Int. J. Nonlinear Anal. Appl. 13 (2022) No. 1, 2115–2126
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2022.5907

Predicting of infected People with Corona virus (covid-19) by using Non-Parametric Quality Control Charts

Heba Mostafa Fawzy[a,*], Asmaa Ghalib[a]

[a]Department of statistics, college of administration and Economics, University of Baghdad, Iraq

(Communicated by Madjid Eshaghi Gordji)

Abstract

Quality control Charts were used to monitor the number of infections with the emerging corona virus (Covid-19) for the purpose of predicting the extent of the disease's control, knowing the extent of its spread, and determining the injuries if they were within or outside the limits of the control charts. The research aims to use each of the control chart of the (Kernel Principal Component Analysis Control Chart) and (K- Nearest Neighbor Control Chart). As (18) variables representing the governorates of Iraq were used, depending on the daily epidemiological position of the Public Health Department of the Iraqi Ministry of Health. To compare the performance of the charts, a measure of average length of run was adopted, as the results showed that the number of infection with the new Corona virus is out of control, and that the (KNN) chart had better performance in the short term with a relative equality in the performance of the two charts in the medium and long rang

Keywords: Quality Control, Control Charts, Average Length of Range, Nearest Neighborhood, Kernel Principal Components Analysis.

1. Introduction & research purpose

Since the beginning of life on the surface of the earth, man has sought, by various means, to monitor the diseases and epidemics that accompanied its appearance to reduce the risk of infection and limit its spread. Many statistical analysts and those interested in the field of health have continued to

*Corresponding author
Email addresses: almaroofhiba@gmail.com (Heba Mostafa Fawzy), dr.asmaa.ghalib@coadec.uobaghdad.edu.iq (Asmaa Ghalib)

Received: May 2021 *Accepted:* October 2021

introduce statistical methods to monitor the spread of diseases, including quality control, as the use of control charts is no longer limited to the industrial field only, but rather includes different areas. Recently it was used on some transitional epidemics that have effects dangerous to human life in particular and the rest of the organisms in general for the purpose of observing the extent of the spread and transmission of diseases between neighboring areas. The research aims to review the multivariate parametric control charts to determine the limits of control for observations of infection with the emerging coronavirus (Covid-19) and to compare the performance of the two charts using the average run length measure.

Hotelling [4] wrote a book on qualitative control in (1947), which is one of the basic references for research and studies specialized in this subject, in which he relied on the theory of multivariate distributions to clarify the statistical methods used in qualitative control in the case of multiple variables, and in (2012) (Walid, Mohamed [5]) used the (K^2-CHART) board on one of the industrial applications and concluded, despite the presence of some problems, that the control process was completed successfully, and that the(K^2 -CHART) revealed small changes in the mean vector. In the year 2020, (Muhammad and Hidayatul [1]) used the kernel PCA control chart that depends on the kernel principal compounds to monitor the qualitative characteristics of the mixed variables and concluded the good performance of the chart in discovering the out-of-control observations.

2. Theoretical side

2.1. Statistical Process Control (SPC)

Statistical control operations are one of the tools and techniques to control production processes, used to improve the quality of production and ensure the availability of the required qualities in the product, examine samples and analyze the capacity of the process ... etc. Therefore, statistical qualitative control is one of the important means in the production process [9].

2.2. Quality Control Chart

The qualitative control panel is considered a statistical means to determine whether the data of a process is subject to ordinary or unusual deviations (ordinary deviations are meant to work under the established limits of control, either unusual indicates out of control and must be reconsidered). In the past ten years, control charts have been applied to improve the quality of production processes to meet the needs and desires of the consumer. The use of control charts serves as an indicator or early warning that reveals the processes outside of control for the purpose of maintaining the continuity of the production process within the limits of control [9]

The control charts usually consist of three limits as follows:

1. Lower control limit (LCL): It represents the least acceptable percentage of defective units.
2. Upper control limit (UCL): It represents the highest percentage acceptable percentage of defective units.
3. Central Limit (CL): It represents the optimum level of production quality

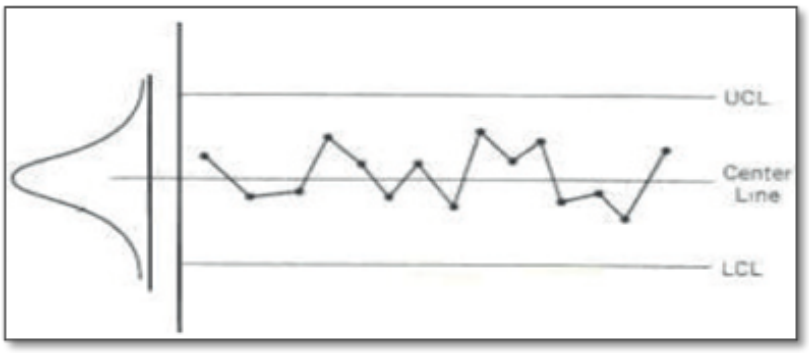

Figure 1: represents the limits of the control charts

2.3. Multivariate Quality Control

The multivariate control charts are used to discover the change in the process in the shortest possible time after the change, which required updating the control charts to monitor or check several variables together instead of examining each characteristic separately [4].
In 1947, (Hotelling) was the first who develop control charts to be used in multivariate processes and explained the methods used in controlling in cases that contain several variables [2]

