the time component and the correlation among the attributes. For example, considering a collection of daily temperature time series collected over a period, intuitively, the missing values are more likely to occur in a continuous sequence of time points. The reason is that if a sensor fails randomly to record a data at time point t, it is more likely to fail repeatedly at time points $t + 1$, $t + 2$, etc, as long as the failure continues, until the sensor is fixed.

Procedure for NMAR data

We recall that the NMAR mechanism holds when the likelihood of missing data depends on the missing values. This happens, for instance, when a sensor fails to record certain values because they are very high or very low. Based on this example, we set an upper and a lower threshold for the attributes, and we omit all the attribute values that are above the upper threshold or bellow the lower threshold. Note that if the attribute domains are different, then we set appropriate thresholds for each attribute.

Table II shows all the incomplete time series datasets created for the experiments. *ndvi-x%* is the real dataset followed by the percentage of missing values, and *WeathFor-x%* is the synthetic dataset followed by the percentage of missing values.

4.1.3 *Methodology.* The experiments are conducted through the following steps:

(1) ***Data preprocessing:*** Before indexing the time series, it is required to use a transformation method that maps them to a lower dimensional feature space. We propose to use the Haar Wavelet Transform as a feature extraction method. Also, we implemented the classical Pyramid Algorithm [Mallat 1989] based on convolutions with filters for Discrete Wavelet Transforms (DWT) to calculate the wavelet coefficients.

(2) ***Indexing:*** For each set of transformed time series obtained at the preprocessing step, we kept the first 20 coefficients to represent each time series and we employed the Euclidean metric to build each index. We used the Slim-tree [Traina et al. 2000] metric access method, available in the Arboretum framework[3], to conduct the experiments.

(3) ***Performing similarity queries:*** For each dataset we selected 500 elements as query centers of which we process a k-nearest-neighbor query (k-NN$_q$) and a range query (r$_q$). The number k of nearest neighbors was set to 50 for the NDVI datasets (0.01% the size of the original dataset), and it was set to 10 for the WeathFor datasets (0.1% the size of the original dataset). The covering radius of the range queries was set to the distance between the query center q and its k^{th} nearest neighbor. To evaluate the performance of the access method when applied on incomplete datasets, we evaluate the precision and recall. Precision denotes the ratio of the number of relevant objects retrieved to the total number of irrelevant and relevant objects retrieved. Recall denotes the ratio of the number of relevant objects retrieved to the total number of relevant objects available. Formally,

$$Precision = \frac{|\text{Relevant objects retrieved}|}{|\text{Total objects retrieved}|}, Recall = \frac{|\text{Relevant objects retrieved}|}{|\text{Total objects relevant}|} \quad (6)$$

We use the query results on the complete datasets as a reference to estimate the effectiveness of the query results on the incomplete datasets. During the experiments, we generated the graphs of precision and recall, and we measured the average number of disk accesses, the average number of distance calculations and the total time required to process all the queries for every dataset. The tests were performed on a machine with a Pentium D 3.4GHz processor and 2Gb of memory RAM.

4.2 Experimental results

Figures (1,2), (3,4) and (5,6) show the query performance over MAR and NMAR NDVI datasets, MAR WeathFor, and NMAR WeathFor datasets, respectively, with different portions of missing data. Notice that in the graphs of the NDVI datasets, the last measurement with 20% of missing values corresponds to the NMAR mechanism and the others to the MAR mechanism.

Figures 1, 3 and 5 show clearly how the effectiveness of the search process over incomplete datasets degenerates with the growing amount of missing values. Under the MAR mechanism, with only 2% of missing data, k-NN$_q$ queries exhibit a fast drop for both precision and recall (see fig 1a and fig 3a), whereas, range queries show a higher precision rate (see fig 1b and fig 3b). Under the NMAR mechanism and up to 10% of missing data, both the k-NN$_q$ and range queries show a slow to moderate decrease of precision and recall rates (see fig 5a and fig 5b) with a higher precision rate achieved by range queries (see fig 5b). Note that, a worsening in precision and recall measurements implies that occurred a significant amount of false hits (i.e., presence of non-qualifying objects) and false dismissals (i.e., missing qualifying objects), respectively. We can also observe that, for the k-NN$_q$ queries, the measurements of precision are identical to the measurements of recall. This occurred because the total number of relevant objects is equal to the total number of retrieved objects, which is set to k (see Equation 6). This is not the case for the range queries because the number of retrieved objects depends on the covering radius.

