Oriented Spatial Box Plot, a New Pattern for Points Clusters

Laurent Etienne

Laboratory of Computer Science, Tours University, Tours, France
E-mail: laurent.etienne@univ-tours.fr

Thomas Devogele

Laboratory of Computer Science, Tours University, Blois, France
E-mail: thomas.devogele@univ-tours.fr

Gavin McArdle

National Centre for Geocomputation, Maynooth University, National University of Ireland Maynooth, Maynooth, Co. Kildare, Ireland
E-mail: gavin.mcardle@nuim.ie

Abstract: Nowadays, an abundance of sensors are used to collect very large datasets containing spatial points which can be mined and analyzed to extract meaningful patterns and information. This article examines patterns which describe the dispersion of 2D data around a central tendency. Several state of the art patterns for point cluster analysis are presented and critiqued before a new pattern, the Oriented Spatial Box Plot, is defined. The Oriented Spatial Box Plot extends the classical one-dimensional box plot for summarizing and visualizing 2D point clusters. The pattern is suitable for detecting outliers and understanding the spatial density of point clusters.

Keywords: Point clusters, Oriented Spatio-Temporal Box Plot, Bagplot, Quelplot, Outlier detection, Spatio-temporal patterns.

Reference to this paper should be made as follows: Etienne, L., Devogele, T. and McArdle, G. (xxxx) 'Oriented Spatial Box Plot, a New Pattern for Points Clusters', *International Journal of Business Intelligence and Data Mining*, Vol. x, No. x, pp.xxx–xxx.

Biographical notes:
Dr Laurent Etienne is an Assistant Professor at University Francois Rabelais, Tours, France. He holds a Ph.D. in geomatics from University of Brest (France) and a M.Sc in Computer Science from University of Rennes (France). His main research interests include spatial data mining and knowledge discovery from moving objects databases, spatio-temporal patterns of mobile objects, outlier detection and maritime Geographic Information Systems (GIS).
Dr. Thomas Devogele is a Professor in computer science at university Francois Rabelais from Tours (France) since 2010. He was previously Associate Professor

Copyright © 2014 Inderscience Enterprises Ltd.

in the French Naval Academy Research Institute. His research interests include Spatio-temporal databases and data mining, GIS, computational geometry, navigation systems. His current research activities are related to moving objects analysis. He is the leader of the French national working group on mobility and GIS (MAGIS research network).
Dr Gavin McArdle is a Research Fellow at the National Centre for Geocomputation at Maynooth University in Ireland. He has Ph.D. and B.Sc degrees in Computer Science from University College Dublin. His research interests include geovisual analytics, location based services, geographic information systems, user interface design and smart city technologies. Currently, Dr McArdle is investigating new techniques for understanding human mobility in an urban setting.

1 Introduction

Soliciting meaning from very large datasets which contain massive amounts of data is now a real problem. A common first step in such analysis is to partition the dataset into clusters by grouping similar data points Jain et al. (1999). Each cluster then consists of objects that are, by some measure, similar to one another and dissimilar to objects of other groups (Berkhin, 2006). A second classical step is to summarize and describe the contents of each point cluster using statistical or artificial intelligence tools. These tools produce more compact and often graphical representations of the data known as patterns. These patterns are useful for:

- Visual analysis of data to compare the patterns of different clusters;
- Outlier detection to identify data points not in a cluster;
- Increasing understanding of changes in space and time.

In the geographical domain, the increasing use of GPS and other tracking technologies, the development of monitoring systems and the emergence of crowd-sourcing have dramatically increased the volume of spatio-temporal data which records the movement of phenomenon through space and time. Clustering techniques and pattern summaries are therefore essential to understand and analyze such data as well as for the detection of outliers. Geographical datasets are complex with points generally consisting of a pair of correlated values (latitude, longitude) or (X, Y) linked to a reference system. Altitude and temporal values can also form part of the complete coordinate.

Many techniques for clustering this type of complex data have been proposed (Berkhin, 2006; Goldberg and Iglewicz, 1992). Several patterns and statistics have been defined to describe the data contained in point clusters. For example, the density of a cluster, according to spatial units can be calculated. Similarly, the dispersion of data around a central tendency can be used to describe the cluster of points. Other techniques, such as the home range kernel density (Worton, 1989), combine these approaches to describe the density of animals in their home range (the area where an animal lives and moves).

This article focuses on using dispersion patterns around a central tendency to produce a compact representation for describing a point cluster using the following criteria:

- The central tendency which describes the normal location of point;
- The spread of a cluster without outlier points;
- The orientation of the point cluster which is linked with the correlation between the X and Y value of the points;
- The shape of the distribution which indicates if the distribution is symmetric or asymmetric.

Using these criteria, the following important tasks can be completed effectively:

- Evaluation of a set of points;
- Visual analysis of point clusters;
- Outlier detection;
- Comparison of point clusters from different types of objects (from two different populations for example) or at different periods;
- Detection of breakpoints where statistical behaviour changes.