2.4. Non-Parametric Multivariate Control Charts

The control charts are often based on the assumption that there is a specific form of the parameter distribution, such as the normal distribution, and it is called the parameter control charts used in many control applications, but there is not enough information or evidence to achieve this assumption, so the performance of many charts parametric control may change in these cases, as it leads to the emergence of shady results and false alarms. To avoided of this problem, we resort to control charts that do not require the assumption of a specific distribution to monitor the production process, and they are called non-parametric control charts or free distribution, as the control charts are a better alternative to the science. You need to know the basic distribution of the production process as it is less affected by outliers [1]

2.4.1. Kernel Principal Component Analysis Control Chart (KPCA)

The analysis of kernel principal component control charts based on
the calculation of Hotelling's statistic (T^2) for the matrix of the kernel principal compounds, which is calculated by applying the algorithm of the analysis of the kernel principal compounds on the variance and covariance matrix to observe the characteristics and properties and discover the external factors influencing depending on the control chart limits. To implement the control chart (Kernel PCA),
we first choose the kernel function, which depends in its calculation on the smoothing parameter (h) or bandwidth, and then we calculate the kernel matrix according to the following general formula [1]:

$$K{=}K_{ij}{=}\Omega\left(x_i\right)\Omega\left(x_j\right) \qquad i=1,\ldots,n, \qquad j=1,\ldots,n \tag{2.1}$$

After determine both the kernel function and the kernel matrix, the kernel components are calculated from the following formula:

$$p_\omega{=}\sum_{j=1}^{n}\alpha_{ij}K\ \ (x_i,x_j) \tag{2.2}$$

and from the first principal compound (l) of p, the Hotelling statistic (T^2) is calculated according to the following formula:

$$\tilde{T}_K^2 = \sum_{i=1}^{l} p_i \lambda_i^{-1} p_i^T \tag{2.3}$$

Where: i=1,....I and λ_i the i-th characteristic roots
The control limits for this chart is determined based on the kernel density estimator, because the distribution of the statistic ($\tilde{T}_K^2$) is unknown and can be expressed through the following formula[11]:

$$\hat{f}_h = \frac{1}{n\hat{h}} \sum_{i=1}^{n} K\left(\frac{T^2 - \tilde{T}_{K?i}^2}{\hat{h}}\right) \tag{2.4}$$

Where:
K: kernel function
$\hat{h}$: smoothing parameter
Also, the distribution of the function $\hat{f}_h$, which is symbolized by the symbol $\hat{F}_h$, can be calculated as follows:

$$\hat{F}_h\left(\tilde{T}_K^2\right) = \int_0^{\tilde{t}_K^2} \hat{f}_h(\tilde{T}_K^2) d(\tilde{T}_K^2) \tag{2.5}$$

We Know $\hat{F}_h\left(\tilde{T}_K^2\right)$ is calculated according to the trapezoid rule, which is one of the numerical integration methods, depends on dividing the integration period into points that are set in advance for the purpose of obtaining approximate mathematical formulas for calculating the integrals as follows:

$$\int_{p_{min}}^{p_{max}} \hat{f}_h(\tilde{T}_K^2) d(\tilde{T}_K^2) \approx \frac{\pi_{max} - \pi_{min}}{2n} \sum_{i=1}^{n} (\hat{f}_h(\tilde{T}_{K,i}^2) + \hat{f}_h(\tilde{T}_{K,(i+1)}^2)) \tag{2.6}$$

Where: π_{max}, π_{min} represents the largest and smallest value of ($\tilde{T}_K^2$), so the upper control limit for this chart is determined by determining the percentage value ($100(1-a)^{th}$), and the false alarm average α, whose value ranges between ($1<\alpha<0$) is calculated according to the following formula [13]:

$$CL = \hat{F}_h(\tilde{t}_k^2)^{-1}(1-\alpha) \tag{2.7}$$

The kernel function adopted in this paper is the (Gaussian kernel) function, whose formula is:

$$k(x_i x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2h^2}\right) \tag{2.8}$$

