Data Clutter Reduction in Sampling Technique

Nur Nina Manarina Jamalludin[1], Zainura Idrus[2], Zanariah Idrus[3], Ahmad Afif Ahmarofi[4], Jahaya Abdul Hamid[5], Nurul Husna Mahadzir[6]

College of Computing, Informatics and Media, Universiti Teknologi Mara (UiTM), Selangor, Malaysia[1, 2]
Faculty of Computer & Mathematical Sciences, Universiti Teknologi MARA Kedah, Kedah, Malaysia.[3, 4, .6]
Kolej Matrikulasi Kedah,Kementerian Pendidikan Malaysia, Changlun, Kedah.[5]

Abstract*—Visualization is a process of converting data into its visual form as such data patterns can be extracted from the data. Data patterns are knowledge hidden behind the data. However, when data is big, it tends to overlap and clutter on visualization which distorts the data patterns. Data is overly crowded on visualization thus, it has become a challenge to extract knowledge patterns. Besides, big data is costly to visualize because it requires expensive hardware facilities due to its size. Moreover, it is timely to plot the data since it takes time for data to render on visualizations. Due to those reasons, there is a need to reduce the size of big datasets and at the same time maintain the data patterns. There are many methods of data reduction, which are preprocessing operations, dimension reduction, compression, network theory, redundancy elimination, data mining, machine learning, data filtering and sampling techniques. However, the commonly used data reduction technique is sampling technique that derives samples from data populations. Thus, sampling technique is chosen as a study for data reduction in this paper. However, the studies are scattered and are not discussed in a single paper. Consequently, the objective of this paper is to collect them in a single paper for further analysis in order to understand them in great detail. To achieve the objective, three interdisciplinary databases which are ACM Digital Library, IEEE Explore and Science Direct have been selected. From the database, a total of 48 studies have been extracted and they are from the years 2017 to 2021. Other than sampling techniques, this paper also seeks information on big data, data visualization, data clutter, and data reduction.

Keywords*—*Sampling technique; probability sampling; non-probability sampling; data clutter; big data; data visualization; data reduction

I. Introduction

Data visualization is a technique to convert data to a visual form to extract knowledge hidden behind the data through data patterns. According to [1], data visualization involves a combination of people with distinct visualization-related skills. Data visualization is also a technology to explore data interactively. Through data exploration, various data patterns can be revealed.

Big data plays a bigger role in our latest technologies today. Communities are particularly depending on the data to gain more information for decision making [2]. The advantage of data visualization is that it supports analysis, identifies issues and tackles problems faster through data patterns [3].

Big data technology is designed to process an enormous dataset for process optimization and decision making [4, 5]. However, an enormous dataset, both structured and unstructured are complex as they deal with an extensive amount of data. Thus, consistently, they are inadequate to operate with conventional processing techniques and algorithms [2, 6]. This is true when data is from various sources with various forms and format and yet they need to be integrated prior to processing.

Other challenges when dealing with big data are the effectiveness and efficiency in understanding, storing, managing, and developing data visualization [7]. Plotting these big data to form visualizations, require high end and expensive hardware and software facilities.

Nevertheless, when data have been successfully plotted and converted to visualization forms, it is common for the data to overlap on top of each other which could lead to data clutter issues. Data clutter can be defined as data that overlapping on top of each other which can lead to a massive number of false detections over the search space that relies on pixel patterns [8]. Fig. 1 below show the example of data clutter.

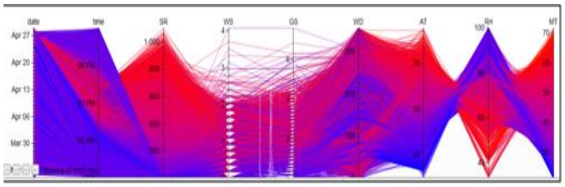

Fig. 1. Example of data clutter (Source: [51])

Another issue of concern regarding data clutter is plotting efficiency. It takes a longer time to plot data into its visual and is defined as computational overhead which is costly [9].

Data clutter also leads to unrecognized data patterns when in fact extracting the patterns is the main objective of data visualization [10] for strategic planning and decision making. In other words, the whole point of data visualization is to extract data patterns in order to uncover the gems of knowledge hidden behind the data.

Thus, there is a need to overcome data clutter and one of the techniques is through data reduction. Data reduction is one of the methods to shrunk computational overhead [11]. Although the dataset is reduced, the original information in the dataset should be preserved without scarifying any data patterns. However, data reduction could somehow remove some information from an original dataset that can lead to an unknown output of the dataset [12]. Nevertheless, data reduction can solve the difficulties that both data and visualization scientists suffer [13].

By reducing data in a dataset, data visualization can be more comprehensible with clearer data patterns. With proper data processing, data reduction can generate accurate visualization without changing data patterns [14].

The sampling technique is about choosing a subgroup from a large dataset and at the same time maintaining its properties or attributes [15, 16]. The sampling technique is a crucial way to analyze a massive dataset where its size is reduced for effective use of equipment and space [17].

Moreover, sampling techniques can lower data error that happens due to human factors when dealing with a large dataset [18]. The sampling technique is an outstanding technique to handle large datasets and when the resources are restricted. It generates results rapidly and accurately as data are smaller in size [19]. Normally, data reduction is implemented along with a machine-learning technique to analyze the size of the dataset to find an accurate output [14].

During analysis, it is common to visualize and analyze smaller datasets at a time as it is easier to identify data patterns [7, 20]. On the other hand, a larger dataset is explored at an early stage of data analysis to get a bird eye view and to identify interesting patterns for further analysis [13, 21]. Besides that, a larger dataset is used to understand the structure and the flow of the dataset [13]. However, it is burdensome for the researchers and the algorithms if a large dataset is used to discover data patterns.