Often, it is challenging to determine which pattern is the most appropriate and descriptive for a particular data set. Therefore, in this paper we describe the merits of several existing techniques to define a compact representation of 2D point clusters. This compact representation is used to visualize and understand the spatial density of a point cluster taking its shape and symmetry into account. The more precise and concise the pattern is, the easier it is to understand and interpret the point clusters. The remainder of this article is organized as follows. The next section presents the state of the art in patterns for point cluster analysis. This is followed by a detailed description of the new Oriented Spatial Box Plot (OSBP) pattern which describes clusters of spatial points by extending the traditional Box Plot. Finally, the paper concludes with a summary of the research and highlights some areas for future research.

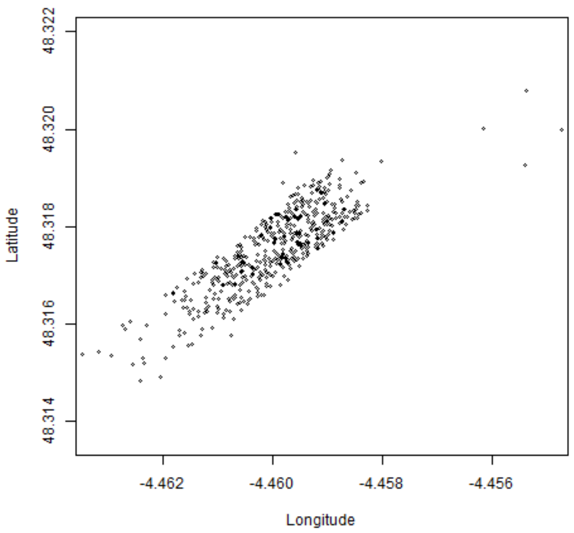
2 State of the art in Patterns for Point Cluster Analysis

2.1 Distribution of data

Point clusters can have very different characteristics. Some clusters have irregular shapes, with large concaves or holes while others are more regular, composed of core and satellite points. This article focuses on presenting techniques for describing regularly shaped clusters of points without holes or large cavities.

To illustrate and compare several techniques and processes for analysing point clusters, we apply each one to a single reference point cloud shown in figure 1. This point cloud contains 519 real geographic points, grouped into a single cluster. Each point is therefore described by its longitude and latitude coordinates. For point clusters, it is common to add a boundary of normality around specific points, called a fence, which separates outliers from the point cluster and allows online classification of new points as they are added to the dataset. In this section, we discuss some classical methods for defining this area of normality for 2D point datasets and discuss the relative benefits of each before presenting a new approach, which extends the traditional box plot for use with 2D point data, in the next section.

Figure 1 A point cloud of 519 geographic points.



2.2 *Patterns for a normal distribution*

The normal distribution, which describes the central tendency of data, is the most prominent probability distribution in statistics. Only two parameters summarize this distribution: the mean as central tendency and the variance or standard deviation which measures how far a set of values spread. For this distribution, bounding fences are not fixed by a rigid mathematical definition but usually represented by a distance of 2 or 3 standard deviations from the mean. With this distance, outliers represent 5% or 0.3% of the dataset. For point data in the form of (X, Y) or (X, Y, Z), the multivariate normal distribution which generalizes the one-dimensional (1D) normal distribution to higher dimensions, could be employed. Multivariate normal distributions are described by a mean vector and a covariance matrix. For the example points depicted in figure 1, the covariance matrix is presented in table 1.

Table 1 Covariance matrix of the point cluster.

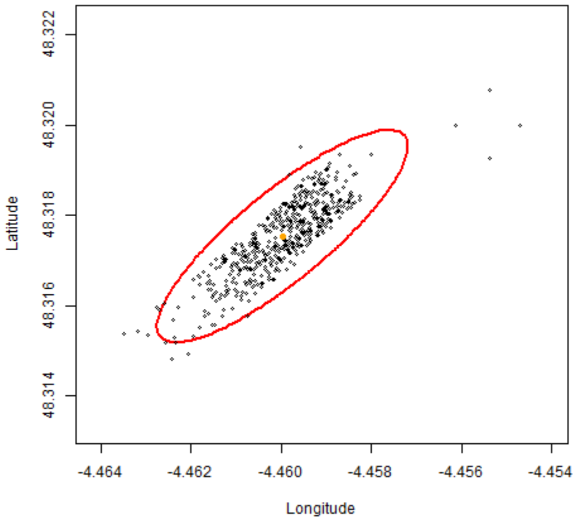
$var(X)$	$covar(X,Y)$	$\rightarrow$	10.042110^{-7}	7.208610^{-7}
$covar(X,Y)$	$var(Y)$		7.208610^{-7}	7.173010^{-7}

If a statistical relationship exists between coordinates, the covariance matrix is not a diagonal one and the coordinates must be considered simultaneously. If covariance is close to zero, coordinates are uncorrelated and each coordinate can be examined separately. Otherwise, the Pearson's Correlation Coefficient of coordinates, also called correlation coefficient (Equation 1), reflects the direction of the linear relationship between coordinates. For the example points in figure 1, the latitude and longitude are correlated and the value of the correlation coefficient is 0.8493. This value represents the strength of this positive linear relationship.

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} \tag{1}$$

A Standard Deviational Ellipse (SDE) or a Cross of Dispersion (Lefever, 1926; Kenward, 1987) extends the normal distribution for two values. This two-dimensional (2D) pattern describes the dispersion of spatial data around the mean. The orientation of the ellipse is defined according to the correlation between X and Y and the lengths of the two axes are defined according to standard deviations. In the same way as for normal distribution, where the two axes are defined with 2 standard deviation values, outlier points outside the ellipse represent 5% of the dataset. Figure 2 gives an example of an SDE.