2.4.2. KPCA Algorithm
The first step: choose the kernel function and then calculate the kernel matrix
The second step: Calculate the kernel principal compounds from Equation (2.2)

$$p_\omega = \sum_{j=1}^{n} \alpha_{ij} K\ (x_i, x_j)$$

Step 3:calculate the statistic $\tilde{T}_K^2$ of the first (l) principal component of p from equation (2.3) i.e.:

$$\tilde{T}_K^2 = \sum_{i=1}^{l} p_i \lambda_i^{-1} p_i^T$$

Step 4: Calculate $\hat{f}_h$ which is the estimator of the density for $\tilde{T}_K^2$
Step 5: Calculate $\hat{F}_h\left(\tilde{T}_K^2\right)$ and then determine the control limits from Equation (2.7), i.e.:

$$CL = \hat{F}_h(\tilde{t}_k^2)^{-1}(1-\alpha)$$

Step6: Draw the chart according to $\tilde{T}_K^2$ and CL .
CL the upper control and a process is considered out of control if it exceeds a statistic value $\tilde{T}_K^2$

2.4.3. K-Nearest Neighbor Control Chart ($\boldsymbol{K^2-CHART}$*)*
The control panel ($\boldsymbol{K^2}$) is based on the nearest neighborhood (K) algorithm, which is one of the non-parametric classification methods that work with unsuprvisor [14], which was used in 1960. KNN solves the classification problem based on local density estimation data using the nearest neighbor algorithm [10]. The KNN algorithm is considered one of the simple methods with direct application, as it only needs a measure to calculate the distance between the sample and the sample closest to it, a set of training samples (used to train a model that helps to get to the nearest neighbor), and the integer (k) which represents the number of the neighbor. The closest [8], and it is determined either by experiment or by calculating the model error for each (k) and then choosing the value with the least error. This chart is characterized by its ability to deal with multi-variable, high-dimensional data [12].
Suppose $NN_i(x)$ represents the training view of the nearest ith adjacent to the point x to be classified. The local density d(x) is calculated for the data point x as follows [5]

$$d(x) = \frac{\frac{i}{N}}{V\left(\|x - NN_i(x)\|\right)} \tag{2.9}$$

Where: V is the size of the field containing the i of the nearest neighbor training views.
N: The size of the training data set
In the same way we calculate the local density of $(NN_i\ (x))$

$$d\left(NN_i(x)\right) = \frac{\frac{i}{N}}{V\left(\|NN_i(x) - NN_i\left(NN_i(x)\right)\|\right)} \tag{2.10}$$

So $NN_i\ (NN_i\ (x))$: It is the training view closest to ith adjacent to $NN_i\ (NN_i\ (x))$ in the same training data set.
(KNN) calculates the average distances K by considering ((i = 1,. . . .,K) by taking the ratio of the local density d(x) to the local density $d\left(NN_i(x)\right)$ which must be greater or equal to one:

$$\frac{d(x)}{d\left(NN_i(x)\right)} = \frac{\sum_{i=1}^{k} \|NN_i(x) - NN_i\left(NN_i(x)\right)\|}{\sum_{i=1}^{k} \|x - NN_i(x)\|} = 1 \tag{2.11}$$

The K^2 statistic that depends on the KNN algorithm is calculated from the following formula [6]

$$k^2=\frac{\sum_{i=1}^{k}\|x-NN_i(x)\|}{k} \tag{2.12}$$

Where: K^2 it represents the average distance between the observation x and the training observations of the nearest neighborhood, and k^2 values are used as statistics in quality control processes. The control limits for this chart are determined based on the value estimated by the (Bootstrap) method, which is one of the sampling methods that are used to re-sampling a set of random samples from the original sample data with return and the same size as the original sample, to calculate the upper control limit we draw the B number of random samples (Bootstrap) whose observations represent the values of K^2statistics

$$k^2{}_{1j}, k^2{}_{2j}, \ldots, k^2{}_{Nj} \qquad j=1,\ldots,B \tag{2.13}$$

Then we determine the value of the percentage $(100(1-a)^{th})$ and the false alarm rate α, whose value ranges between $(1<\alpha<0)$ for each sample (Bootstrap), and therefore the upper control limit is calculated from the following formula [15].

$$CL=\sum_{j=1}^{M}k_{ij}^2/B \tag{2.14}$$

2.4.4. k^2-Chart Algorithm

-Step 1: Determine the k parameters of the KNN algorithm from the set of training samples.

- Step 2: Calculate the values of the statistic $k^2$from equation (2.12)

$$k^2=\frac{\sum_{i=1}^{k}\|x-NN_i(x)\|}{k}$$

- Step 3: Established the upper limit of control on the basis of the general average of k^2 values and as a percentage of $(100(1-a)^{th})$ for the bootstrap samples that number B from Equation (2.14):

$$CL=\sum_{j=1}^{M}k_{ij}^2/B$$

- Step 4: Draw the chart according to k^2 and CL .

