ADRI
INTERNATIONAL JOURNAL
OF ENGINEERING AND
NATURAL SCIENCE

VOLUME 5 | NUMBER 2 | NOVEMBER 2020
ISSN: 2656-1174 (online)

K-Means Clustering of COVID-19 Cases in Indonesia's Provinces

Fitria Virgantari and Yasmin Erika Faridhan
Department of Mathematics, Pakuan University
Bogor 16143, Indonesia
fitriav12@gmail.com, yasmin.faridhan@gmail.com

Abstract

The novel coronavirus disease (COVID-19) has been rapidly spreading, causing a severe health crisis all around the world, including Indonesia. As expected, due to Indonesia's diverse topography and population, there are variations in the number of cases amongst its provinces. Therefore clustering is needed to develop a map of COVID-19 cases to enable optimal handling of this pandemic. The provinces are clustered using K-means method according to their respective COVID-19 case numbers. Data taken from Indonesian Ministry of Health in November 2020 is used in this study, covering COVID-19 cases in Indonesia's 34 provinces. K-means results in seven optimal clusters with variance ratio of 0.185. Clusters 1 to 3 cover most provinces in Java, including DKI Jakarta in Cluster 1 as the province with the most cases. Each of Clusters 4 and 5 consists of 5 provinces, while each of Clusters 6 and 7 consists of 10 provinces. Cluster 7 comprises provinces with lowest cases of COVID-19.

Keywords
Clustering, Covid-19, Indonesia's provinces, K-means

1. Introduction

The novel coronavirus disease SARS-CoV-2, or widely known as COVID-19, has been declared a pandemic by WHO on 11th March 2020 . Countries around all the world battle to contain the highly contagious virus, and numerous research regarding COVID-19 have been conducted in many aspects and levels. Regardless of the case levels, whether worldwide, regional, national or even states/provinces as well as cities within a country, clustering is seen as one way to map COVID-19 cases in order to assist in handling the pandemic. Some of the studies including Zarikas, Poulopoulos, Gareiou, & Zervas, (2020) who propose a hierarchical algorithm in clustering worldwide countries, and Mahmoudi, Baleanu, Mansor, Tuan, & Pho, (2020) who employs Fuzzy clustering on seven high-risk countries.

In terms of Indonesia, with its diverse topography and population, variations in the number of COVID-19 cases do occur amongst its 34 provinces. Clustering local cases using K-means method seems quite popular. Using this method, Dwitri, Tampubolon, Prayoga, Zer, & Hartama, (2020) cluster Indonesia's COVID-19 cases per May 2020. Solichin & Khairunnisa, (2020) also exercise this method in clustering COVID-19 spread in Jakarta.

In this paper, K-means method will be employed in clustering COVID-19 cases in Indonesia's provinces using November 2020 data. Different pre-defined number of clusters will be attempted to determine optimal number of clusters. While Dwitri, Tampubolon, Prayoga, Zer, & Hartama, (2020)'s study is more technical and demonstrates a software to run the clustering method, our study will look more into the statistical side. Visual map of the cases' spread in Indonesia will also be included.

The structure of this paper is as follows. Section 2 describes the clustering method, particularly K-means method. In Section 3 variables and data source are described. Section 4 reports the findings, and Section 5 concludes the paper.

2. Cluster Analysis

There have been many variants of clustering methods proposed in recent decades. Broadly speaking, each clustering method may be seen belonging into either hierarchical or non-hierarchical categories. Hierarchical methods are used to structurally group observations based on their similarities; the number of desired clusters is not determined in advance. On the other hand, non-hierarchical methods start by determining the number of desired clusters (Ma & Chow, 2004),(Mattjik & Sumertajaya, 2011).

Both hierarchical and non-hierarchical clustering categories comprise several sub-categories. K-means belongs to non-hierarchical category, specifically in partitioning sub-category. This particular clustering method will be discussed further in the next subsection.

2.1 K-Means Clustering

Of several sub-categories in non-hierarchical clustering category, partitioning is the most widely used. K-Means is the most popular method in the partitioning sub-category. Other methods such as K-Modes, K-Prototypes, and Fuzzy C-Means are its variants (Berkhin, 2006), (Ma & Chow, 2004). This method assigns each object to the cluster that has the nearest centroid, or mean.