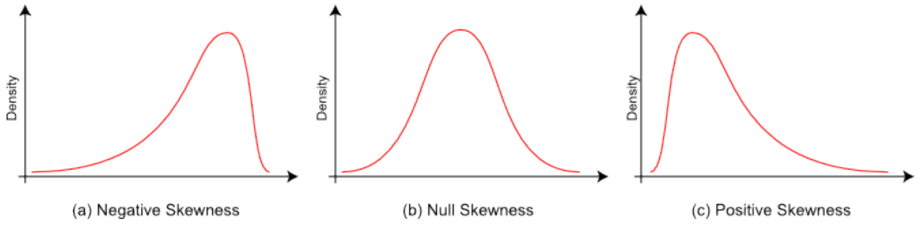
Figure 2 Standard Deviational Ellipse (SDE) for the point cluster.



The SDE is not a robust method as it can only be used with symmetric distributions which are close to a normal distribution. For example, if the distribution is asymmetric, the mean is different to the median and so the mean is not a good estimation of central tendency and the median is preferred. In the same way variance must be replaced by quartiles. To measure the symmetry of a distribution, skewness is employed Ferreira and Steel (2006). If the distribution is symmetric then the mean is equal to the median and the distribution will have a close to zero skewness (figure 3.b). If skewness is negative, the left side (left tail) is longer than the right side (figure 3.a). Reciprocally, if skewness is positive, the right tail is longer (Figure 3.c).

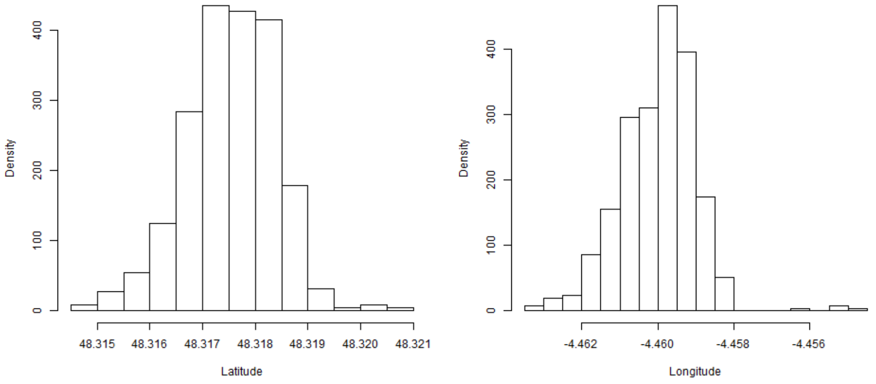
For the cluster sample depicted in figure 1, the histograms in figure 4 show that longitude is more asymmetric than latitude. Indeed, the skewness of longitude is -0.818 and the skewness of latitude is -0.149. Unfortunately the skewness of the two coordinates cannot

Figure 3 Skewness and shape of the distribution.



be studied simultaneously. Nevertheless, these two values show that the point distribution is asymmetric.

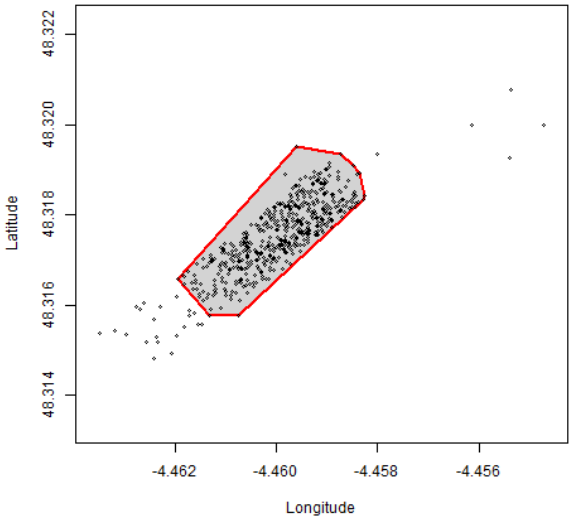
Figure 4 Histograms of longitude and latitude.



2.3 *Usual patterns for asymmetric distribution*

The second class of pattern often used to summarize a point cluster is a Minimum Convex Polygon (MCP) (Mohr, 1947). The MCP is the smallest convex polygon around the dataset. Figure 5 illustrates the MCP of the point cluster with a value of 5% for outlier data. To eliminate outliers, the simplest method by peeling is to exclude points sequentially that are farthest from a central tendency (Kenward, 1987). Another method called Nearest Neighbour Convex Hulls (NNCH) aggregates points. For example, Getz and Wilmers (Getz and Wilmers, 2004) propose to aggregate small MCPs with k-1 nearest neighbours. This last pattern is not convex and holes are possible. MCP and derived patterns are often employed to define the home range. Unfortunately, these patterns are not compact representations due to the large subset of points required to define the fence. Moreover, MCP and derived patterns are very complex to compute. Some proposals combine SDE and MCP to study the evolution of phenomena during time or to compare two different clusters (Thériault et al., 1999).

Figure 5 Minimum convex polygon (MCP) for the point cluster (95%).