CL the upper control and a process is considered out of control if it exceeds a statistic value k^2

2.5. Average Run Length(ARL)

The average length of the range is one of the commonly used measures to compare the performance of multivariate qualitative control charts. It is also known as the number of samples that must be determined before the process is registered as out of control [7] and the average run length is used to detect whether the process suffers from deviations, and it is related to the probability of the occurrence of a type-I error (α). It can be defined as the false alarm rate, as it gives an indication that the process is out of control [3].

The average run length is calculated from the following formula:

$$ARL=\frac{1}{\alpha} \tag{2.15}$$

So: α represent type 1 false alarm rate [3].

The ARL value is compared with its approximate value, which is calculated as the ratio of the number of observations out of control to the total number of observations.

3. Application side

The two control charts (KPCA-Chart) and (K^2-Chart) were applied to the data on the numbers of coronavirus infection taken from the official website of the Public Health Department of the Iraqi Ministry of Health, based on the daily epidemiological situation for a period of 91 days from 01/21/21. 15 to 4/14/2021 for all governorates of Iraq, represented by (18) governorates.

3.1. Coronavirus Brief

Corona virus (covid-19) belongs to the family of coronaviruses that originate in animals and then are transmitted to humans. The new corona virus (covid-19) appeared in the Chinese city of Wuhan in the year (2019) and then spread widely to cover infections all over the world. The virus is transmitted through personal contact with an infected person or upon contact with contaminated surfaces because the virus can live on them for some hours. Symptoms of the Corona virus vary according to the degree and period of infection, symptoms may appear two days or 14 days after infection and include (fever - diarrhea - fatigue - loss of taste and smell - headache - muscle pain ...). Some cases show serious symptoms of the virus such as (shortness of breath - chest pain - inability to move or speak) [16].

3.2. Application Non-Parametric Control Charts

The multivariate non-parametric control charts were drawn and determined their control limits for the following levels of significance (0.5, 0.1, 0.05, 0.01, 0.005) to choose the best among them to control the numbers of infection with the Corona pandemic (covid-19)

1. Control charts at the significant level of 0.5

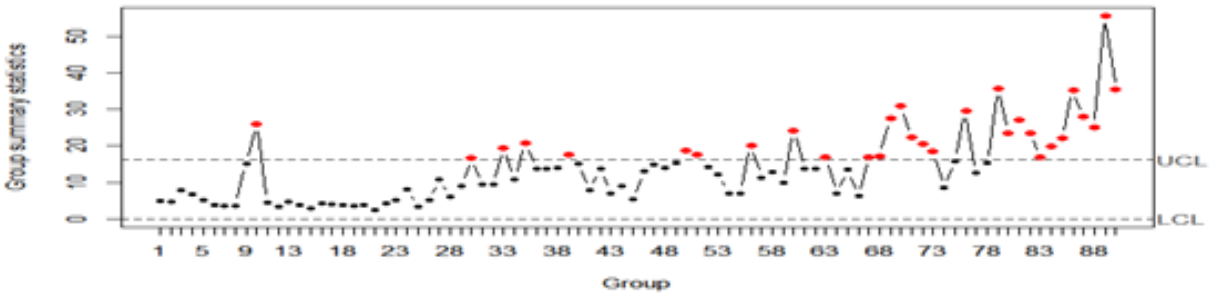

Figure 2: Control KPCA-Chart at the significant level of 0.5

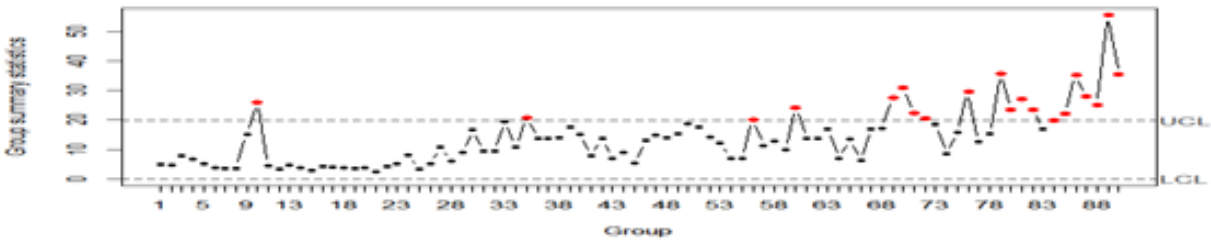

Figure 3: Control $K^2-Chart$ at the significant level of 0.5

We see from Figure 2 and Figure 3 that some values are out of control, meaning that the monitoring process is out of control at the level (0.5)

