INGENIERÍAS USBMed

https://revistas.usb.edu.co/index.php/ingUSBmed

ARTÍCULO DE INVESTIGACIÓN

Un *boxplot* bidimensional en coordenadas polares

A Two-dimensional Boxplot in Polar Coordinates

Carlos Gaviria[1], Carlos Márquez[2], César Guerra Villa[3], Walter Acevedo Nanclares[4]

[1]Ciencias Básicas, Facultad de Ingenierías, Universidad de San Buenaventura, Medellín, Colombia. Email: carlos.gaviria@usbmed.edu.co

[2]Ciencias Básicas, Facultad de Ingenierías, Universidad de San Buenaventura, Medellín, Colombia. Email: carlos.marquez@usbmed.edu.co

[3] Facultad de Ciencias Empresariales, Universidad de San Buenaventura, Medellín, Colombia. Email: cesar.guerra@tau.usbmed.edu.co

[4]Ciencias Básicas, Facultad de Ingenierías, Universidad de San Buenaventura, Medellín, Colombia. Email: walter.acevedo@tau.usbmed.edu.co

Tipo de Artículo: Investigación científica y tecnológica.
Recibido: 01-06-2019.
Revisado: 25-07-2019.
Aprobado: 01-08-2019.
Doi: 10.21500/20275846.4218

Referenciar así: C. Gaviria, C. Márquez, C. Guerra & W. Acevedo. "Un *boxplot* bidimensional en coordenadas polares". *Ingenierías USBMed*, 10(2), pp.2-7, 2019.

Declaración de disponibilidad de datos: Todos los datos relevantes están dentro del artículo, así como los archivos de soporte de información.

Conflicto de intereses: los autores han declarado que no existen conflicto de intereses.

Editores: Yohana López Rivera, Universidad de San Buenaventura, Medellín, Colombia. Alfonso Insuasti Rodríguez, Universidad de San Buenaventura, Medellín, Colombia. Erika Solange Imbett Vargas, Instituto Tecnológico Metropolitano. Eliana Zapata Ruiz, Instituto Tecnológico Metropolitano. José Fernando Valencia Grajales, Universidad Autónoma Latinoamericana.

Resumen. En este artículo realizamos una propuesta de un *boxplot* bidimensional, una extensión del *boxplot* univariado de Tukey. Este gráfico está conformado por dos polígonos convexos que están orientados en la dirección de una linea recta ajustada, llamada "línea de Tukey". El *boxplot* bidimensional tiene una caja interior que contiene el 50% de los datos; un punto interior a dicha caja que denotamos como la "mediana" y una caja externa que separa los valores atípicos. El *boxplot* bidimensional representa ubicación, extensión, correlación y asimetría de los datos.
Palabras Clave. *Boxplot* bidimensional, línea de Tukey, líneas cuartiles, vallas, cajas interior y exterior.

Abstract. This article was a proposal of a two-dimensional boxplot, an extension of the univariate Tukey boxplot. This graph consists of two convex polygons, which are oriented in the direction of an adjusted straight line, called "Tukey line." The two-dimensional boxplot has an inner box, which contains 50% of the data; a point inside the box, called "median" and an external box, which separates the outliers. The two-dimensional boxplot represents location, extension, correlation, and asymmetry of data.
Keywords. Two-dimensional *Boxplot*, Tukey line, quartile lines, fences, and inner and outer boxes

I. Introducción

En la historia se han estudiado diversos gráficos bivariados, cada uno de ellos con la intención de representar de manera adecuada las características asociadas a un conjunto de datos proveniente de dos variables aleatorias continuas. Gráficos de esta naturaleza son importantes porque, en primer lugar, permiten hacer extensiones de gráficos univariados –lo cual representa una ganancia en términos teóricos–; en segundo lugar, permiten captar la esencia de la relación entre dos variables aleatorias a simple vista, cuestión que responde a las exigencias que se hace en la estadística aplicada [1, 2]; y en tercer lugar son útiles en aplicaciones–a la ingeniería, por ejemplo–. Un gráfico bivariado particular es el *boxplot* (diagrama de caja) en dos dimensiones, el cual es una extensión del *boxplot* univariado de Tukey.

Backetti y Gould [3] propusieron un *rangefinder boxplot*. Luego, Lenth (1988) sugirió una versión mejorada. Goldberg y Iglewicz [4] propusieron otro tipo de diagramas de caja: un Relplot y Quelplot; Hyndman (1996) hizo lo propio con una gráfica de regiones de más alta densidad; Rousseeuw, Ruts y Tukey [5] presentaron otra extensión bivariada del *boxplot* univariado, una *bagplot*. Tongkumchum [6] propuso una extensión del *boxplot* univariado de Tukey.