Various techniques have been used for data reduction such as pre-processing operations, dimension reduction, compression, network theory, redundancy elimination, data mining, machine learning, data filtering and sampling technique However, in this study, the sampling technique is chosen as a data reduction technique. It is chosen because it is a well-known technique and consistently gives a positive result in scaling down the number of massive datasets [11].

However, there are various types of sampling techniques. Each is with its own strength and weakness. However, the discussion of these sampling techniques is scattered and not in a single paper. It is inconvenient to gather information from different papers. Thus, this review paper combines various sampling techniques into a single paper to ease their comparison and contrast for further analysis. Besides, the analysis could be used in choosing the right sampling technique for optimum outcome.

To be specific, the objective of this review paper is to understand the various sampling techniques for data reduction. This study is intended to find the answers to three main research questions (RQ) which are: RQ1: What are the various types of sampling techniques to reduce data clutter? RQ2: What are the distinct behaviors of these sampling techniques? Finally, RQ3: What is the outcome of these techniques in various dataset and applications?

To achieve the objectives, this paper starts with an introduction, followed by research method. Then, the paper continues with the synthesis of results, followed by discussion. Finally, the last section summarizes and presents conclusions.

II. Research Method

The aim of this paper is to explore the various types of sampling techniques that can assist in encountering the problem of data clutter and provide a snapshot to a direct future design and research.

Since there are huge research papers available in various areas of research, they need to be filtered in order to focus only on related papers that give most input. Thus, this section focuses on the three steps on how the filtering process is designed.

The first step is to identify research questions, followed by a description for each of the questions. This phase is vital as it gives direction to the filtering process. Next, is to identify related keywords based on the research questions. These keywords become the basis for research paper filtering. Three online databases have been identified that are suitable for cluster analysis. They are ACM Digital Library, IEEE Explore and Science Direct. Lastly, inclusion and exclusion criteria for the research paper are identified. These criteria are used to further filter research papers based on the keywords selected. It is to ensure that the research papers are qualified and within the scope of the research.

A. Research Questions

The first step of the research paper filtering process is to formulate research questions. Thus, to achieve this paper's objectives, it has to answers three research questions as depicted in Table I.

TABLE I. The Research Questions Addressed

Research Questions	Motivation
RQ1: What are the various types of sampling techniques to reduce data clutter?	To identify various sampling techniques that have been used to reduce data.
RQ2: What are the distinct behaviors of these sampling techniques?	Identify the distinct behaviors or attributes of these sampling techniques and how they are different from each other.
RQ3: What are the outcomes of these techniques in various dataset and applications?	Find out how these sampling techniques are being applied in previous research to view it from a bigger perspective.

B. Search Strategy

This paper covers English language articles that have been published from 2017 to 2021. The primary collection method is through online databases. Three popular online research databases have been selected which are ACM Digital Library, IEEE Explore, and Science Direct.

The next step is to identify keywords. There are several sets of keywords that have been built based on the research questions.

Next, the keywords are embedded with Boolean operators which are 'OR' and 'AND' operators. Table II shows the research questions and their related keywords.

TABLE II. KEYWORDS FOR RESEARCH QUESTION

Research Question	Keywords
What are the various types of sampling techniques to reduce data clutter?	"Types" OR "Sampling techniques" AND "Reduce" AND "Data clutter" OR "Data overlap"
What are the distinct behaviors of these sampling techniques?	"Distinct" OR "Behaviors" AND "Sampling techniques"
What are the outcomes of these techniques in various dataset and applications?	"Sampling Techniques" AND "Application"

After the keywords have been built, they are then used to filter the three online databases. The first search string retrieves 85,729 search results on ACM Digital Library, IEEE Explore returns 97,668 search results and Science Direct fetches 2,218,712 results. The second search string receives 138,341 from ACM Digital Library, IEEE Explore returns 241,709 results and Science Direct fetches 2,648,324. The last search string receives 144,579 from ACM Digital Library, IEEE Explore fetches 680,554 results and Science Direct receives 11,397 results. The total for all keywords in three different online databases is 6,267,013 articles.

C. Inclusion and Exclusion Criteria

Once the databases have been filtered, the next stage is to further filter the articles through inclusion and exclusion criteria. This step is to ensure the articles fall within the paper scope. Table III shows the inclusion and exclusion criteria for this research.

TABLE III. CRITERIA OF INCLUSION AND EXCLUSION

Inclusion Criteria	Exclusion Criteria
The research papers published in the English language are included	Papers written other than English language are not included
Primary studies like original research papers are selected	Papers that fail to answer the research questions are excluded
Research papers, book chapters that are relevant to the main topic are selected	Elimination of duplicated papers
Research papers ranging from 2016 to 2021 are included in the studies	Research paper less than two pages are removed

The inclusion and exclusion criteria are applied to the 6,267,013 articles filtered earlier. Then, the references from previous systematic reviews are also added, in order to collect as many papers as possible. Finally, a total of 52 papers have been identified suitable for the review. From the 52 articles, they are then grouped into their related sampling techniques. Fig. 2 shows the tabulation of the papers and their topics on sampling techniques.

The total number of papers on simple random sampling is nine while six papers focusing on systematic random sampling. Stratified random sampling is found in seven papers. Moreover, multi-stage sampling has been discussed in three papers while convenient sampling appears in four papers as well as seven papers on snowball sampling. Lastly, quota and judgmental sampling are each discussed in other two papers.

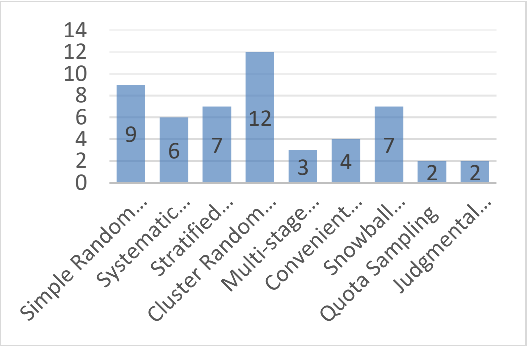

Fig. 2. Histogram for the articles used for each sampling technique

III. SYNTHESIS OF RESULTS

This section answers the research questions as depicted on Table I, by synthesizing and analyzing the knowledge collected from the filtered research papers.