Johnson & Wichern, (2007) elaborates K-means algorithm in its simplest version, starting with defining *K* initial centroids as seed points. Using Euclidean distance, the distance between object *i* and cluster *l* is defined as

$$d(i,l) = \left(\sum_{j=1}^{p}[x(i,j) - \bar{x}(i,j)]^2\right)^{\frac{1}{2}} \quad (1)$$

with

$$E[p(n,k)] = \sum_{i=1}^{n} D[i,l(i)]^2 \quad (2)$$

where $\bar{x}(i,j)$ denotes the average of variable *j* in cluster *l*; $E[p(n,k)]$ is the partition error, *1(i)* is the cluster containing object *i*; and *D[i,1(i)]* is the distance between object *i* and average of the cluster containing the object *i* (Nugroho, 2008). This initial selection partly determines the final assignment of objects the final clusters. It is advisable to re-run the algorithm with different initial partition.

It should be noted that pre-defining the number of clusters *K* must be approached with caution. Johnson & Wichern, (2007) summarise pitfalls that might happen as a result of pre-defining *K*:

1. If two or more seed points happen to lie within a single cluster, the resulting clusters will be poorly differentiated.
2. Should there be an outlier, it might yield at least one group with dissimilar objects.
3. As the sampling method may not select data from the rarest group, forcing data into pre-defined *K* groups would result in nonsensical cluster.

2.2 Measuring Cluster Validity

Comparison on cluster methods includes measuring validity on the resulting clusters, which is whether the clusters are well separated. As proposed by Ray & Turi, (2000) an ideal cluster should minimize the variance within clusters while maximizing the variance between clusters. Ali Ridho Barakbah & Kohei Aarai, (2004) give the formula as follows. The variance of cluster *k* can be determined using

$${V_k}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_k)^2 \quad (3)$$

Given N as the number of members in all clusters, the variance within clusters can be defined as

$${V_w}^2 = \frac{1}{N-k}\sum_{i=1}^{k}(n_i - 1)\,{V_i}^2 \quad (5)$$

Variance between clusters (${V_b}^2$) is defined as

$${V_b}^2 = \frac{1}{(k-1)}\sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2 \quad (6)$$

where $\bar{x}$ is the grand mean of all clusters.

An ideal cluster should have a minimum ${V_w}^2$and a maximum ${V_b}^2$; in other words, a minimum v, also known as variance ratio, where

$$v = \frac{{V_w}^2}{{V_b}^2} x100\% \quad (7)$$

3. Data and Variables

K-means method is employed to COVID-19 data in Indonesia's 34 provinces, provided by Indonesia's Ministry of Health (Kementrian Kesehatan Republik Indonesia, 2020) as per 8th November 2020. There are three variables originally exercised in this paper, i.e. numbers of COVID-19 confirmed cases, recovered, and death. One variable i.e. case fatality rate (CFR) is added to show the measure of severity of the disease (Scientific Brief, 2020). The case fatality rate is simply calculated as the number of confirmed deaths divided by the number of confirmed cases (Our World in Data, 2020).

Table 1 shows descriptive statistics results of COVID-19 cases in Indonesia's 34 provinces. Indonesia's CFR as of this date is slightly higher than worldwide CFR at the same day[1], i.e. 2.94 per cent compared to 2.51%, respectively. The highest CFR at 7.15 per cent in East Java is quite alarming, as well as the fact that there are 15 provinces with CFR above 3 per cent. As per 27th November 2020, Indonesia's CFR rises to 3.2 per cent, while worldwide CFR decreases to 2.4 per cent (Kementrian Kesehatan Republik Indonesia, 2020).

VOLUME 5 | NUMBER 2 | NOVEMBER 2020
ISSN: 2656-1174 (online)

Table 1. Descriptive statistics of COVID-19 cases in Indonesia per 8th November 2020, N = 34 provinces

Cases	Minimum	Mean	Maximum	Standard Deviation
Confirmed	680	12,813	111,000	21,223
Recovered	556	10,726	101,000	18,921
Deaths	7	428	3,884	787
Case Fatality Rate (%)	1.03	2.94	7.15	1.49

Figure 1 depicts the multiple box-plots for four variables of COVID-19 measures in Indonesia's 34 provinces, based on respective z-scores. There are outliers in confirmed cases, recovered, as well as deaths. The top four provinces with most confirmed cases are all located in Java, i.e. DKI Jakarta, East Java, West Java, and Central Java. Interestingly, DKI Jakarta is also the province with the most recovered cases; followed by the same three of Java's provinces. In terms of deaths, East Java holds the highest number of 3,884; DKI Jakarta and Central Java are second and third, respectively, with both numbers of deaths above 1,000. West Java comes fourth with less than 800 deaths.