El objetivo en el presente trabajo es construir un *boxplot* bidimensional que tenga alguna diferencia frente a aquellos que ya cuentan con trayectoria, como es el caso de los citados en el párrafo anterior. Inicialmente se hará la construcción del *boxplot* bidimensional, para lo cuál se hablará en primer lugar de la línea de Tukey que le dará orientación a las cajas interior y exterior de la *boxplot* bidimensional. En segundo lugar, se definirán las líneas cuartiles y las vallas en términos de coordenadas polares, y con base en ellas se construirán las cajas interior y exterior; nuestro *boxplot* así construido es más simple que el *bagplot* por su construcción y más general que el *boxplot* bidimensional propuesto en [6], pues el nuestro contempla polígonos convexos que no son paralelogramos. Por último, se presentará un ejemplo que muestra nuestra construcción.

En la siguiente sección se mostrarán los elementos básicos relacionados con el *boxplot* Univariado de Tukey y los *boxplot* bidimensionales más representativos que se han construido hasta el momento. Todo esto es necesario para hacer una propuesta de un *boxplot* en el caso de 2 variables aleatorias continuas.

A. *Boxplot* Univariado de Tukey

John Tukey introdujo el *boxplot* y los bigotes como parte de su conjunto de herramientas para el análisis exploratorio de datos [7], pero no llegó a ser ampliamente conocido hasta la publicación formal en 1977. En términos concretos, el diagrama de caja, o *boxplot*, es un resumen gráfico que describe varias de las características más relevantes de un conjunto de datos; estas características incluyen: el centro, la dispersión, el grado y naturaleza de cualquier alejamiento de la simetría, y la identificación de las observaciones extremas inusualmente alejadas del cuerpo principal de los datos. Como un solo valor puede afectar de forma notable a $\overline{x}$ y s, entonces una gráfica de caja está basada en medidas resistentes a la presencia de unos cuantos valores apartados, la mediana y una medida de variabilidad llamada "dispersión de los cuartos".

Para construir un diagrama de caja se ordenan las observaciones de la más pequeña a la más grande, y se separa la mitad más pequeña de la más grande; se incluye la mediana $\widetilde{x}$ en ambas mitades si n es impar. En tal caso, el cuarto inferior es la mediana de la mitad más pequeña; y el cuarto superior, la mediana de la mitad más grande. Una medida de dispersión que es persistente a los valores apartados es la dispersión de los cuartos f_s, dada por:

$$f_s = \text{cuarto superior} - \text{cuarto inferior}$$

En general, la dispersión de los cuartos no se ve afectada por las posiciones de las observaciones comprendidas entre el 25% más pequeño o el 25% más grande de los datos. Por esta razón, la dispersión es resistente a valores apartados.

La gráfica de caja más simple se basa en los números: x_i más pequeños, cuarto inferior, mediana, cuarto superior y x_i más grandes.

Para construir el diagrama de caja se procede así:

1. Trazar una escala de medición horizontal.
2. Colocar un rectángulo sobre este eje horizontal. El lado izquierdo del rectángulo está en el cuarto inferior, y el derecho en el superior. De esta manera, el ancho de la caja es f_s.
3. Colocar una línea vertical en la ubicación de la mediana. La ubicación de esta última da información sobre asimetría.
4. Trazar bigotes fuera de ambos extremos del rectángulo, hacia las observaciones más pequeñas y más grandes.

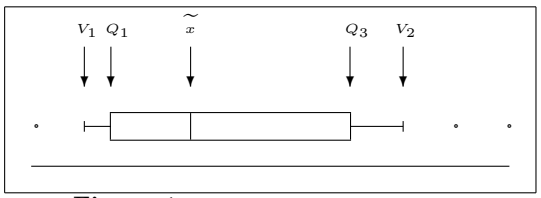

Figura 1. *Boxplot* univariado de Tukey

En la figura 1, $\widetilde{x}$ es la mediana, Q_1 es el primer cuarto, Q_3 es el tercer cuarto y V_1, V_2 son las vallas. Cualquier observación a más de $1.5f_s$ del cuarto más cercano es un valor apartado o atípico. Un valor apartado es extremo si se encuentra a más de $3f_s$ del cuarto más cercano; y moderado si ocurre lo contrario.