2.4 Box Plot Extensions

For 1D data, when the dataset does not belong to any particular statistical distribution and when gaps between the mean and median values are important, the Box Plot (Tukey, 1977) is a useful descriptive pattern. This kind of distribution is very common for spatial data points. The Box Plot graphically describes groups of numerical data using five important sample percentiles:

- The sample minimum (smallest observation);
- The lower quartile (Q1);
- The median;
- The upper quartile (Q3);
- The sample maximum (largest observation).

50% of data are between Q1 (the lower hinge) and Q3 (the upper hinge). To identify outlier data, the interquartile range (IQR) is used. The IQR is a measure of dispersion and is equal to the difference between the upper and lower quartiles (Q3 â˘’ Q1). An outlier datum is defined as the lowest datum still within 1.5 IQR of the lower quartile (Q1-1.5 IQR), or the highest datum still within 1.5 IQR of the upper quartile (Q3+1.5 IQR). The 1st decile and the 9th decile are also often used as a boundary, called the whiskers or the fences. Figure 6 shows the Box Plot bounded by Q1 and Q3, with median values inside the box and the whiskers.

Box Plots are useful for visualizing and comparing the shape and âŁœpeakednessâŁž of a distribution which is generally described by Kurtosis. Kurtosis is classed as leptokurtic

Figure 6 Box Plot of one dimension dataset.

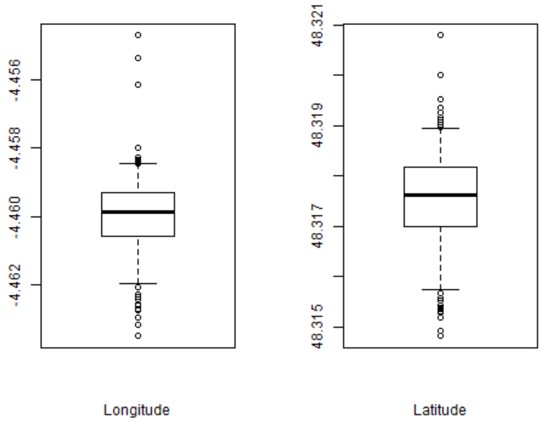
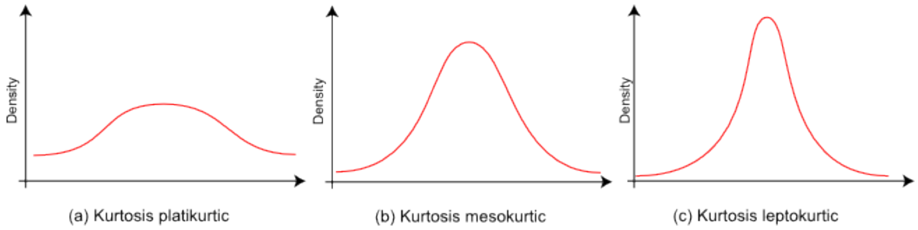


Figure 7 Shape of the distribution and Kurtosis.



(a) Kurtosis platikurtic
(b) Kurtosis mesokurtic
(c) Kurtosis leptokurtic

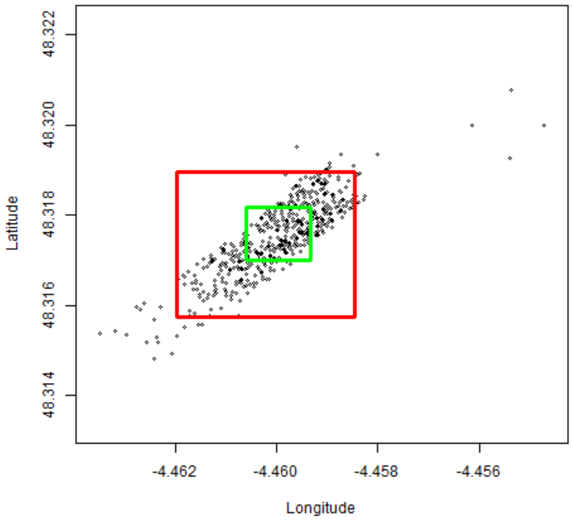
with a high Kurtosis distribution (value >1) characterized by sharper peaks and longer, fatter tails (figure 7.c). Reciprocally, a low kurtosis distribution (value <1) has more rounded peaks and shorter, thinner tails and the kurtosis is described as platykurtic (figure 7.a). Finally, the Kurtosis value of normal distribution is 1 and characterized as mesokurtic (figure 7.b). The Box Plot can effectively describe shapes using the IQR, while the SDE and MCP can be extended to incorporate shape through the addition of a second ellipse or bag containing 50% of points. However these approaches are only capable of describing data in a 1D distribution.

The Box Plot pattern can be extended to several dimensions (Rangefinder Box Plot (Becketti and Gould, 1987)), only if values of each dimension are not correlated. Unfortunately, if coordinates are correlated, the N-dimensional Box Plot is too large. Figure 8 presents a 2-dimension Box Plot. Since some parts of this cube are very empty (left upper corner and right lower corner in this example), it can be seen that this box plot does not produce a good fit for correlated point coordinates.