There is no outlier in case fatality rate; however as the CFR distribution is skewed to the right, this means that there are a number of provinces with quite large CFR, far beyond the mean of 2.94 per cent. While the local governments in these provinces have been attempting hard to contain the pandemic in their respective areas, it is hoped that frequent publications of these statistics will inform the residents in raising their awareness of the COVID-19 spread and risks.

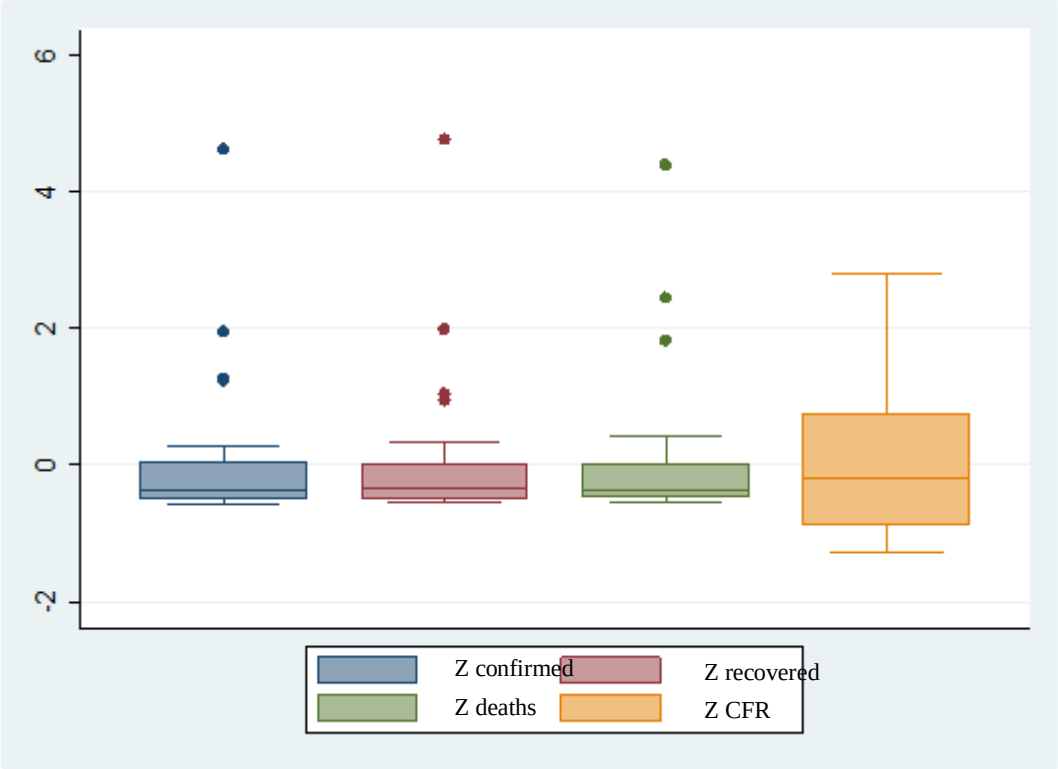

Figure 1. Box-plots of COVID-19 cases in Indonesia per 8th November 2020, N = 34 provinces (standardized values)

VOLUME 5 | NUMBER 2 | NOVEMBER 2020
ISSN: 2656-1174 (online)