B. *Rangefinder boxplot*

La propuesta de Backetti y Gould [3] simplemente superpone seis segmentos de línea en el diagrama de dispersión, dos de los cuales forman una cruz. A partir de los *boxplot* univariados de las variables aleatorias X y Y, las medianas, los cuartiles y los puntos extremos de cada conjunto de bigotes se utilizan para dibujar líneas horizontales y verticales en el diagrama de dispersión. En la versión mejorada de Lenth (1988), los *boxplot* univariados de las variables X y Y se incorporan a lo largo de los ejes. Estos dos *boxplot* en dos dimensiones no muestran las relaciones bivariadas de forma y correlación de los datos.

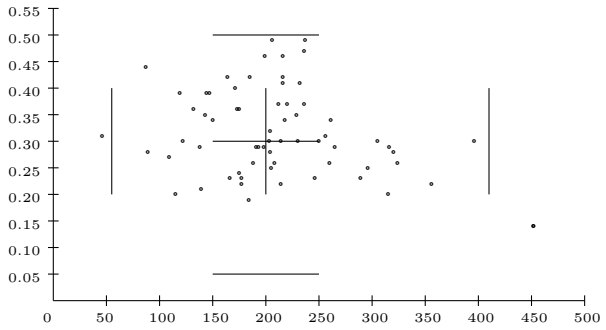

Figura 2. Rangefinder *boxplot*

C. *Relplot y quelplot*

Goldberg y Iglewicz [4] propusieron otro tipo de *boxplot* bivariado: *relplot* (gráfica elíptica robusta) y *quelplot* (gráfica de un cuarto elíptico). La primera se construyó para los conjuntos de datos que se suponen simétricos elípticamente; en este caso, las elipses se obtienen ajustando una distribución bivariada Gaussiana. Para los conjuntos de datos que no son simétricos, las elipses son reemplazadas por cuatro cuartos de elipses separados, coincidentes en sus ejes mayores y menores, basados en estimaciones robustas de localización y escala. En contraste con el *rangefinder boxplot*, estos gráficos son realmente bivariados: muestran la forma bivariada de los datos mediante un par de elipses concéntricas, o *quels* que sirven como una bisagra y una valla. La *relplot* se centra en la media, mientras que la *quel* se enfoca en el centro de la probabilidad; la región interior contiene 50% de los datos, mientras que la externa delinea los valores atípicos potenciales. Los gráficos muestran la ubicación y la escala de los datos por medio de la intersección de dos segmentos de línea, ya sea en las líneas de regresión o en los ejes mayor y menor. Sin embargo, este enfoque estima parámetros basados en un modelo estadístico supuesto, en lugar de ser de distribución libre.

D. Regiones de más alta densidad

En 1996, Hyndman [8] propuso una *highest density regions* (HDR), con la idea de resumir la distribución de probabilidad de una región del espacio muestral que cubre una probabilidad especificada, seleccionando una región que contiene relativamente alto nivel de densidad de probabilidad. El HDR es, en esencia, un gráfico bivariado de contorno que se construye mediante el uso del 50% y 99% de las regiones de más alta densidad, y está centrado en la moda de los datos. Para construir el gráfico HDR es necesario estimar la densidad bivariada de los datos (utilizando un método de Kernel, por ejemplo), y entonces el 50% y 99% de las regiones de más alta densidad se suponen en el diagrama de dispersión. El 50% y 99% de las regiones de más alta densidad están dadas por los contornos de densidad que comprenden 50% o 99% de la masa de probabilidad. Este tipo de gráfico es muy adecuado para la visualización de datos multimodales.

E. *Bagplot*

Rousseeuw, Ruts y Tukey [5], así como Rousseeuw y otros [9], proponen otra extensión bivariada de la gráfica de caja univariada de Tukey: una *bagplot*. El concepto clave es la ubicación de la profundidad de un punto, la cual extiende el concepto de univariado rango. La ubicación de la profundidad ldepht(θ, Z) de un cierto punto $\theta \in \mathbb{R}^2$ relativa a una nube de datos bivariados $Z = z_1, z_2, \cdots, z_n$ fue introducida por Tukey (1975). Es el número más pequeño z_i contenido en cualquier espacio cerrado con una cota a través de θ. La región de profundidad D_k es el conjunto de todos los θ con ldepth$(\theta, Z) \geq k$. Las regiones de profundidad son polígonos convexos, y $D_{k-1} \subseteq D_k$. La ubicación más profunda es la mediana. La profundidad mediana de Z se define como el θ con mayor ldepth si solo hay un θ; de lo contrario, se define como el centro de gravedad de la región más profunda. Para obtener la bolsa, lo primero que se determina es el valor de k para el número de puntos de datos en $D_k \leq [n/2] <$ el número de puntos de datos en D_{k-1}, y luego se interpola linealmente entre D_k y D_{k-1} relativo a la mediana. La bolsa contiene $n/2$ de las observaciones con mayor profundidad que rodea la media. La “valla" es definida por aumento de la bolsa por un factor de 3. Los puntos fuera de las vallas se marcan como valores atípicos.