A. RQ1: What are the Various Types of Sampling Techniques to Reduce Data Clutter?

Sampling technique can be divided into two types which are probability sampling and non-probability sampling. Probability sampling is also known as random sampling or representative sampling.

Unlike probability sampling, non-probability sampling does not have a random selection of the sample.

1) Probability sampling: Probability sampling chooses the sample data randomly. The data population must be defined precisely to make the sample selection process easier. The advantage of probability sampling is that it can reduce the chances of systematic error during the sampling process [22]. A systematic error is caused by incorrect measurement of data. Moreover, the sampling technique under-probability sampling can reduce bias that is a common problem in sampling technique [22]. It can select sample data fairly without bias. Moreover, sampling techniques in probability sampling can generate a better sample as compared to non-probability sampling [22].

However, probability sampling demands a lot of training. In addition, to ensure the sample data collected is fair and generates the desired outcome, some calculations need to be applied to the technique. Other than that, probability sampling is timely to generate samples because of the calculation and the implementation of the technique happens layer by layer [23]. There are a few techniques that fall under probability sampling which are simple random sampling, systematic random sampling, stratified random sampling, cluster sampling, and multi-stage sampling.

2) Simple random sampling: One of the sampling techniques that fall under probability sampling is simple random sampling. Simple random sampling is a sampling where each data in the population has a similar chance to be selected as the sample. Each data in the population can participate to be selected as a sample without exception. Usually, the researcher uses computer-generated random numbers to select the sample. Simple random sampling is suitable to be used when the entire population is available and the researcher has a list of all subjects from the population [23]. Fig. 3 below shows how the simple random sampling selects samples from the population.

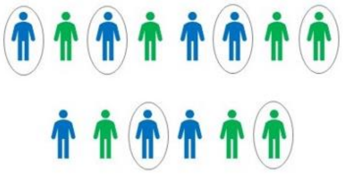

Fig. 3. Selection of sample using simple random sampling

However, using simple random sampling techniques may reduce the possibilities of bias. It is because the dataset is selected randomly, hence there is no bias. The sample from the population is a righteous representative of the whole population [22]. For example, if there exist a few categories in the dataset, all the categories will reform to represent the whole categories. Simple random sampling lowers the vulnerability suitable to the finite size of the sample [23].

Simple random sampling might be pricey and time-consuming if the population involved a broadly spread geographical location. It needs a lot of attempts when a dataset used is large [22]. However, if there is a minority in the population, they will be diminished.

3) Systematic random sampling: The next technique of probability sampling is systematic random sampling which makes use of a specific formula to select the sample. The formula is also known as regular interval [24]. The process of data selection is initiated by selecting a random data and then the selection continues at regular intervals [23].

A population can be described based on any characteristics that are suitable for the studies. The characteristics can be age, gender, race, location, and others, as long as there are different characteristics.

Fig. 4 below shows how systematic random sampling selects the sample from the population.

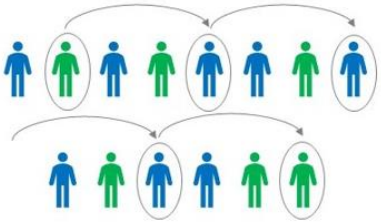

Fig. 4. Selection of sample using systematic random sampling

Systematic sampling is to ensure that the whole population participates in the sampling selection and there is no exception [22]. Thus, there is certainly a sample from each data category since all the categories are involved. Similar to simple random sampling, systematic sampling chooses the sample randomly and the location of the sample is not important as long as all the elements are included [22]. Systematic sampling is less expensive [24].

Due to the calculation that is needed for this technique, it might be timely and requires lots of effort especially if the population involved is scattered in the widely spread geographical area and it is also a challenge to access the population [22].

4) Stratified random sampling: One of the sampling techniques that uses strata or subgroups of the population is stratified random sampling. Stratified random sampling uses a subgroup to give an equal possibility to select data randomly from the strata [21]. The population is grouped into similar characteristics and the sample is selected randomly from the subgroup [22]. Stratified sampling is derived from the simple random sampling; hence the sample frame is needed. Fig. 5 below shows how stratified random sampling is selected from the population.

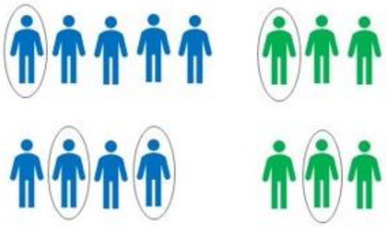

Fig. 5. Selection of sample using stratified random sampling

The advantage of stratified random sampling is the researcher can collect samples from each of the strata and the sample size will be different from each stratum [22]. Stratified random sampling also collects samples from the minority population, hence there is no exception between the majority population and minority population [22]. Thus, stratified sampling can represent the actual data population [25].

However, stratified random sampling is costly, timely and requires a lot of effort due to the subgroup process and randomly selected samples [20, 25]. If the population is not sub grouped into the same characteristic, the entire research may be useless [20]. Hence, before proceeding to select the samples, the researcher must make sure that the population has been sub grouped correctly based on the characteristics.

5) Cluster sampling: Cluster sampling is a sampling technique where the samples of the population are from a geographical area that is spread and possible to be accessed simultaneously [21]. The researcher splits the population into clusters based on the geographical area and later extracts the sample from the cluster. Cluster sampling is similar to simple random sampling as the samples are randomly selected from the cluster. Fig. 6 shows how samples are selected from the population using cluster sampling.