4. Results and Discussion

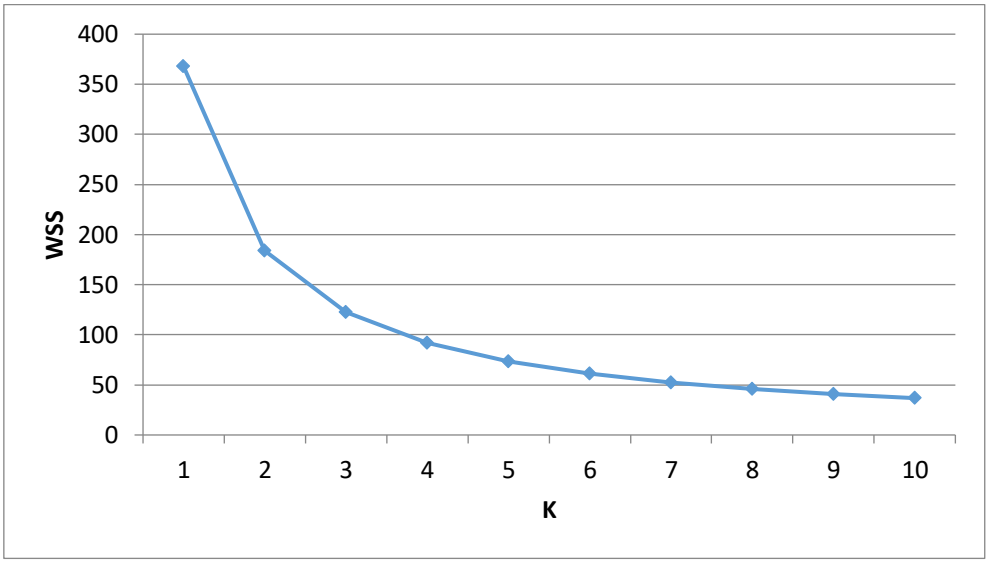

Figure 2. Within cluster sum of squares for every pre-defined number of clusters, *K*

In this study, K-means algorithms have been re-run several times using different pre-defined number of clusters to determine the optimal number of clusters in analyzing Indonesia's COVID-19 data. Figure 2 displays each attempted number of clusters *K* against their respective within cluster sum of squares (WSS). $K = 7$ is preferred over larger numbers 8, 9, and 10, due to small reduction in WSS for $K > 7$. Another consideration is that eight or more clusters in grouping 34 provinces might make many clusters with only a few members. The variance ratio for $K = 7$ is 0.185.

Table 2. Description of COVID-19 confirmed cases in each of the seven clusters formed by K-means

Cluster	n	Mean	Std. Dev.	Min	Max
1	1	111,000	0	111,000	111,000
2	1	54,349	0	54,349	54,349
3	2	39,347	378	39,080	39,614
4	5	15,832	1,807	13,665	18,683
5	5	10,390	1,737	8,267	12,181
6	10	4,715	1,252	3,037	7,661
7	10	1,333	567	680	2,240

Table 2 describes basic statistics of COVID-19 confirmed cases for each of the seven clusters formed by K-means. Each of Clusters 1 and 2 only has one member. Clusters 4 and 5 both have five members. Clusters 6 and 7 both have ten members, with Cluster 7 consists of provinces with the lowest confirmed cases. Cluster 7 also has smaller variance compared to Clusters 4, 5, and 6.

VOLUME 5 | NUMBER 2 | NOVEMBER 2020
ISSN: 2656-1174 (online)

Table 3. Members of clusters formed by K-means

Cluster	Member(s)
1	DKI Jakarta
2	East Java
3	Central Java, West Java
4	Riau, East Kalimantan, South Sulawesi, West Sumatra, North Sumatra
5	South Sumatera, Bali, Papua, Banten, South Kalimantan
6	Gorontalo, Maluku, DI Yogyakarta, West Papua, Central Kalimantan, Aceh, South East Sulawesi, West Nusa Tenggara, Riau Islands, North Sulawesi
7	North Maluku, Lampung, West Kalimantan, Jambi, Bengkulu, West Sulawesi, Central Sulawesi, North Kalimantan, East Nusa Tenggara, Bangka Belitung Islands

Table 3 indicates the provinces that belong to each cluster. Members of Clusters 1 to 3 are outliers shown in Figure 1. The numbers of confirmed cases in these provinces seem to correspond with their respective population. The capital province DKI Jakarta as well as the three biggest Java provinces, i.e. West Java, Central Java, and East Java, are the most populated provinces in Indonesia. On the other hand, members of Clusters 4 to 7 are more widespread. Cluster 4 comprises nearby provinces North Sumatera, West Sumatera and Riau, as well as two other provinces outside Sumatera, i.e. East Kalimantan and South Sulawesi.