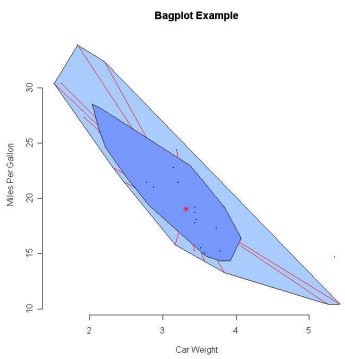

Figura 3. *Bagplot*

F. *Two-dimensional boxplot*

Tongkumchum [6] propone una extensión del *boxplot* univariado de Tukey: *two-dimensional boxplot*. Este gráfico está formado por un par de paralelogramos orientados en la dirección de una línea recta ajustada, con símbolos que denotan los valores extremos. La línea recta que muestra la relación entre las dos variables es la línea de Tukey. Los componentes principales del gráfico son una caja interior, que contiene 50% de los puntos de proyección de las observaciones en la línea de ajuste; un punto medio en la caja interior, y una caja externa que separa los valores atípicos. El *two-dimensional boxplot* permite visualizar la ubicación, extensión, correlación y asimetría de los datos.

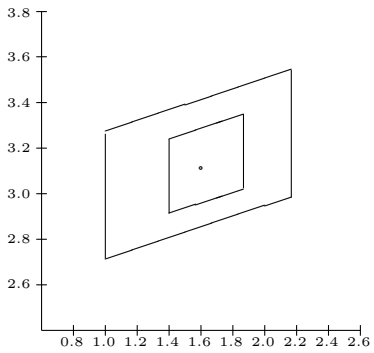

Figura 4. *Two-dimensional boxplot*

II. Construcción del *boxplot* bidimensional.

Un concepto básico en un *boxplot* bidimensional radica en ajustar una línea recta para el gráfico de dispersión, de tal suerte que se construya un polígono que rodea la línea ajustada. En el caso que nos ocupa, se utilizó el procedimiento que se describe a continuación para construir un *boxplot* bidimensional.

A. Línea de ajuste de Tukey.

Sean (x_i, y_i) el conjunto de datos, para $i = 1, 2, \cdots, n$. Se Dividen estos puntos en tres regiones disjuntas, de acuerdo con sus valores de x, de modo que cada región contiene igual o casi el mismo número de puntos, aproximadamente un tercio de las observaciones. Se calculan la mediana de los valores de x y la mediana de los valores de y en cada una de las regiones exteriores, valores denotados por (x_1, y_1) y (x_3, y_3), respectivamente. La pendiente b de la línea que une los puntos (x_1, y_1) y (x_1, y_1) es:

$$b = \frac{y_3 - y_1}{x_3 - x_1}$$

y la intersección a de la línea con el eje y es:

$$a = \text{mediana}\{y_i - bx_i\}.$$

De esta manera se obtiene la línea de Tukey $\widehat{y} = a + bx$. La mediana se define como el punto medio entre (x_1, y_1) y (x_3, y_3).

B. Líneas cuartiles y vallas

Se encuentra el par (r_i, θ_i) asociado con cada par (x_i, y_i). Se denota por $Q_{\theta(j)}$ el j-ésimo cuartíl de θ_i. Las líneas cuartiles, a su turno, se definen como $\theta = Q_{\theta(1)}$ y $\theta = Q_{\theta(2)}$, donde $Q_{\theta(1)}$ y $Q_{\theta(2)}$ son los cuartiles inferior y superior de las θ_i. Se Denota por D el rango intercuartil de las θ_i, y se definen las vallas como $\theta = Q_{\theta(1)} - D$ y $\theta = Q_{\theta(2)} + D$. La figura 7 muestra los cuartiles por medio de líneas rojas, y las vallas con azules.