2. Control charts at the significant level 0.1

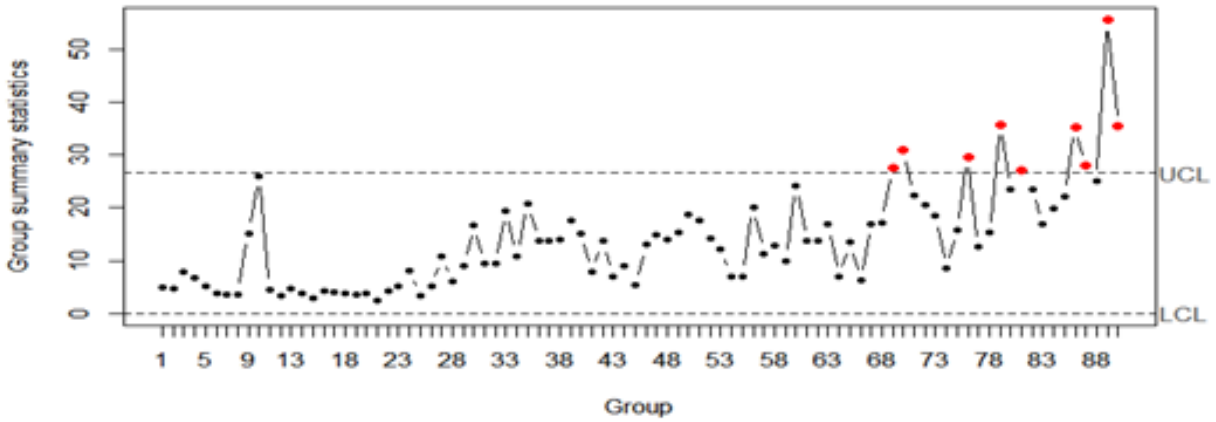

Figure 4: Control KPCA-Chart at the significant level 0.1

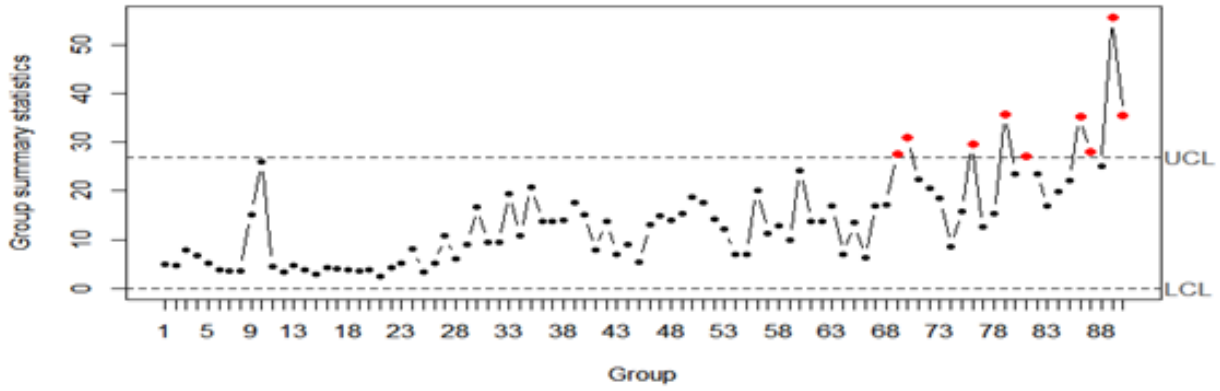

Figure 5: Control $K^2-Chart$ at the significant level of 0.1

We find that in (KPCA-Chart)) nine observations are outside the upper limit of control, meaning that the examination process was out of control, and $(K^2-Chart)$ indicates as well that the process is out of control in the same significant level (0.1) .