Under the MAR mechanism, the precision of the range queries is better than that of the k-NN$_q$ queries because, probably, there are less false hits in the query responses. However, it does not mean that range queries outperform k-NN$_q$ queries, because the associated recall that reflects the completeness of the query answer is relatively poor. In fact, unlike false hits, missing qualifying objects are difficult to recover once they occurred, and thus, recall turns out to be more problematic than precision.

[3] $http://www.gbdi.icmc.usp.br/old/arboretum$

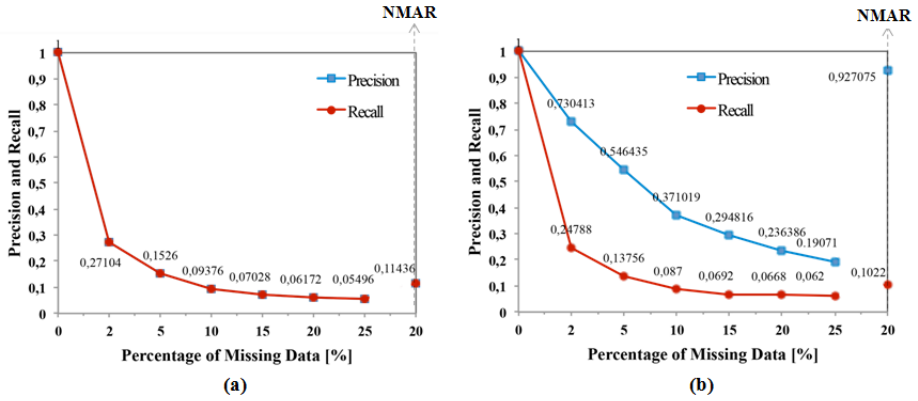

Fig. 1. The query performance (given by a precision and recall) for (a) k-nearest neighbor query, (b) Range query, where the plots refer to the NDVI datasets.

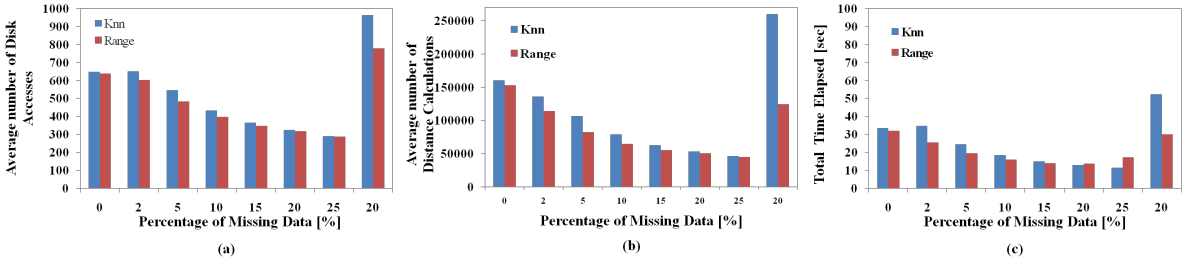

Fig. 2. The query performance (given by an average number of disk accesses, average number of distance calculations and a total time elapsed (Sec)) for the MAR and NMAR NDVI datasets.

The results obtained so far were the expected ones because, intuitively, the more relevant attributes are used to represent the objects, the higher is the ability to identify similar objects and more accurate the query answer. But, when attribute values are missing, the distances between objects tend to be uncertain as there is no warranty that close objects are close in the complete data space. The experimental results show that when data are MAR, some objects in the clusters where missingness occurred are likely to move randomly to other clusters changing completely the neighborhood and, subsequently, causing false hits and false dismissals in the query responses, which can explain the poor quality of the query performance. However, when missing data are NMAR, the objects of the clusters where missingness occurred are more likely to remain in the same clusters. Although, moving closer to each other as the pairwise distances become smaller, because NMAR missingness can either underestimate or overestimate the distances among objects of the cluster, which can justify the better quality of the corresponding query.

So far, we have seen that MAR data are able to cause a significant change in data distribution of the metric space even with only 2% of missing data, and NMAR data tend to follow the same distribution as the complete data space with a sizable portion of missing data (up to 10% of missing data, k-NN$_q$ and range queries achieve more than 50% of precision and recall).