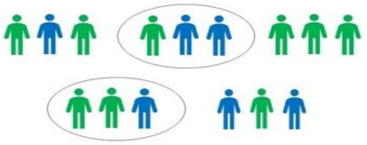

Fig. 6. Selection of sample using stratified random sampling

If the population is widely spread over a geographical area, cluster sampling can lessen the cost, time, and efforts compared to other sampling techniques [26]. Cluster sampling requires less effort and time because the population can be visited once. Moreover, cluster sampling does not need to define the number of clusters prior to processing [27]. The number of clusters is defined after the whole population has been clustered. According to [28], cluster sampling does not require additional information while the algorithm is applied to the population. By using hybrid models, it can enhance the variances for the data [29].

However, cluster sampling might not perform the true diversity of the population [30]. Moreover, biases and systematic errors could happen sometimes. In addition, after the cluster sampling has been applied to the population, it might be challenging to do a correction to the sample [27, 28].

6) Multi-stage sampling: The last sampling technique of probability sampling is multi-stage sampling. Multi-stage sampling is similar to cluster sampling where the population is sub grouped into the same cluster and a sample from the cluster is selected in the next process [21].

The advantages of multi-stage sampling are time and cost efficiency. This is because the population is sub grouped into the same category and samples from the subgroup are selected at a later stage [22] which reduced the process flow. The disadvantage of multi-stage sampling is the sample does not represent the population if the selected clusters do not capture the characteristics of the population [22].

7) Non-probability sampling: Another type of sampling technique is non-probability sampling where selection is not random. The dataset for this type of sampling does not need to be precisely defined. Unlike probability sampling, non-probability sampling can be used either for specific and general categories.

The advantage of using the non-probability type of sampling is that they require less effort and less time to generate the sample [22]. However, their disadvantage is that they are easily exposed to systematic errors and bias issues [20]. Thus, the samples at times might not be an accurate representation of the population [20].

There are a few sampling techniques that fall under non-probability sampling, which are convenient sampling, snowball sampling, quota sampling, and judgment sampling.

8) Convenient sampling: Convenient sampling is the process of selecting sample subjects based on their availability and accessibility [24]. It is common for researchers to conduct interviews from the available pool of respondents as shown in Fig. 7. It is not the researchers' concern if the selected data fail to represent the population.

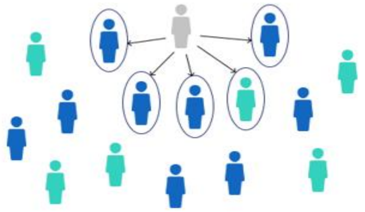

Fig. 7. Selection of sample using convenient sampling

With this kind of method, data could be collected through online channels where respondents are those who are willing to spend time in the data collection process [31]. Thus, it takes less effort. However, convenient sampling is easily facing problems of biasness and systematic error.

9) Snowball sampling: Another non-probability sampling technique is snowball sampling. Snowball sampling is where the researcher makes the first arrangement with a small group of people that are relevant to the subject and uses them as the criterion to contact other people [21]. In other words, data are collected form a small group of respondents and through them, more responded are identified as illustrated in Fig. 8.

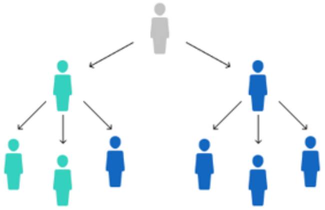

Fig. 8. Selection of sampling using snowball sampling

The technique is suitable when the population is not located in a particular area [24]. It is also suitable for scarce and very small population [22]. Thus, there may occur biases and systematic errors due to non-random network connection [20].

10) Quota sampling: Quota sampling is also categorized as non-probability sampling. Quota sampling is commonly used if the elements of the population is not matched with another characteristic of the criteria that has been defined. The population is sub grouped into its same elements and the quota is set for each subgroup.

The advantage of quota sampling is every single subject in the population has its own subgroup. Compared to stratified sampling, quota sampling is less time consuming and inexpensive [22]. However, quota sampling might not be the best method to represent the whole population. Hence, it cannot counter issues where generalizability needs to be made [20].

11) Judgmental sampling: Judgmental sampling is the sampling where the subjects of the population are selected by the researcher [20]. The process starts with the researcher generally evaluating the population's characteristics. From there, samples are selected with the aim that they represent the whole population as illustrated in Fig. 9.

There are not many researchers who use this sampling technique because the researcher's judgment might be biased [22].

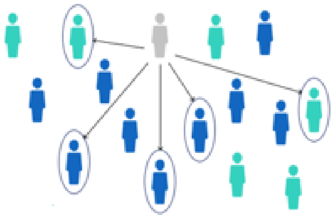

Fig. 9. Selection of sample using judgmental sampling

In short there are various sampling techniques, and they can be grouped into probability and non-probability. Each has its own behaviors and characteristics. Thus, the next section answers the RQ2 where the sampling techniques differences are analyzed.

B. *RQ2: What are the Distinct behaviours of these Tecniques?*

To answer research question 2 (RQ2) about the distinct behaviors of the sampling techniques, this section compares and contrasts these sampling techniques. Table IV shows their comparison.

TABLE IV. COMPARISON OF SAMPLING TECHNIQUES

Behaviors Sampling Techniques	Use sample	Costly	Bias problem	Timely	Sample represent population	Lots of efforts	Systematic error	Size of population
Simple Random Sampling	Yes	Depends	No	Depends	Yes	No	-	Large
Systematic Random Sampling	Yes	No	-	Yes	Yes	Yes	-	Large
Stratified Random Sapling	Yes	Yes	-	Yes	Yes	Yes	-	Large
Cluster Sampling	Yes	No	Yes	No	Yes	No	Yes	Large
Multi-stage Sampling	Yes	-	-	-	-	-	-	Large
Convenient Sampling	No	No	-	-	-	No	-	From small to large
Snowball Sampling	No	No	Yes	Yes	-	Yes	Yes	From small to large
Quota Sampling	Yes	No	-	No	No	No	-	Small
Judgmental Sampling	No	No	-	No	-	Yes	-	From small to large

The comparison involves a group of probability and non-probability sampling techniques which have been introduced in the earlier sections. The comparison involves a few attributes which are sample, cost, bias problem, time-consuming, population representation, efforts used, systematic error and size of population.