2.5 Hybrid Patterns for Analyzing 2D Points

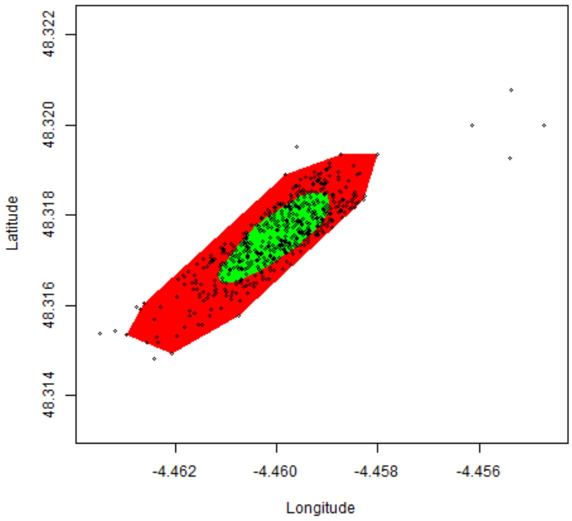
To improve these classical 2D patterns, their respective advantages have been combined to provide hybrid solutions. For example, a Bagplot (Rousseeuw et al., 1999) summarizes the cluster by a depth median consisting of a bag containing 50% of points and a fence

Figure 8 The two rectangles and one median point defined by the Rangefinder Box Plot method.



(magnifying the bag by a factor 3). Figure 9 gives an example of a Bagplot applied to the original set of points described in figure 1.

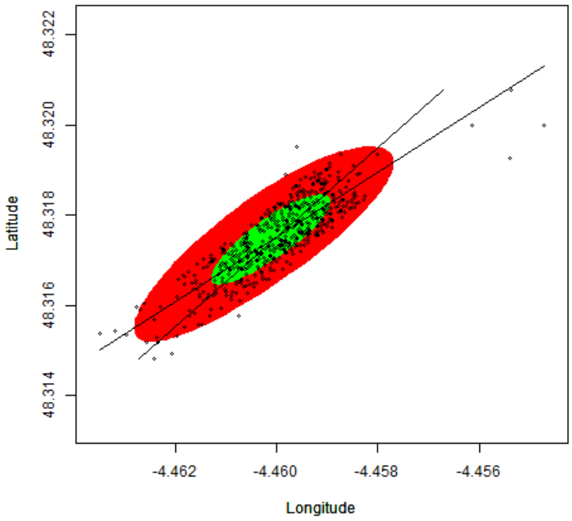
Figure 9 Bagplot for the point cluster.



This representation mixes the MCP and Box Plot approaches and visualizes the location, shape, spread, correlation and skewness simultaneously. Unfortunately, two large sets of points are necessary to define the bag and the fence which makes it an inefficient and impractical approach which is difficult to manipulate.

Similar to the Bagplot, the Quelplot (a quarter elliptic plot) (Goldberg and Iglewicz, 1992) replaces the ellipses by four separate quarter ellipses matched on their major and minor axes. While the Quelplot is a compact pattern to summarize an asymmetric dataset, the approach estimates parameters to compute the hinge and the fence according to statistical model assumptions, rather than the real dataset. Asymmetric real data distributions are modelled as complex deformations of a normal distribution. Figure 10 shows the Quelplot for the example point cluster.

Figure 10 Quelplot for the point cluster.



The Box Plot is the most popular pattern to summarize a distribution. Several extensions of the Box plot have been proposed (Potter, 2006; Wickham and Stryjewski, 2011). Unfortunately, the standard Box Plot is not suited for 2D data. Tongkumchum (Tongkumchum, 2005) proposes a Bivariate Box Plot, defined by a pair of trapeziums oriented in the direction of a single fitted axis. The trapeziums are defined according to the X-axis only with 2 edges are parallel to the Y-axis but for points, the X-axis and the Y-axis are equally important.

Table 2 summarizes the different patterns described above by comparing the characteristics of the data points which they can effectively describe. The table shows that while SDE, MCP and Box Plot are relevant patterns for describing 2D point clusters, weaknesses exist. For example, no classical solution has the ability to summarize an asymmetric correlated coordinate cluster. The table reinforces the need for the new approach, described in the next section, for this type of data.

Table 2 Comparative summary of 2D patterns.

	SDE	MCP	Rangefinder Box Plot	Bagplot	Quelplot	Bivariate Box Plot
2D data	+++	+++	+	+++	+++	++
Central point	+	-	+	+	+	+
Skewness (Symmetry)	-	+	+	++	++	++
Kurtosis (shape)	-	-	-	+++	+	++
Outlier detection	+	++	-	+++	+++	++
Compact pattern	++	-	++	-	+	++
X & Y axis symmetry	++	++	++	++	++	–

3 Oriented Spatial Box Plot (OSBP)

This section defines a new 2D pattern for concisely describing spatial point clusters. As with other 2D Box Plots, the new pattern, called the Oriented Spatial Box Plot (OSBP), relies on specific parameters to synthesize the point cluster. These are:

- Central tendency;
- A central shape that encompasses 50
- A fence shape separates the normal data from potential outliers (outer fence).

With this technique, rectangular shapes are selected to represent the inner and outer fences. Rectangles are efficient to compute, only requiring 4 angles to describe this shape. Similarly, it is easy to compare two different rectangles and to check if a point is inside or outside the rectangle. The principle idea is to minimize the size of the two rectangles to match the cluster shape. While it is not necessary for the rectangles to be parallel to the X and Y axes, their relative orientation must be equivalent.