Now, if we observe the efficiency parameters (see fig 2, 4 and 6), we can notice that for MAR datasets, the average number of disk accesses along with the average number of distance calculations and the total time required decrease with increasing amount of missing data, while NMAR datasets require higher values for all the parameters, in particular for the k-NN$_q$ queries.
Actually, the degradation in NMAR data is due to the skew of the metric data space. In fact, when missing data are NMAR, neighboring objects in the complete data space are more likely to preserve their positions in the clusters where missingness occurred, while moving in the direction that

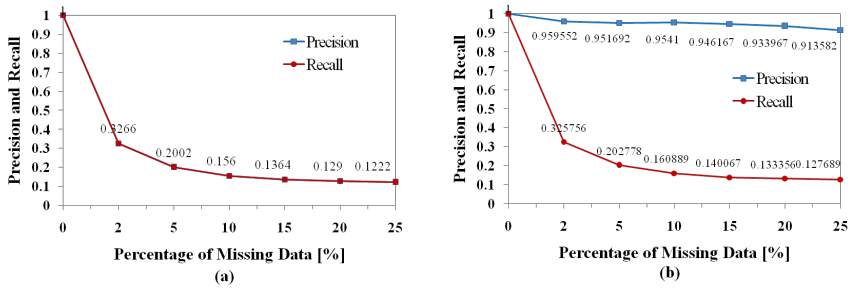

Fig. 3. The query performance (given by a precision and recall) for (a) k-nearest neighbor query, (b) Range query, where the plots refer to the MAR WeathFor datasets.

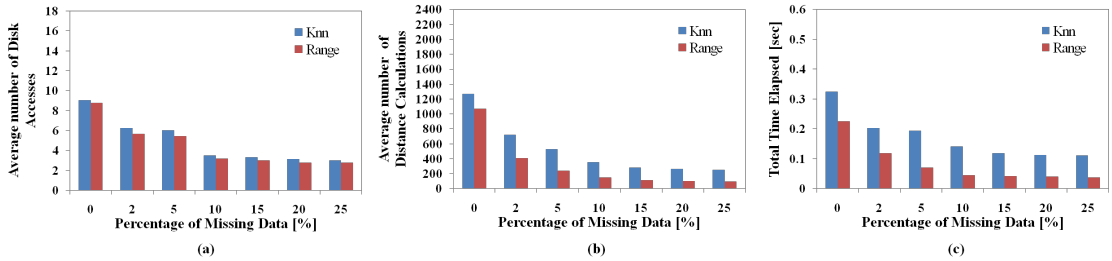

Fig. 4. The query performance (given by an average number of disk accesses, average number of distance calculations and a total time elapsed (Sec)) for the MAR WeathFor datasets.

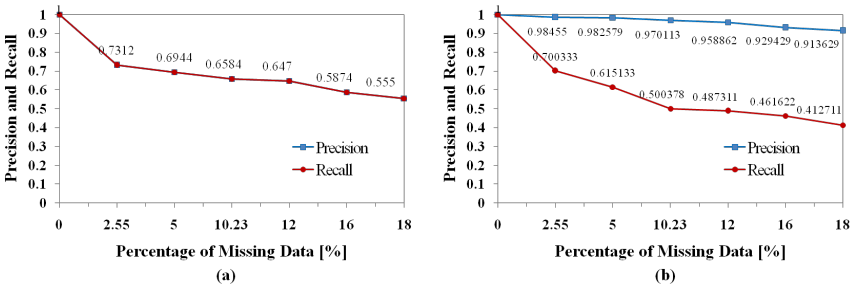

Fig. 5. The query performance (given by a precision and recall) for (a) k-nearest neighbor query, (b) Range query, where the plots refer to the NMAR WeathFor datasets.