Since K-means is not a hierarchical method, there is no dendrogram for the clustering results in this study. Alternatively, as the clustering results can be geographically mapped, members of the clusters listed in Table 3 can be shown in a map. Figure 3 displays Indonesia's map of COVID-19 confirmed cases as per 8th November 2020, based on K-means clustering method. Each cluster is represented by a different colour.

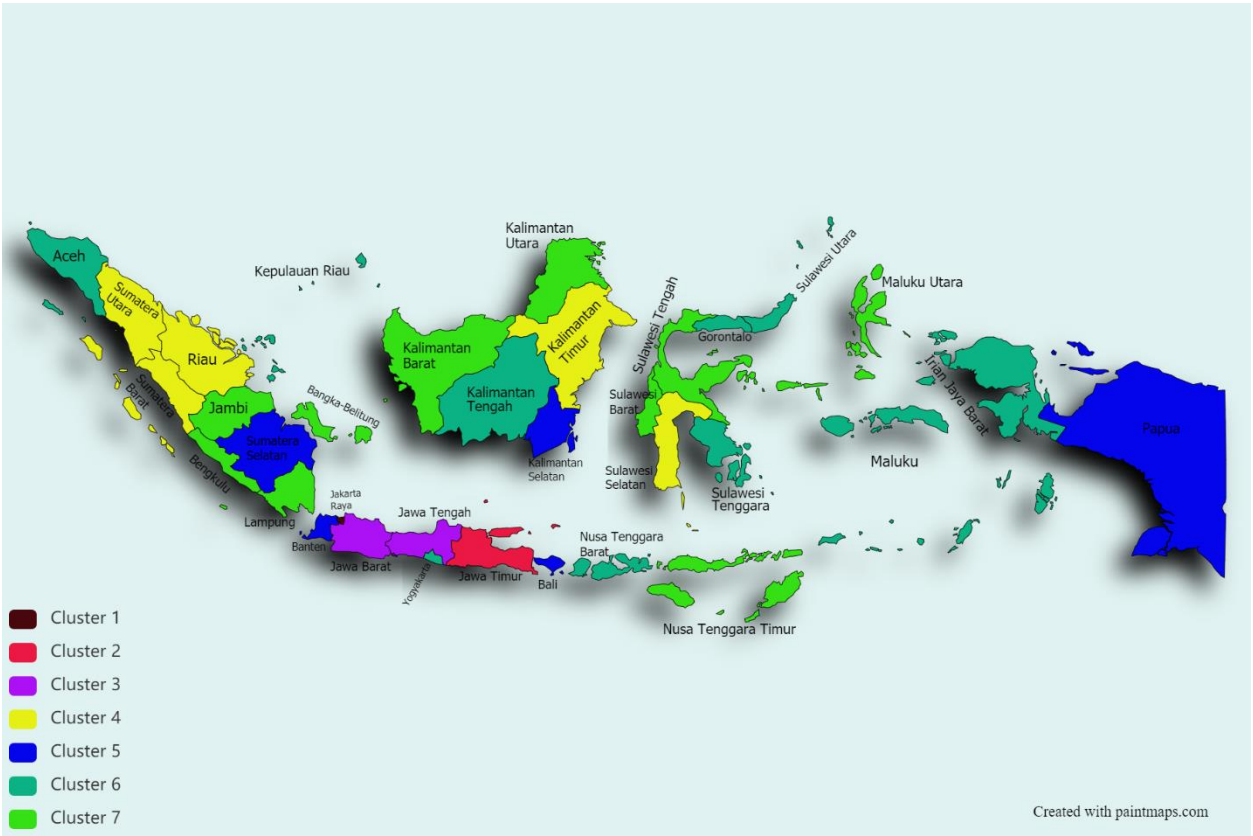

Figure 3. Indonesia's map of COVID-19 cases as per 8th November 2020, based on K-Means clustering

5. Conclusion and Future Works

This paper was intended to map COVID-19 cases in Indonesia's provinces using K-means clustering method. It is an initial attempt in order to inform the public and raise the awareness of the disease spread. It is hoped that this study may assist towards optimal handling of the pandemic in Indonesia.

As a future work, the most immediate would be to add other clustering methods to the analysis and the results could be compared. As the pandemic has started since early 2020, time series clustering could also be considered.