3. Control charts at the significant level of 0.05

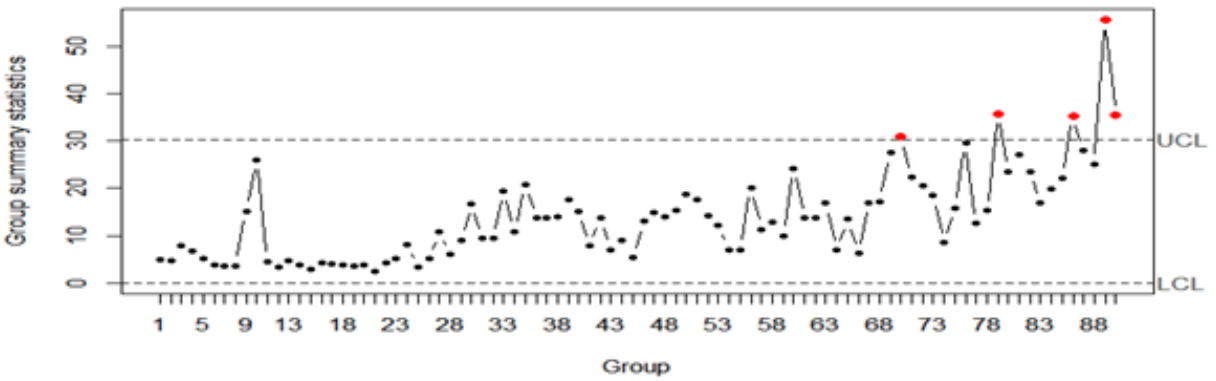

Figure 6: Control KPCA-Chart at the significant level 0.05

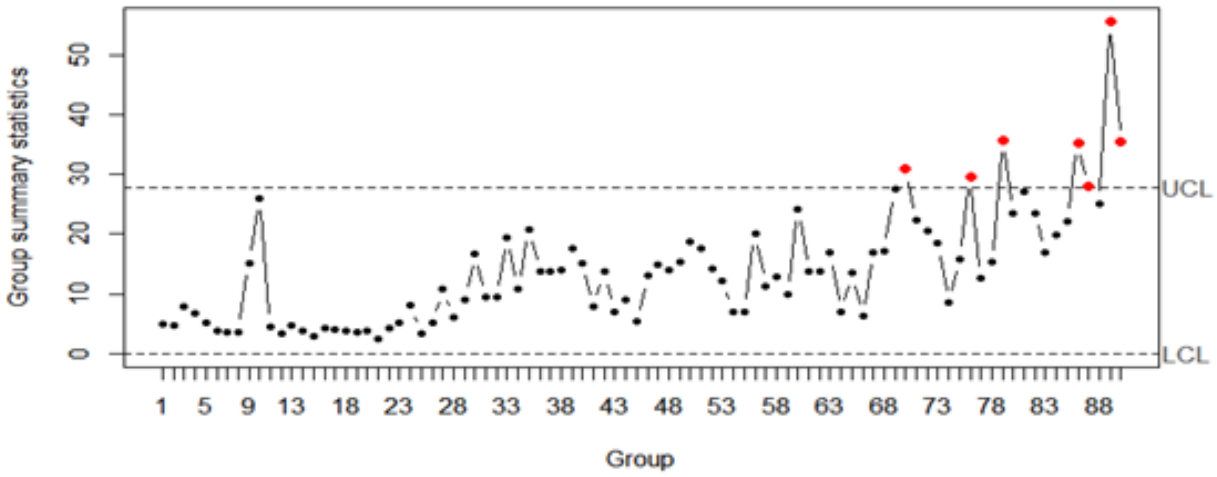

Figure 7: Control $K^2-Chart$ at the significant level of 0.05

From Figure 6 (KPCA-Chart), it is clear that the examination process is still out of control at this level, as there are not all observations within the upper limit of control, as well as for the K^2-Chart we note the exit of seven observations that is mean the process is out of control at this level also in both charts.