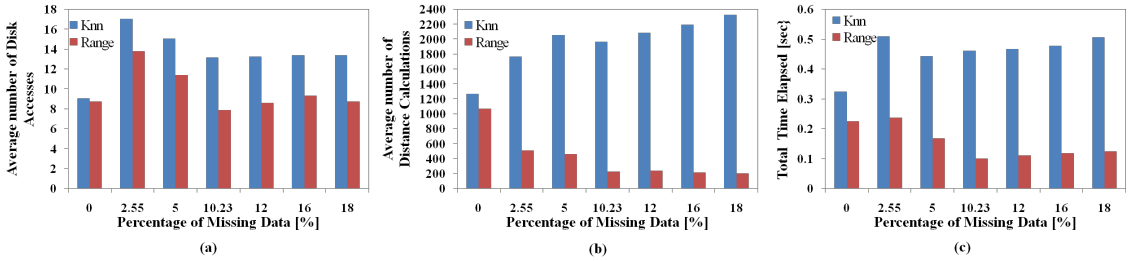

Fig. 6. The query performance (given by an average number of disk accesses, average number of distance calculations and a total time elapsed (Sec)) for the NMAR WeathFor datasets.

makes close objects become closer, because the pairwise distances are skewed (in this case and more precisely the distances are underestimated). This condition is similar to the *distance concentration* (also well known as *dimensionality curse*), where distance differences among pairs of objects become insignificant. So, when neighboring objects move closer, the nearest neighbor becomes very close and a small increase in the query radius will include several other objects. Recall that distance concentration phenomenon favors the node overlap in the metric indexes and makes the search process more expensive in terms of time search, number of disk accesses and number of distance calculations. Contrariwise, when processing the MAR data, the efficiency improves with the increasing amount of missing data. Theoretically, the results are unexpected because if MAR data do not cause skew in the metric space we should obtain the same query performance as from complete data. One possible motive is just that missing objects reduces the size of the dataset and that MAR data do not cause skew in the metric space.

To better understand the latter results, we plot the probability density function (i.e., concentration) of the pairwise distances among the objects to analyze their behavior with regard to missingness mechanisms in both data and transform domains (see fig 7 and fig 8). In the data domain, the graph of the distance concentration has almost the same shape, height, and width for all MAR datasets with different portions of missing data. However, for NMAR data, the graph becomes higher and narrower as missing data increases, that is, the distance concentration becomes increasingly important. This result shows that the problem of distance concentration that causes skew in the data space affects NMAR data but not MAR data.
In the transform domain, the plot of the distance concentration becomes gradually wider and low-lying for both MAR and NMAR datasets. Interestingly, the plot for MAR data tends to preserve its shape over a sizable amount of missing data (up to 10%), whereas the plot of NMAR data worsens quickly with the increasing amount of missing data, as we can observe a development of many peaks that become increasingly numerous and homogeneous, spanning over a wider range of distance values. These peaks represent the distance concentrations caused by the NMAR process.

When data are missing at random, the objects with missing attribute values move randomly in the space causing change in the distribution of the metric space. However, when the distance concentration does not already exist in the complete metric space, the random movement of the objects in the incomplete data space does not introduce concentration of the distances that can potentially cause skew in the space, even at high amounts of missing data (at least 10%). On the other hand, when data are missing not at random, the objects with missing attribute values do not move randomly in the space. Instead, close objects become closer and distant objects move further, dividing the data space into several subspaces. With the increasing amount of missing data, the distance differences between pairs of objects of each subspace become insignificant, favoring the phenomenon of concentration. However, more distance concentration yields a highly skewed data space.

5. CONCLUSION

In this article, we analyzed the problem of missing data when indexing incomplete high-dimensional databases. First, we presented the fundamental concepts related to missing data and we provided a classification of the mechanisms that lead to missingness. Thereafter, we formalized the problem of missingness in metric access methods and, based on the theoretical framework established by the statisticians, we demonstrated that missing data have a direct effect on the metric spaces and we, experimentally, evaluated the kind and amount of distortion that is imposed on the metric indexing methods. We conducted a set of experiments to evaluate the performance of a metric access method when applied to incomplete datasets and analyzed the role of each missingness mechanism in skewing the metric space. The outline of our results are the following:

—for MAR data, the parameter performance of the search process related to precision and recall degrade significantly with the growing amount of missing data. However, NMAR data show a