Figure 11 The process used to define the Oriented Spatial Box Plot.

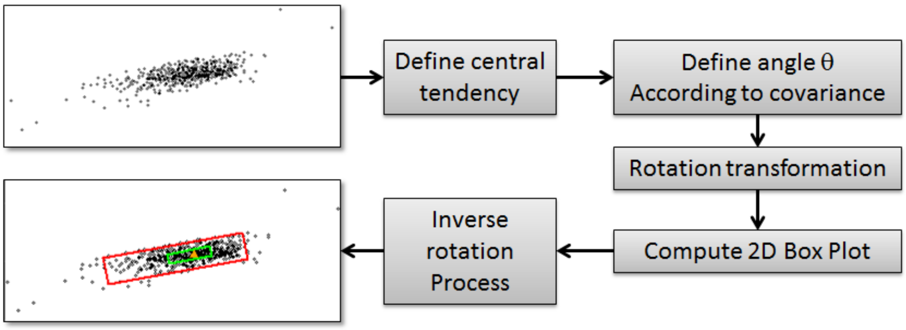
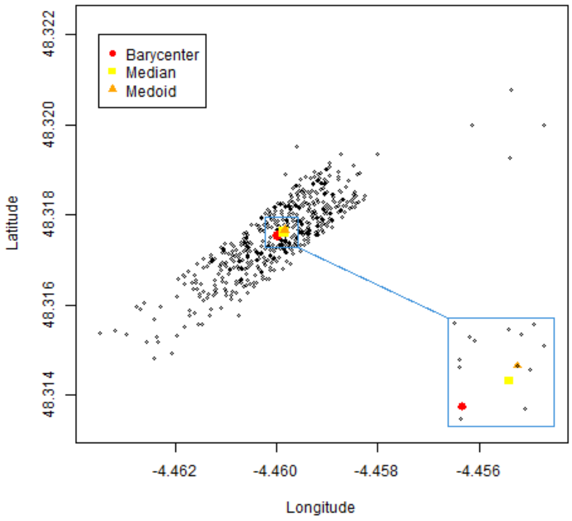


Figure 11 provides an overview of the process used to compute the OSBP[1] , while the following sections describe the steps in more detail. A maritime application of OSBP is proposed in (Devogele et al., 2013).

3.1 Defining Central tendency

In the 1D Box Plot concept, the central tendency (Small, 1990) of the dataset is represented using the median value. However, for 2D values, three different approaches can be used to define and visualize this tendency measure. Firstly, the mean point of all the points in all of the coordinate directions is called the Barycenter. As this value is computed using the arithmetic mean of all the points of the cluster, it is very sensitive to outliers. Another approach computes and combines median values for each coordinate to create a geometric median. While, the geometric median can be used to resolve the issue of outliers, the geometric median does not correspond to a real point in the cluster.

Figure 12 Barycenter, Medoid and Geometric median for the point cluster.



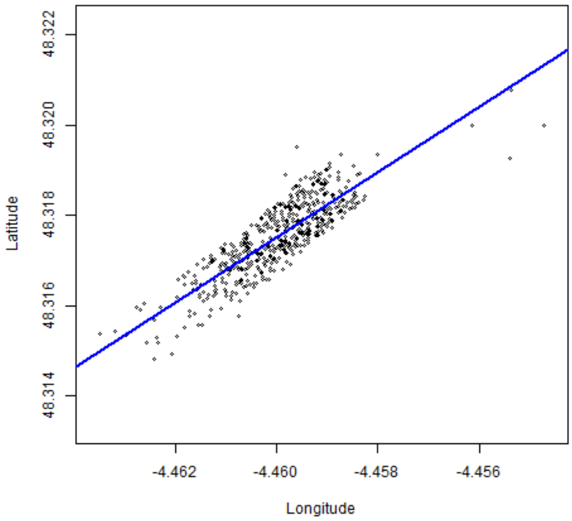
Finally, another technique to define the central density of a point cluster is to use the value of the medoid. The medoid is the point of the cluster whose average distance to all the points in the cluster is minimal. The geometric median is faster to compute but represents a virtual point, whereas the medoid represents a real point. Moreover, if the shape of the cluster is particular; with a large hole or the shape of cluster contains a deep concave; the geometric median could be distant to other points. Figure 12 highlights these three particular points. The geometric median is represented as a yellow square, the medoid of the cluster of points is represented as an orange triangle and the Barycenter is highlighted with a red circle. In this particular example, the different reference points are very close together, however, the Medoid is the only point that corresponds to a real point of the cluster.

3.2 Defining the angle

The edges of the classical Box Plot are parallel to an axis. In the Range Finger Box Plot, four edges are parallel to the axes while for the Bivariate Box Plot only two edges are parallel to the X axis. To minimize the size of the 2D Box Plot and to fit the two rectangles to the shape formed by the points of the cluster, this parallel constraint is removed. To achieve this, the orientation of the rectangles needs to be defined. Pearson's correlation coefficient (ρ) between coordinate values is chosen as the key factor. This coefficient is efficient to compute and frequently used for problems of this type, for example, it is used for Principal Component Analysis (PCA). Indeed, the original concept of Pearson (Pearson, 1901) was to determine a straight line to represent the best fit for a set of data points. In the Oriented Box Plot problem, we define a rectangle which best fits the points of a specific cluster. This coefficient is symmetric ($\rho X, Y = \rho Y, X$). While, this coefficient is not robust if a large set of outliers are present, for common point clusters, the number of outliers is generally small. To translate the Oriented Box Plot problem to a Range Finger Box Plot with edges parallel to the axes, a rotation transformation is used. The angle of rotation (θ) is the same as the angle between the X axis and the first principal component of PCA (Equation 2).

$$\theta = tan^{-1}\left(\frac{cov(X,Y)}{var(X)}\right) \tag{2}$$

Figure 13 The best fit straight line for the point cluster of the example.