In terms of costing, stratified technique is more costly as compared to others. On the other hand, costing for simple random sampling depends on the dataset size. The bigger the size the more cost involved. Sampling techniques that are less costly are systematic random sampling, cluster sampling, convenient sampling, snowball sampling, quota sampling and judgmental sampling.

Next attribute that is important in the area or sampling is bias. Sampling techniques that are prone to bias are cluster sampling and snowball sampling. Meanwhile, simple random sampling is the least with the issue of bias.

Timely is another attribute for comparison. Sampling techniques that are timely to be executed are systematic random sampling, stratified random sampling and snowball sampling. On the other hand, sampling techniques that are not commonly related to timely issues are cluster sampling, quota sampling and judgmental sampling. While simple random sampling depends on the size of population or dataset. The bigger size, the longer sampling process.

In terms of accurately representing population, the sampling techniques are simple random sampling, systematic random sampling, stratified random sampling and cluster sampling. Meanwhile, quota sampling is not included in this group of accurately representing the population.

Besides, effort is another attribute of sampling techniques that differ from each other. Systematic random sampling, stratified random sampling, snowball sampling and judgmental sampling require more effort for implementation. On the contemporary level, simple random sampling, cluster sampling, convenient sampling and quota sampling are categorized as less effort when it comes to implementation.

Systematic error is another vital attribute of sampling technique that needs to be compared. Cluster sampling and snowball sampling are prone to systematic error. While for other sampling techniques, it could not be identified whether they are prone to systematic error.

Lastly, population size varies with sampling techniques. Different techniques are suitable for different population sizes. Techniques that use large populations are simple random sampling, systematic random sampling, stratified random sampling, cluster sampling and multi-stage sampling. On the other hand, quota sampling is suitable for a small population. However, convenient sampling, snowball sampling and judgmental sampling need two steps. Thus, the first step is to make use of the small population and the population will grow with time.

C. *RQ3: What is the Outcome of these Techniques in Various Dataset and Application*

This section is to identify the various applications of the sampling techniques based on the previous research.

The discussion starts with probability sampling and continues with a non-sampling group of techniques.

1) Probability sampling: This section discusses the application of various probability sampling in various areas.

2) Simple random sampling: Many previous studies used simple random sampling as the technique in their studies. According to [32], to find the random information that is adjacent to the expected distribution, the anticipated distribution and the sampling distribution are extricated frequently by using simple random sampling. The purpose of extracting frequently is to make sure that the population is placed randomly and not follow any arrangement. By combining the simple random sampling technique with Arithmetic Mean which is then known as Random Sampling-Arithmetic Mean (RS-AM), it could reduce the data conflict and increase the sampling accuracy [32]. The benefit of using RS-AM is that an efficient calculation method could be achieved if the distance function is used effectively [32].

Based on [33], block pool is the output of I-sampling. It is also known as the block-based sampling method. It is used with data that has almost similar probability distribution.

The sampling process start with splitting a large dataset into non-overlap data. Next, block pools are created. Finally, the data blocks are randomly be selected from the block pool. This is to ensure data are selected fairly and without bias.

In addition, by using distinct data sizes to randomly extract the documents especially in medical, the chances of certain words appearing in the documents and sentences are high. Identifying the distance between target words in different dimensions is used to categorize the documents and to differentiate the documents. Each alphabet has its own size and distance between each stroke. For the word centroids distance training, random sampling is applied to verify the accuracy of the output and the variety of the selected documents [30, 48].

Simple random sampling can also be used to allocate resources to the network and all subsystems in the Markov chain [31, 34]. It has been proven that simple random sampling can be used in the network and not limited to dataset. Due to the power supply and network restraint, it is suitable to make a schedule especially a wireless network system that incorporates the immense number of nodes. Scheduling is used to ensure the stability of performance specification in the network systems. Scheduling performs independent control loops based on the requirement and overall network supplies.

3) Systematic random sampling: Systematic sampling technique is used in an alternative method to estimate the Banzhaf-Owen value in a large class of TU games with scientific [32, 34]. The systematic sampling technique is suitable with the arrangement of a priori unions. Systematic sampling is used as a technique in this study because this technique can handle memory size issues and can reduce massive situations.

In addition, Stochastic bilevel programming is a program that has randomness in the problem, hence it will face problems with computationally expensive and challenging [33, 34]. The randomness that happens in bilevel programming is the randomness property and hierarchical nature of optimization. Systematic sampling is chosen as the algorithm because this technique will determine a representative from the leader's opinion and hybrid particle swarm optimization procedure.

Another application of systematic sampling is to measure the correlation of two objects. The correlation is defined as the max-min distance between two points set and could identify the parallel between two points [34, 48]. A Measure two-point set is known as Hausdorff distance or HD. HD is difficult to figure out because of a very massive scale point set and at the same time to assure the certainty of HD. Systematic sampling is chosen as the algorithm because it spends less time to choose the sample and achieve maximum dispersion of the sample and the starting point of this sampling is selected random samples [34, 48].

4) Stratified random sampling: Stratified sampling is a widely used tool for variance reduction used for failure probability estimation. This study merges the stratified sampling with importance sampling and stratified importance sampling. Stratified sampling is used to reduce the samples to approximately a failure probability with the same coefficient of variation [34, 35].