4. Control charts at the significant level of 0.01

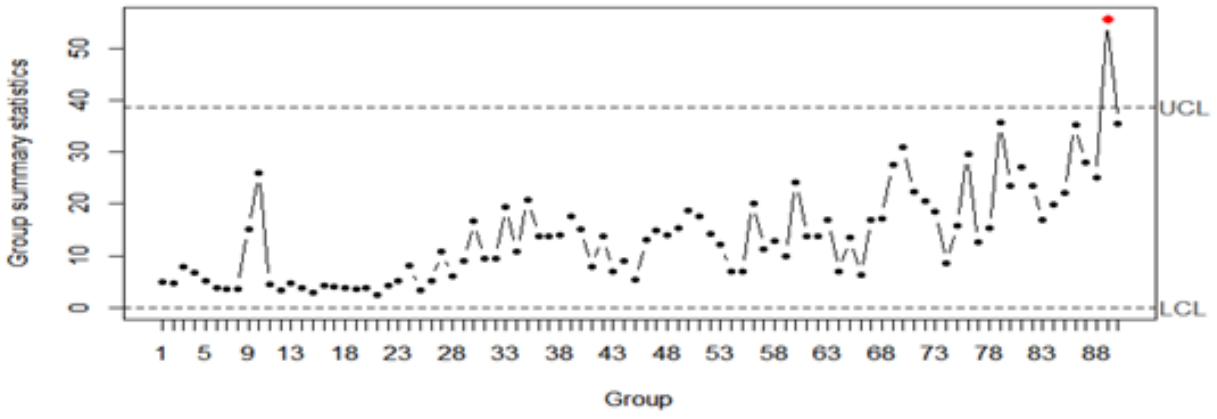

Figure 8: Control KPCA-Chart at the significant level 0.01

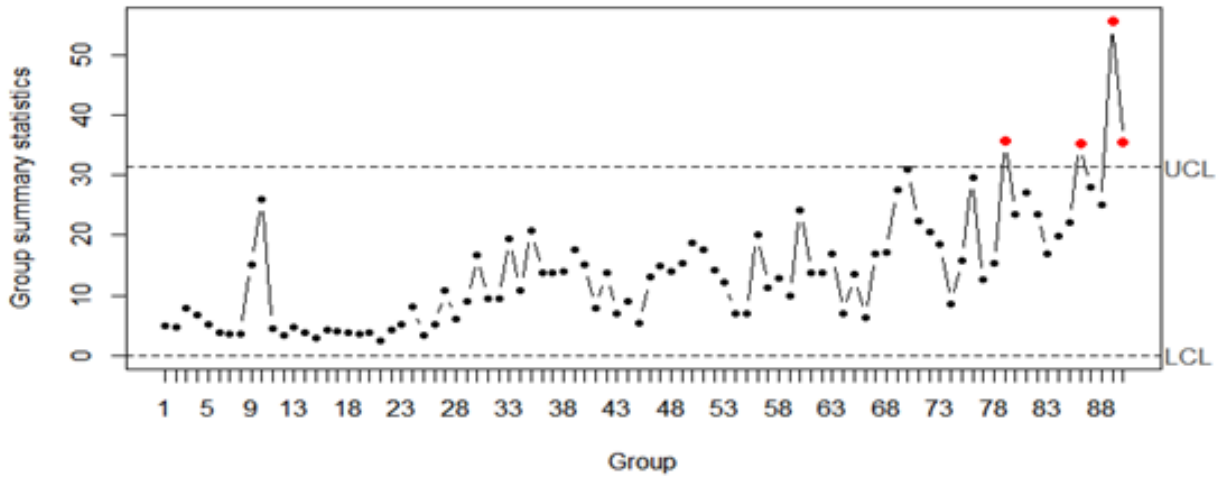

Figure 9: Control $K^2-Chart$ at the significant level of 0.01

From Figure 8 and Figure 9, we see that some of the observations went out of the upper limit of control, and this means that the monitoring process for the preparation of the Corona virus was out of control for both charts at this level.

5. Control charts at the significant level of 0.005

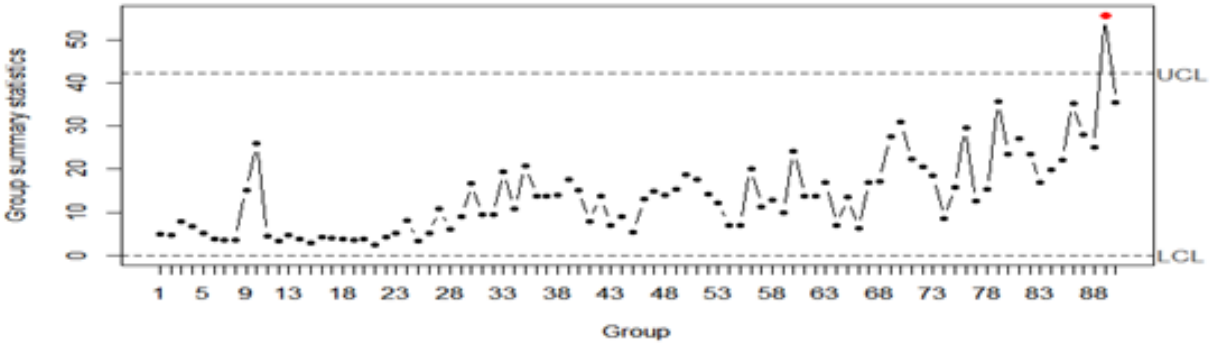

Figure 10: Control KPCA-Chart at the significant level 0.005

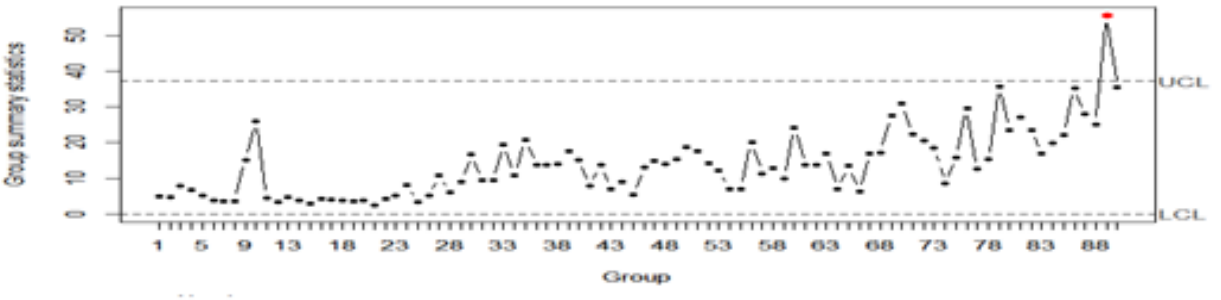

Figure 11: Control $K^2-Chart$ at the significant level of 0.005

We find that both KPCA-Chart and K^2-CHART have detected one observation out of control at this level, meaning that the monitoring process was also out of control.