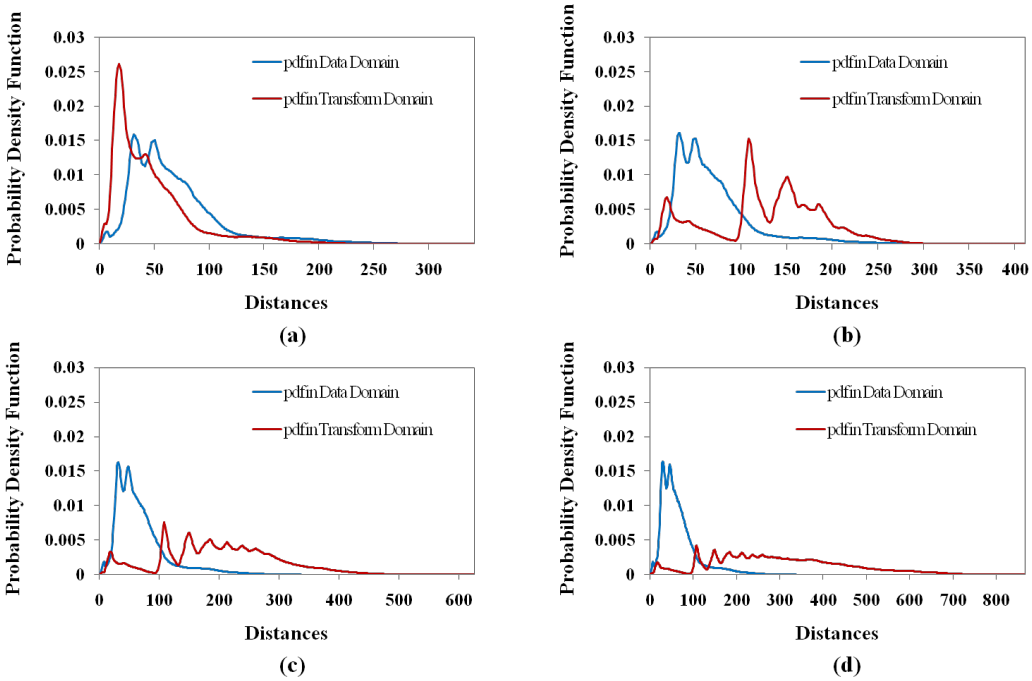

Fig. 7. Probability Density Function of the pairwise distances for MAR data: (a) Complete dataset. (b) 2% missing data. (c) 5% missing data. (d) 10% missing data.

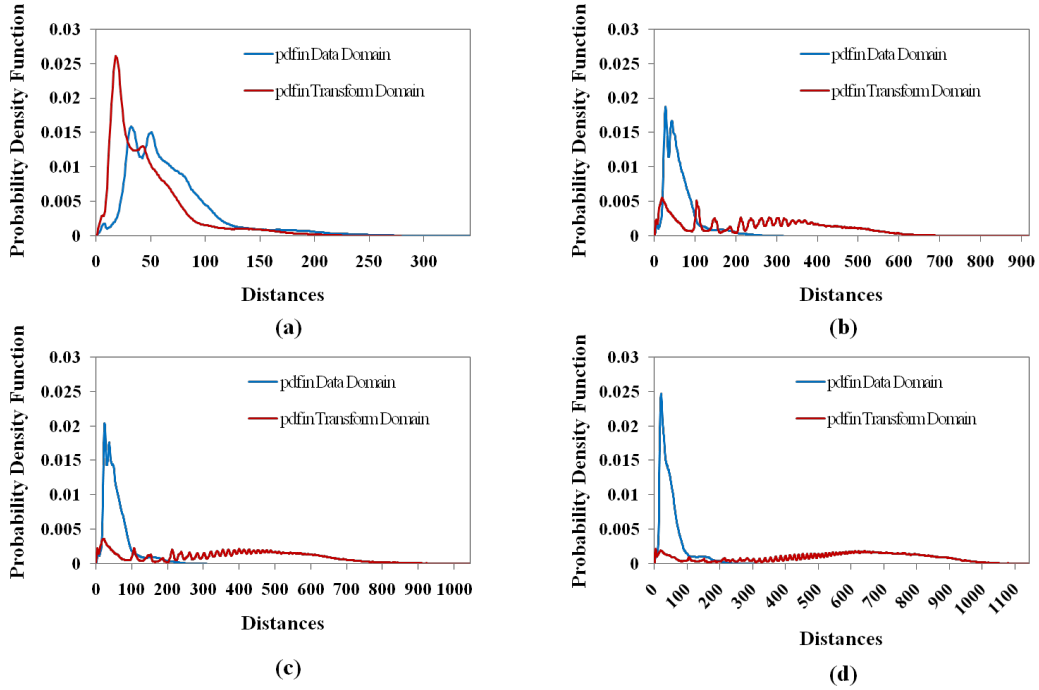

Fig. 8. Probability Density Function of the pairwise distances for NMAR data: (a) Complete dataset. (b) 2.55% missing data. (c) 5% missing data. (d) 10.23% missing data.