The common use of stratified sampling is to subgroup the population [34, 36]. Then samples are selected from the subgroup. The sample size is to resolve the strata by acknowledging the compatibility between sampling design and load research objectives. The population from the study consists of electricity tariffs, contract power, geographical area, and region type. Stratified random sampling subgroups the population into the same categories.

In the study conducted by [37, 48], stratified sampling is used to preprocess the wastewater condition dataset that has a sophisticated nonlinear relationship, performance, and MLR models that are not good. The researcher partitions the dataset into different subgroups and selects data points from distinct strata for different purposes. Stratified random sampling trains and tests the dataset that contains the same proposition of each class label [37].

5) Cluster sampling: A study conducted by [38] stated that the cluster sampling technique is applied to get a sample from Java software that consists of a similar system and to display the differences between the clusters. The software is grouping into the cluster using the CrossSim algorithm to observe the similarities. After finding out the similarities, the software is clustered using cluster sampling. Lastly, to extract the description of the project and group the systems based on the extraction, Python implementation of the Latent Dirichlet Allocation (LDA) is used.

Moreover, the clustering technique can be used as an alternative technique for subdividing the input space especially involving the high-dimension input spaces and to model MFs into partitions. This study uses the cluster sampling approach initialized ANFIS and MF which can take the entire advantage of intrinsic data distribution. The cluster sampling is used to develop ANFIS nearest-neighbourhood and to allow the online generation of advanced rules by excluding the nearest-neighbourhood that is not effective anymore [20, 39].

On the other hand, conditional cluster sampling is used as the approach for the current pandemic around the world that is COVID-19. This study is conducted by [40] where conditional cluster sampling is used to test patients in pools rather than individual testing by using a numerical method, statistical data, and machine learning. After the output has been generated for each pool, the researcher makes a decision either to continue the testing or abort it. Conditional cluster sampling is applied to cluster the population depending on the patients' condition. The major reason behind this study is the possibility of COVID-19 critical patients is higher than other patients' diseases.

Lastly [41] stated that the cluster sampling is not efficient as SRS, but it is cost-efficient in the statistical information that is to come across a wide geographical area. Cluster sampling is the most economical when a group of the population element establishes a sampling unit than a single element [41].

6) Multi-stage sampling: There are a few studies that used multi-stage sampling as the approach. One of the studies is conducted by [42], where the multistage sampling is the extension to the acceptance sampling technique where the inspection happened several times and only accepted if it travelled as such it covered all stages. This approach is called the multi-stage acceptance technique when the starting test is pursued by the next inspections.

7) Non-probability sampling: This section discussed the non-probability type of sampling technique.

8) Snowball sampling: The study conducted by [43] stated that snowball sampling is used when it is difficult to approach subjects with distinct characteristics. The study used qualitative research which coordinates the approach of describing people's experiences and internal feelings. This research collects data with a different approach such as interviews, observations, focus groups, narratives, notes, reports, and a review of archives. Snowball sampling assembles the information to approach specific groups of people.

Moreover, [44] using snowball sampling to create samples from simulated networks and cut down the distance from network statistics across network sizes. In this study, snowball sampling generates samples with an identical number of waves and seeds as the samples taken from simulated population networks.

Besides, snowball sampling is used to identify the effectiveness of assumptions about the existence of effects such as network closure and attribute homophiles [45]. This study uses snowball sampling to generate specimens of nodes in a network by applying the network structure itself that can be represented as follows.

Lastly, snowball sampling starts with the entity that has preference characteristics and uses that individual's connection to attract other people with the same characteristics [46]. This study uses a snowball sampling to gain information about the mothers with children that have developmental disabilities. Mothers are requested to pass the information to other mothers that might have a child that suffered from developmental disabilities. However, previous studies for judgmental sampling, quota sampling, and convenient sampling are quite difficult to find.

IV. Conclusion

There are two main types of sampling techniques which are probability sampling and non-probability sampling. The sampling techniques that fall under probability sampling are simple random sampling, systematic random sampling, stratified random sampling, cluster sampling, and multi-stage sampling.

The population involved in the probability sampling can be large as the sampling techniques take the samples directly from the population. Hence, the large population should ensure that the output generated is correct and accurate. Each sampling technique in probability sampling has its methods or calculation to take samples from the population.

The sampling techniques for the non-probability sampling are convenience sampling, snowball sampling, quota sampling, and judgment sampling. The majority of the sampling techniques in the non-probability sampling are suitable for small populations. Thus, most researchers use this sampling technique in their research because of the population size.

On the other hand, non-probability sampling is not quite popular among researchers because of the small population. This can lead to inaccurate output. Therefore, in this paper, there are sampling techniques in non-probability sampling that do not have previous studies. Besides, non-probability sampling majorly uses primary data which could become a burden to researchers.

Probability sampling is usually used by the researcher because the data can be primary and secondary. Hence, the researcher can choose how to gather the data (primary or secondary). Moreover, probability sampling uses a large population that can lead to accurate results because of the number of the population involved in the research.

To make researchers to understand the pattern well, researcher can visualize the graph. One of the well-known visualization graphs is parallel coordinates as it can be used for large and multi-dimensional data set visualization [46, 47]. One of benefits of using parallel coordinates graph is the ability to identify the relationship of multivariate data [20].

References

[1] Walny, J., Frisson, C., West, M., Kosminsky, D., Knudsen, S., Carpendale, S., & Willett, W. “Data Changes Everything : Challenges and Opportunities in Data Visualization Design Handoff.”, *IEEE Transactions On Visualization And Computer Graphics, 26*(1), 12–22, 2020.

[2] Taylor-sakyi, K. “*Big Data : Understanding Big Data*.”, January 2016.