C. Cajas interior y exterior

La línea fuerte de Tukey arroja la orientación de la caja interior. Por tanto, se considera el polígono convexo que está acotado por las líneas cuartiles, dos rectas paralelas a la recta de Tukey y que contiene el 50% de los datos. La caja exterior se construye de manera similar a la interior.

III. Implementación numérica

Con el ánimo de ejemplificar la construcción del *boxplot* bidimensional se toman de [10] los siguientes 65 datos, correspondientes a las variables aleatorias: X: Concentración de bifenilos policlorados (PCB) y Y : Espesor de la cáscara de huevo, de huevos de pelícano.

Conc	Esp	Conc	Esp	Conc	Esp	Conc	Esp
452	0,14	184	0,19	115	0,2	315	0,20
139	0,21	177	0,22	214	0,22	356	0,22
166	0,23	246	0,23	177	0,23	289	0,23
175	0,24	296	0,25	205	0,25	324	0,26
260	0,26	188	0,26	208	0,26	109	0,27
204	0,28	89	0,28	320	0,28	265	0,29
138	0,29	198	0,29	191	0,29	193	0,29
316	0,29	122	0,30	305	0,30	203	0,30
396	0,30	250	0,30	230	0,30	214	0,30
46	0,31	256	0,31	204	0,32	150	0,34
218	0,34	261	0,34	143	0,35	229	0,35
173	0,36	132	0,36	175	0,36	236	0,37
220	0,37	212	0,37	119	0,39	144	0,39
147	0,39	171	0,40	216	0,41	232	0,41
216	0,42	164	0,42	185	0,42	87	0,44
216	0,46	199	0,46	236	0,47	237	0,49
206	0,49						

Tabla 1. Datos concentración vs. espesor [10]

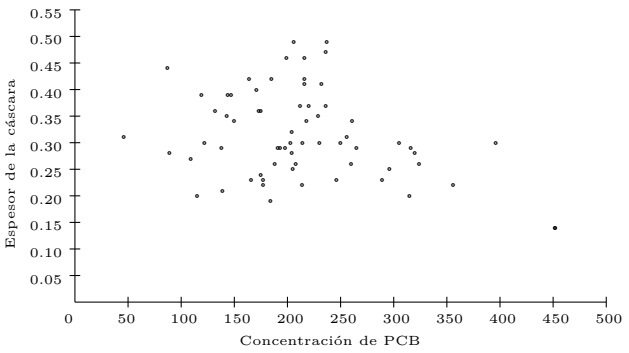

Figura 5. Diagrama de dispersión

La figura 5 corresponde al gráfico de dispersión para estos datos.

Considerado el primer paso para la construcción del *boxplot* bidimensional se obtienen la línea de Tukey y la mediana que se muestran en la figura 6.

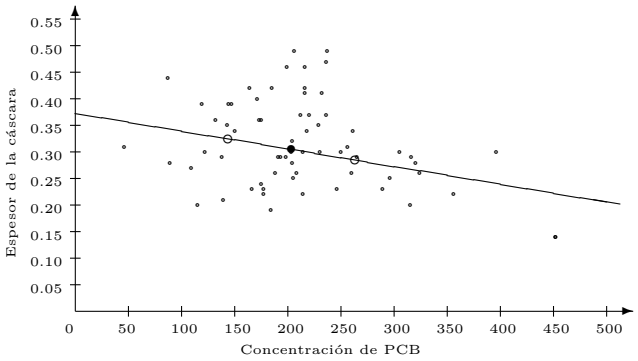

Figura 6. Línea de Tukey

Para el cálculo de las líneas cuartiles y las vallas se lleva a cabo el paso 2 para la construcción del *boxplot* bidimensional. En la figura 7 se ilustran las líneas cuartiles con rojo y las vallas con azul.

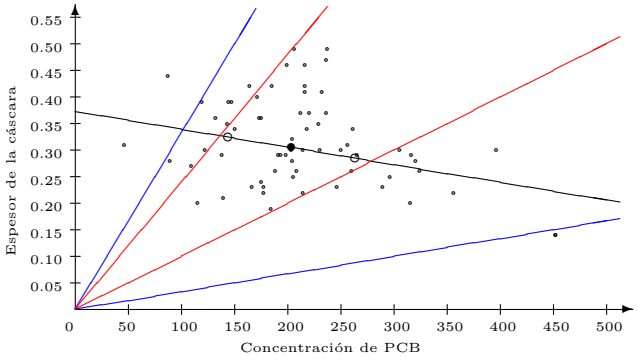

Figura 7. Cuartiles y vallas

Mediante el paso 3 de la construcción del *boxplot* bidimensional, se construyen la caja interior y exterior (figura 8).