For the example, the value of θ is $11.56°$. Figure 13 shows the best fit straight line for this point cluster.

3.3 Rotation transformation

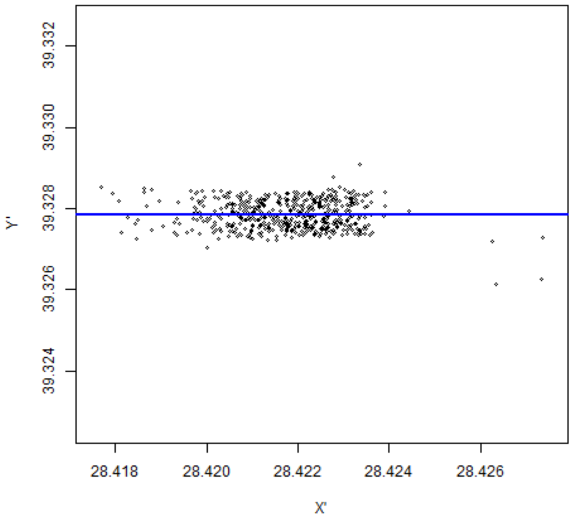
Classically, the coordinates of the point after rotation (x', y') can be computed using these formulae:

$$x' = xcos(\theta) - ysin(\theta) \tag{3}$$

$$y' = xsin(\theta) + ycos(\theta) \tag{4}$$

These axes are also called the first principal direction (X') and the orthogonal direction (Y'). After this rotation, the first principal direction is horizontal. In the same way, the value of $\rho X, Y$ is equal to zero. The new coordinates of points after the rotation can be easily examined separately. In Figure 14, it is seen that the shape of the cluster is horizontal.

Figure 14 Coordinates of the point of cluster after rotation.



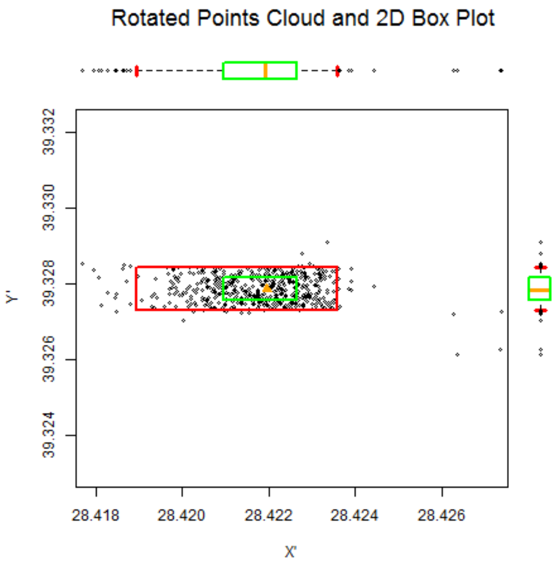
3.4 Compute 2D Box Plot

In this step, the Box Plot of x' values and the Box Plot of y' values are separately computed. The process is the same as the Range Finger Box Plot process using 2.5% as the threshold for outlier data. This produces the following descriptive statistics:

- 1.25% of outlier point is above of the fence rectangle;
- 1.25% is below;
- 1.25% is at the right;
- 1.25% at the left.

The 1.25% value is used in the same way as in the central rectangle. The edges of the two rectangles are parallel to the X' and Y' axis. Figure 15 shows the computed Box Plot for the x' values (above the plot) and the one for the y' values (right of the plot). Using this information, it is now possible to provide a compact description of the point cluster. For example, the two Box Plots are combined to define the fence while the central rectangles and the medoid, which is a real point of the cluster, are used to define the central tendency.

Figure 15 2D Box Plot to the cluster of rotated points.



3.5 Inverse rotation process

The last step involves using the inverse rotation transformation to rotate the points and box plot back to the original coordinate system. The corner coordinates of the two rectangles are computed (Equation 5 and Equation 6) and the results are shown in Figure 16.

$$x = x' cos(-\theta) - y' sin(-\theta) \tag{5}$$

$$y = x' sin(-\theta) + y' cos(-\theta) \tag{6}$$

4 Properties of Oriented Spatial Box Plot

Overall, it has been demonstrated how the OSBP extends the Box Plot concept for use with 2D data. The OSBP has the same compact properties for 2D data as Box Plot for 1D data. The OSBP pattern summarizes and eases the visualization of important information about the 2D statistical distribution of a point cluster while the central tendency is enhanced using the medoid value.