[3] Arockia Panimalar.S, Komal M.Khule, Karthika.S, N. K. “Data Visualization Tools and Techniques For Datasets In Big Data., *International Research Journal of Engineering and Technology(IRJET)*, 4(8). Retrieved from https://irjet.net/archives/V4/i8/IRJET-V4I8296.pdf, 2017

[4] Babu, A. G. L., Reddy, S. G., Agarwal, & Swathi. " An effective approach for Visualizing Big Data.", *International Journal of Innovations in Engineering and Technology*, *7*(2), 77–81,2016.

[5] Samuel, S. A., & Anthonia, A. "An overview of big data visualization techniques in data mining.", *International Journal of Computer Science and Information Technology Research*, *4*(3), 105–113. Retrieved from https://www.researchgate.net/publication/305905594, 2016

[6] Bryner, D., Huffer, F., Rosenthal, M., Tucker, J. D., & Srivastava, A. "Estimation of linear target-layer trajectories using cluttered point cloud data.", *Computational Statistics and Data Analysis*, *102*, 1–22. https://doi.org/10.1016/j.csda.2016.04.002, 2016

[7] Idrus, Z., Zainuddin, H., Ja'afar, A.D.M.: Visual analytics: designing flexible filtering in parallel coordinate graph. J. Fundam. Appl. Sci. **9**, 23–32. https://doi.org/10.4314/jfas.v9i5s.3, 2019

[8] Li, T., De la Prieta Pintado, F., Corchado, J. M., & Bajo, J. "Multi-source homogeneous data clustering for multi-target detection from cluttered background with misdetection.", *Applied Soft Computing Journal*, *60*, 436–446. https://doi.org/10.1016/j.asoc.2017.07.012, 2017

[9] Yongjoo Park, Michael Cafarella, B. M. "*Visualization-Aware Sampling for Very Large Databases*.", *2*, 5,2015.

[10] Bum Chul Kwon and Janu Verma, Peter J. Haas, C. D. "Visualization Viewpoints.", *IEEE Computer Society*, (February), 100–108, 2017.

[11] Li, S. "*Data Reduction Techniques for Scientific Visualization and Data Analysis", *Vol. 36, 2017.

[12] Rojas, J. A. R. "Sampling Techniques to Improve Big Data Exploration.", *IEEE Symposium on Large Data Analysis and Visualization*, 26–35, 2017.

[13] Bhardwaj, P. "Types of Sampling in Research.", *2019 Journal of the Practice of Cardiovascular Sciences*, 157–163. https://doi.org/10.4103/jpcs.jpcs, 2019

[14] Iqbal Jeelani M, F. D. and M. G. "A Review on the Recent Development on the Cluster Sampling.", *Biostatics and Biometrics Open Access Journal*, *5*(5), 146–150. https://doi.org/10.19080/BBOAJ.2018.05.555673, 2018

[15] Bruno, P., Calimeri, F., Kitanidis, A. S., & De Momi, E. "Data reduction and data visualization for automatic diagnosis using gene expression and clinical data.", *Artificial Intelligence in Medicine*, *107*(November 2019), 101884. https://doi.org/10.1016/j.artmed.2020.101884, 2020

[16] Hepworth, K. "Big Data Visualization : Promises & Pitfalls.", *Communication Design Quarterly*, *4*, 7–19, 2017.

[17] Alvi, M. H. "*A Manual for Selecting Sampling Techniques in Research Mohsin Hassan Alvi",* 2016.

[18] Rahi, S. "Research Design and Methods : A Systematic Review of Research Paradigms, Sampling Issues and Instruments Development.", *International Journal of Economics & Management Sciences*, *6*(2), 1–5. https://doi.org/10.4172/2162-6359.1000403, 2017

[19] Imad Rida, S. A. "AN ENSEMBLE LEARNING METHOD BASED ON RANDOM SUBSPACE SAMPLING FOR PALMPRINT IDENTIFICATION.", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2047-2051, 2018.

[20] Qing Xu, G. S. "Generalizing systematic adaptive cluster sampling for forest ecosystem inventory.", *Forest Ecology and Management*, 1-10, 2021.

[21] Elfil, M., & Negida, A. "Sampling methods in Clinical Research ; an Educational Review.", 5(1), 3–5, 2017.

[22] Bashar I. Ahmad, A. T. "Spectral Analysis of Stratified Sampling: A Means to Perform Efficient Multiband Spectrum Sensing.", *IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS*, 178-187, 2012.

[23] Saumya Singh, S. S. "Review of Clustering Techniques in Control System.", International Conference on Smart Sustainable Intelligent Computing and Application under ICITETM 2020, 272-280, 2020.

[24] Sumanta Das, J. C. "UAV-Thermal imaging and agglomerative hierarchical clustering.", *ISPRS Journal of Photogrammetry and Remote Sensing*, 221-237, 2021.

[25] Aly, H. H. "A proposed intelligent short-term load forecasting hybrid models of ANN, WNN and KF based on clustering techniques for smart grid.", *Electric Power Systems Research*, 1-13, 2021.

[26] Ly, T., Cockburn, M., & Langholz, B. "Cost-efficient case-control cluster sampling designs for population-based epidemiological studies.", *Spatial and Spatio-Temporal Epidemiology*, *26*, 95–105. https://doi.org/10.1016/j.sste.2018.05.002, 2018

[27] Yu Han, Q. S.-Y.-L. "A convenient sampling and noninvasive dried spot method of uric acid in human saliva: Comparison of serum uric acid value and salivary uric acid in healthy volunteers and hyperuricemia patients.", *Journal of Chromatography B*, 1-8, 2021.