Table 1: represents the values of the measure ARL For control charts at different significant levels

Alpha	KPCA	K^2
0. 5	1.486	1.611
0. 1	8.194	8.975
0.05	18.4	18.317
0.01	98.889	95.556
0.005	197.778	197.778

From the above table, we see that when ($\alpha = 0.5$), the closest value of ARL was (1.6111) for the ($K^2 - Chart$) followed by KPCA-Chart as it reached (1.486). Also, the value of K^2- chart approached at the level of significance ($\alpha = 0.1$), followed by the KPCA-Chart at ($\alpha = 0.05$), it is the closest equal to the value of ARL in both charts. At level 0.01, the closest ARL was according to the KPCA-Chart method, followed by (K^2). Also, when the value of ($\alpha = 0.005$), the value of the ARL was equal on both charts with a value of 197.778. We see that the performance of the ($K^2 - Chart$) is better at both the level (0.5) and (0.1), and we see that the performance of both charts is equal at the level (0.05), but at the level (0.01), the performance of the KPCA-Chart was good compared to the ($K^2 - Chart$), while both charts performed well at the level of significance (0.005)

4. Conclusions and recommendations

4.1. Conclusions

1. We conclude that the process of controlling the epidemic was completely out of control.
2. The ability of the two charts used to discover the observations that exceeded the upper limit of control at all levels significant
3. The performance of both charts was good in monitoring process deviations.

4.2. Recommendations

1. The use of non-parametric multivariate control charts with data that are not normally distributed to detect deviations in the process progress
2. Using other multivariate-parametric charts

References

[1] M. Ahsan, M. Mashuri, H. Khusna and M. H. Lee, *Multivariate Control Chart Based on Kernel PCA for Monitoring Mixed Variable and Attribute Quality Characteristics*, Symmetry, 12 (11) (2020) 1838 .

[2] M. E. Camargo, A. I. d. S. Dullius, W. P. Filho, S. L. Russo, M. R. Cruz, A. Galelli, G. F. da Silva , *Multivariate quality control basead on discriminant analysis-ajusted variables*, Aust. J. Basic Appl. Sci., 6 (1) (2012) 207-212.

[3] N. Das *Non-parametric Control Chart for Controlling Variability Based on Rank Test*, Econ. Qual. Control, 23 (2)(2008) 227-242 .

[4] M. Frisén, *On multivariate control charts*, Produção, 21(2)(2011) 235-241.

[5] W. Gani, M. Limam, *Performance Evaluation of One-Class Classification-based Control Charts through anIndustrial Application*, Qual. Reliab. Eng., 29(16) (2012) 841-854.

[6] W. Gani, M. Limam,*A One-Class Classification-Based Control Chart Using the ??-Means Data Description Algorithm*, J. Qual. Reliab. Eng., 2014 (2014) 1-9 .

[7] G. Han, K. M. B. Chong, *A Study on the Median Run Length Performance of the Run Sum S Control Chart*, Int. J. Mech. Eng. Rob. Res., 8 (6) (2019) 885-889.

[8] Q. P. He, J. Wang, *Fault Detection Using the k-Nearest Neighbor Rule for Semiconductor Manufacturing Processes*,IEEE Trans. Semicond. Manuf. , 20 (4) (2007) 345-354.

[9] A. GH. Jaber and F. H. Enad, *The Using of Multivariate Parametric Hottling -T^2 and Non-Parametric Bootstrap Charts in Quality Control Using Simulation*, Muthanna Journal of Administrative and Economic Sciences , 10 (3)(2020) 8-22.

[10] S. Kazemi and S. Niaki , *Monitoring image-based processes using a PCA-based control chart and a classification technique*, Decis. Sci. Lett., 10 (1)(2020) 39-52 .

[11] H. Kuswanto, M. Ahsan,*Multivariate control chart based on PCA mix for variable and attribute quality characteristics*, Prod. Manuf. Res., 6(1) (2018) 364-384 .

[12] W. Li, C. Zhang, *Nonparametric monitoring of multivariate data via KNN learning*,Int. J. Prod. Res. ,2 (2020) 1-16 .

[13] M. Mashuri, H. Haryono, *Tr(R2) control charts based on kernel density estimation for monitoring multivariate variability process*, Cogent Eng., 6 (1) (2019) 1-37 .

[14] G. Verdier and A. Ferreira , *Adaptive Mahalanobis Distance and k-Nearest Neighbor Rule for Fault Detection in Semiconductor Manufacturing Adaptive Mahalanobis Distance and k-Nearest Neighbor Rule for Fault Detection in Semiconductor Manufacturing* , IEEE, 24(1) (2009) 1-21 .

[15] T. Sukchotrat, S. B. Kim and F. Tsung, *One-class classification-based control charts for multivariate process monitoring*, IIE Trans. , 42 (2) (2010) 107- 120.

[16] `https://www.mayoclinic.org/ar/diseases-conditions/coronavirus/symptoms-causes/syc-20479963`.