References

Ali Ridho Barakbah, & Kohei Aarai. (2004). Determining Constraints of Moving Variance to Find Global Optimum and Make Automatic Clustering. *Conference: Industrial Electronics Seminar (IES)*, 409–413. Retrieved from https://www.researchgate.net/publication/274137730_Determining_Constraints_of_Moving_Variance_to_Find_Global_Optimum_and_Make_Automatic_Clustering

Berkhin, P. (2006). *Survey of Clustering Data Mining Techniques*.

Dwitri, N., Tampubolon, J. A., Prayoga, S., Zer, F. I. R. ., & Hartama, D. (2020). Penerapan Algoritma K-Means Dalam Menentukan Tingkat Penyebaran Pandemi COVID-19 di Indonesia. *JurTI (Jurnal Teknologi Informasi)*, *4*(1), 128–132. https://doi.org/10.36294/JURTI.V4I1.1266

Johnson, R. A., & Wichern, D. W. (2007). *Sixth Edition Applied Multivariate Statistical Analysis*.

Kementrian Kesehatan Republik Indonesia. (2020, October). Infeksi Emerging Kementerian Kesehatan RI. Retrieved March 5, 2021, from Infeksiemerging website: https://infeksiemerging.kemkes.go.id/

Ma, E. W. M., & Chow, T. W. S. (2004). A new shifting grid clustering algorithm. *Pattern Recognition*, *37*(3), 503–514. https://doi.org/10.1016/j.patcog.2003.08.014

Mahmoudi, M. R., Baleanu, D., Mansor, Z., Tuan, B. A., & Pho, K. H. (2020). Fuzzy clustering method to compare the spread rate of Covid-19 in the high risks countries. *Chaos, Solitons and Fractals*, *140*. https://doi.org/10.1016/j.chaos.2020.110230

Mattjik, A. A., & Sumertajaya, I. M. (2011). *Sidik Peubah Ganda*. Bogor: IPB Press.

Nugroho, S. (2008). *Statistik Multivariat Terapan*. Bengkulu: UNIB Press.

Our World in Data. (2020, November). Mortality Risk of COVID-19 . Retrieved March 5, 2021, from Our World in Data website: https://ourworldindata.org/mortality-risk-covid

Ray, S., & Turi, R. H. (2000). *Determination of Number of Clusters in K-Means Clustering and Application in Colour Image Segmentation* (pp. 137–143). pp. 137–143. Retrieved from https://research.monash.edu/en/publications/determination-of-number-of-clusters-in-k-means-clustering-and-app

Scientific Brief. (2020, August). Estimating mortality from COVID-19. Retrieved March 5, 2021, from World Health Organization website: https://www.who.int/news-room/commentaries/detail/estimating-mortality-from-covid-19

Solichin, A., & Khairunnisa, K. (2020). Klasterisasi Persebaran Virus Corona (Covid-19) Di DKI Jakarta Menggunakan Metode K-Means. *Fountain of Informatics Journal*, *5*(2), 52. https://doi.org/10.21111/fij.v5i2.4905

Zarikas, V., Poulopoulos, S. G., Gareiou, Z., & Zervas, E. (2020). Clustering analysis of countries using the COVID-19 cases dataset. *Data in Brief*, *31*, 105787. https://doi.org/10.1016/j.dib.2020.105787

Biographies

Fitria Virgantari is a senior lecturer, and Head of Department of Mathematics in Faculty of Mathematics and Science, Pakuan University, Bogor, Indonesia. She earned Ir. in Statistics, Masters of Science in Statistics, and PhD in Agricultural Economy, all from IPB University, Bogor, Indonesia. She has published journal and conference papers. Her research interests include clustering, data development, and statistical modelling. She is member of IndoMS and IORA.

Yasmin E. Faridhan is a lecturer in Department of Mathematics in Faculty of Mathematics and Science, Pakuan University, Bogor, Indonesia. Ms. Faridhan holds a Bachelor of Science degree in Statistics and a Master of Science degree in Statistics, both from IPB University, Bogor, Indonesia. She has published journal and

VOLUME 5 | NUMBER 2 | NOVEMBER 2020
ISSN: 2656-1174 (online)

conference papers. Her research interests include classification and decision tree methods, as well as survey sampling.