[28] Tian, S., Zhang, J., Chen, L., Liu, H., & Wang, Y. "Random Sampling-Arithmetic Mean: A Simple Method of Meteorological Data Quality Control Based on Random Observation Thought.", *IEEE Access*, *8*, 226999–227013. https://doi.org/10.1109/ACCESS.2020.3045434, 2020

[29] He, Y., Huang, J. Z., Long, H., Wang, Q., & Wei, C. "I-Sampling: A New Block-Based Sampling Method for Large-Scale Dataset.", *Proceedings - 2017 IEEE 6th International Congress on Big Data, BigData Congress 2017*, 360–367. https://doi.org/10.1109/BigDataCongress.2017.53, 2017

[30] Xie, H., Bin Ahmadon, M. A., Yamaguchi, S., & Toyoshima, I. "Random Sampling and Inductive Ability Evaluation of Word Embedding in Medical Literature.", *2019 IEEE International Conference on Consumer Electronics, ICCE 2019*. https://doi.org/10.1109/ICCE.2019.8662022, 2019

[31] Lu, Z., Zhuang, Y., & Yuan, L. "Random sampling and performance analysis for networked systems.", *Proceedings off the 36th Chinese Control Conference*, 7847–7851. https://doi.org/10.23919/ChiCC.2017.8028597, 2017

[32] Saavedra-nieves, A. "Assessing systematic sampling in estimating the Banzhaf – Owen value.", *Operations Research Letters*, *48*(6), 725–731. https://doi.org/10.1016/j.orl.2020.08.015, 2020

[33] Goshu, N. N., & Kassa, S. M. "Computers and Operations Research A systematic sampling evolutionary (SSE) method for stochastic bilevel programming problems.", *Computers and Operations Research*, *120*, 104942. https://doi.org/10.1016/j.cor.2020.104942, 2020

[34] Ryu, J., & Kamata, S. "An efficient computational algorithm for Hausdorff distance based on points-ruling-out and systematic random sampling.", *Pattern Recognition*, *114*, 107857. https://doi.org/10.1016/j.patcog.2021.107857, 2021

[35] Xiao, S., Oladyshkin, S., & Nowak, W. "Reliability analysis with stratified importance sampling based on adaptive Kriging.", *Reliability Engineering and System Safety*, *197*(January), 106852. https://doi.org/10.1016/j.ress.2020.106852, 2020

[36] Raeisi-Gahrooei, Y., Khodabakhshian, A., & Hooshmand, R. A. "A new stratified random sample customer selection for load research study in distribution networks.", *International Journal of Electrical Power and Energy Systems*, *97*(July 2017), 363–371. https://doi.org/10.1016/j.ijepes.2017.11.029, 2018

[37] Fu, Z., Cheng, J., Yang, M., Batista, J., & Jiang, Y. "Wastewater discharge quality prediction using stratified sampling and wavelet de-noising ANFIS model.", *Computers and Electrical Engineering*, *85*. https://doi.org/10.1016/j.compeleceng.2020.106701, 2020

[38] Capiluppi, A., Di Ruscio, D., Di Rocco, J., Nguyen, P. T., & Ajienka, N. "Detecting Java software similarities by using different clustering techniques.", *Information and Software Technology*, *122*(September 2019), 106279. https://doi.org/10.1016/j.infsof.2020.106279, 2020

[39] Leonori, S., Martino, A., Luzi, M., Frattale Mascioli, F. M., & Rizzi, A. "A generalized framework for ANFIS synthesis procedures by clustering techniques.", *Applied Soft Computing Journal*, *96*, 106622. https://doi.org/10.1016/j.asoc.2020.106622, 2020

[40] Zoha, N., Ghosh, S. K., Arif-Ul-Islam, M., & Ghosh, T. "A numerical approach to maximize the number of testing of COVID-19 using conditional cluster sampling method.", *Informatics in Medicine Unlocked*, *23*, 100532. https://doi.org/10.1016/j.imu.2021.100532, 2021

[41] Ly, T., Cockburn, M., & Langholz, B. "Cost-efficient case-control cluster sampling designs for population-based epidemiological studies.", *Spatial and Spatio-Temporal Epidemiology*, *26*, 95–105. https://doi.org/10.1016/j.sste.2018.05.002, 2018

[42] Sommer, A., & Steland, A. "Journal of Statistical Planning and Inference Multistage acceptance sampling under nonparametric

dependent sampling designs.", *Journal of Statistical Planning and Inference*, *199*, 89–113. https://doi.org/10.1016/j.jspi.2018.05.006, 2019

[43] Naderifar, M., Goli, H., & Ghaljaie, F. "Snowball Sampling : A Purposeful Method of Sampling in Qualitative.", *Strides in Development of Medical Education*, *14*(3). https://doi.org/10.5812/sdme.67670.Research, 2017

[44] Rolls, D. A., & Robins, G. "Minimum distance estimators of population size from snowball samples using conditional estimation and scaling of exponential random graph models.", *Computational Statistics and Data Analysis*, *116*, 32–48. https://doi.org/10.1016/j.csda.2017.07.004, 2017

[45] Stivala, A. D., Koskinen, J. H., Rolls, D. A., Wang, P., & Robins, G. L. "Snowball sampling for estimating exponential random graph models for large networks.", *Social Networks*, *47*, 167–188. https://doi.org/10.1016/j.socnet.2015.11.003, 2016

[46] Lee, J., & Spratling, R. "Recruiting Mothers of Children With Developmental Disabilities : Adaptations of the Snowball Sampling Technique Using Social Media.", *Journal of Pediatric Health Care*, *33*(1), 107–110. https://doi.org/10.1016/j.pedhc.2018.09.011, 2019

[47] N. N. S. 'Asri, Z. Idrus, H. Zainuddin and Z. Idrus, "Parallel Coordinates Graph in Bundling Technique," 2019 1st International Informatics and Software Engineering Conference (UBMYK), pp. 1-6, https://doi.org10.1109/UBMYK48245.2019.8965511, 2019

[48] Idrus, Zainura & Zainuddin, H. & Ja'afar, A.D.M. Visual analytics: designing flexible filtering in parallel coordinate graph. Journal of Fundamental and Applied Sciences. 9. 23. https://doi.org/10.4314/jfas.v9i5s.3, 2018