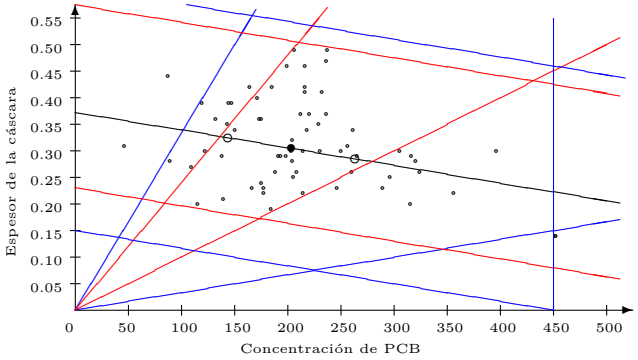

Figura 8. Construcción de cajas interior y exterior

La figura 9, a su turno, explicita la *boxplot* bidimensional.

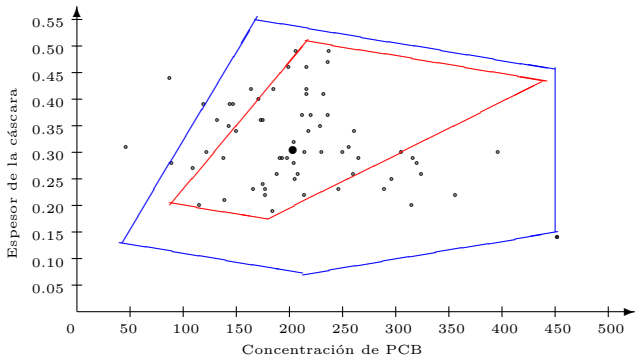

Figura 9. *Boxplot* bidimensional

IV. Conclusiones

El presente artículo ha mostrado el desarrollo de un estudio de las construcciones de *boxplot* en el caso de tener 2 variables aleatorias continuas. En particular, el trabajo que nos ocupa fue iniciado para *boxplot* en el caso de tener una variable aleatoria continua.

En este trabajo se construye un *boxplot* en dos dimensiones que extiende el *boxplot* univariado de Tukey. Este gráfico está conformado por dos polígonos convexos que están orientados en la dirección de una linea recta ajustada (línea de Tukey). El *boxplot* bidimensional representa ubicación, extensión, correlación y asimetría de los datos.

El *boxplot* bidimensional basado en coordenadas polares es más simple que el *bagplot* en su construcción y es más general que el *two-dimensional boxplot*, pues el nuestro no considera paralelogramos sino polígonos convexos más generales.

A través del problema implementado numéricamente, se muestra la construcción del *boxplot* bidimensional.

Referencias

[1] J. Correa and N. González, *Gráficos estadísticos con R*. Medellín: Universidad Nacional, sede Medellín, 2002.

[2] C. Gaviria and C. Márquez, *Estadística descriptiva y probabilidad*. Medellín: Editorial Bonaventuriana, 2019.

[3] S. Becketti and W. Gould, "Rangefinder box plots: A note," *Amer. Statist*, vol. 41, no. 2, p. 149, 1987.

[4] K. Goldberg and B. Iglewicz, "Bivariate extension of the boxplot," *Technometrics*, vol. 34, no. 3, pp. 307–320, 1992.

[5] P. Rousseeuw, I. Ruts, and J. Tukey, "Bivariate extension of the boxplot," *Amer. Statist*, vol. 53, no. 4, pp. 382–387, 1999.

[6] P. Tongkumchum, "Two-dimensional box plot," *Songklanakarin J. Sci. Technol.*, vol. 27, no. 4, pp. 859–866, 2005.

[7] J. Tukey, *Exploratory Data Analysis*. Massachusetts: Addison-Wesley Publishing, 1977.

[8] R. J. Hyndman, "Computing and graphing highest density regions," *The American Statistician*, vol. 50, no. 2, pp. 120–126, 1996.

[9] K. Miller, S. Ramaswwami, P. Rousseeuw, J. Sellarés, D. Souvaine, I. Streinu, and A. Struyf, "Efficient computation of location depth contours by methods of computational geometry," *Statistics and Computing*, vol. 13, pp. 153–162, 2003.

[10] D. McNeil, "Epidemiological research methods," *Jhon Wiley and Sons ltd*, p. 78, 1996.