better performance;

—for NMAR data, the search process evaluated measuring execution time, number of disk accesses and number of distance calculations degenerates quickly with the growing amount of missing data, whereas, MAR data do not show much difference;

—unlike MAR data, NMAR data are susceptible to distance concentration in both data and transform domains.

—although, missing values do not appear in the feature vectors, any potential for skew held by the missingness in the original data is preserved in the feature vectors and, consequently, affects the query performance of the metric access method.

Degeneration of the precision and recall is not an indicator of skew in the metric space, because it is natural to obtain different results when queries are performed on different datasets. However, degeneration of the efficiency measurements (i.e., time search, number of disk accesses and number of distance calculations) is not natural unless there is a reason that leads to the degeneration of the metric access method. We identified that the reason is the effect of distance concentration caused by NMAR processes. In fact, distance concentration increases the node overlap of the MAM and degenerates the performance of the search process.

Thus, we claim that missing data at random do not cause skew in the metric space, not even with 25% of missing values, but data missing not at random can cause severe skew with only 2% of missing values. This result is useful to guide the development of two types of solutions for handling missing data: first, a solution to make the metric access methods able to index incomplete datasets when missing attribute values occur in the feature vectors, without needing to predict the missing values, in order to perform the similarity queries on the available data. Second, a solution to reduce the skew caused by NMAR missingness when missing attribute values occur in the original data, in order to speed up the search process and improve the query performance. We highlight that our results are the first to allow the development of indexing techniques that take into account the effects caused by missing data to really allow bypassing them in the index structure.

REFERENCES

Aggarwal, C. and Parthasarathy, S. Mining Massively Incomplete Data Sets by Conceptual Reconstruction. In *Proceedings of the 7th International Conference on Knowledge Discovery and Data Mining*. a, New York, USA, pp. 227–232, 2001.

Allison, P. D. *Missing data*. SAGE Publications, Inc, London, Uk, 2001.

Böhm, C., Berchtold, S., and Keim, D. Searching in high-dimensional spaces: Index structures for improving the performance of multimedia databases. *ACM Computing Surveys* vol. 33, pp. 322–373, 2001.

Canahuate, G., Gibas, M., and Ferhatosmanoglu, H. Indexing Incomplete Databases. In *Proceedings of the 10th international conference on Advances in Database Technology*. a, Munich, Germany, pp. 884–901, 2006.

Cheng, W., Jin, X., Sun, J.-T., Lin, X., Zhang, X., and Wang, W. Searching Dimension Incomplete Databases. *IEEE Transactions on Knowledge and Data Engineering* vol. 26, pp. 725–738, 2014.

Graham, J. W. Missing data analysis: making it work in the real world. *Annual Review of Psychology* 60 (1): 549–576, 2009.

Hjaltason, G. R. and Samet, H. Index-driven similarity search in metric spaces. *ACM Transactions on Database Systems* vol. 28, pp. 517–580, 2003.

Mallat, S. G. A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. *IEEE Trans. Pattern Anal. Mach. Intell.* vol. 11, pp. 674–693, 1989.

Ooi, B. C., Goh, C. H., and Tan, K.-L. Fast high-dimensional data search in incomplete databases. In *Proceedings of the 24rd International Conference on Very Large Data Bases*. a, New York, USA, pp. 357–367, 1998.

Rubin, D. B. Inference and missing data. *Biometrika* 63 (3): 581–592, 1976.

Schafer, J. L. and Graham, J. W. Missing data: Our view of the state of the art. *Psychological Methods* vol. 7, pp. 147–177, 2002.

Traina, C., Traina, A. J. M., Seeger, B., and Faloutsos, C. Slim-Trees: High Performance Metric Trees Minimizing Overlap Between Nodes. In *Proceedings of the 7th International Conference on Extending Database Technology: Advances in Database Technology*. a, London, UK, pp. 51–65, 2000.