In addition to describing the central tendency and spread of the point cluster, the distance between the edges of the two rectangles and the medoid (see figure 17) can be used to describe the symmetry coefficient (s). The symmetry coefficient (equation 7) is defined as a simple coefficient to measure the symmetry of the 2D distribution.

$$s = \frac{(A+B)}{(A+B+C+D)} * \frac{|A-B|}{(A+B)} + \frac{(C+D)}{(A+B+C+D)} * \frac{|C-D|}{(C+D)} \tag{7}$$

Figure 16 Oriented Spatial Box Plot for the point cluster.

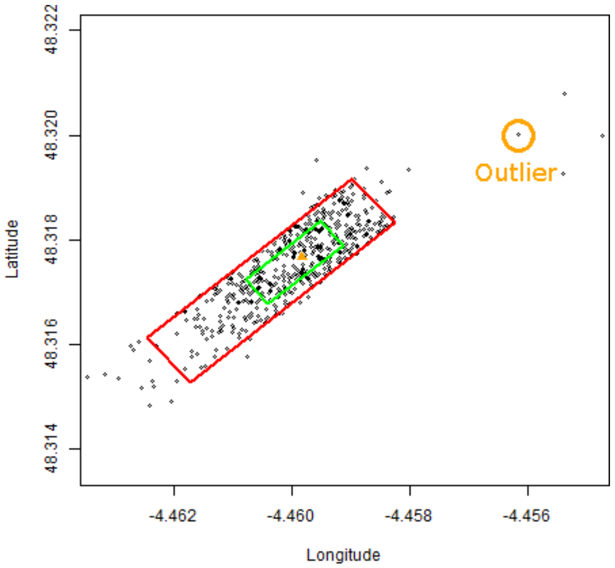
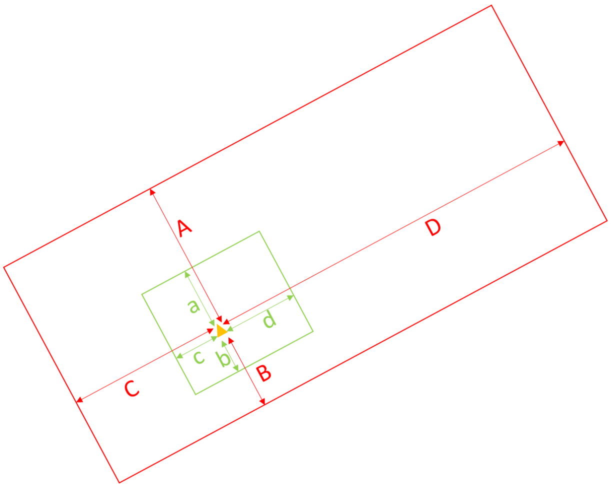


Figure 17 Distances used to compute the symmetry and the peakedness coefficients.



This coefficient is a weighted sum. The absolute differences between the right part and the left part of the rectangular fence (C and D) and between the higher and the lower part of the rectangle (A and B) are computed. These two absolute differences are normalized. Finally, weights are defined according to the length of these parts. Figure 17 shows that distances A and D are greater than B and C respectively and so the 2D distribution is asymmetric. If the coefficient is very close to 0, the distribution is symmetric. Alternatively, the more the coefficient raise, the more the distribution is asymmetric. In the same way, the shape or the âŁœpeakednessâŁž (k) of the distribution can be observed by examining the ratio between the surface of the central rectangle and the surface of the outer rectangular fence (equation 8).

$$k = \frac{((a+b)*(c+d))}{((A+B)*(C+D))} \tag{8}$$

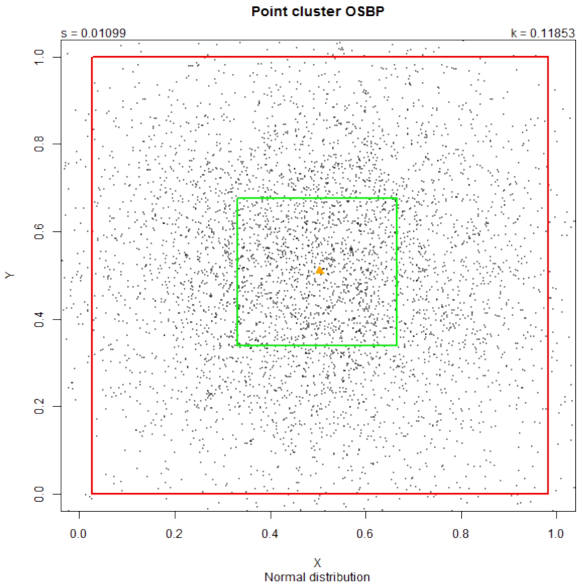
This ratio is called the peakedness coefficient. If this coefficient is low, a large part of data is grouped around the medoid and the âŁœtailsâŁž of the distribution are very long, so the peak is high. If this ratio is high, the distribution is more uniform. The OSBP approach is compact and robust to outliers. Outliers can be detected as points appearing outside the fences. In the same way, the OSBP is invariant under affine transformations of the plane. Moreover, the symmetry and peakedness coefficient are simple to compute.

Within this process, one analytical problem remains. The values of 50% of the data outside the central rectangle and 5% outside the outer rectangular fence are upper estimates. Indeed, an outlier could for example, fall below the outer fence (bottom limit outlier) and also to the right of the fences (right limit outlier). This point is then considered as an outlier twice, as illustrated in the circle of figure 16. Empirically, the outlier percentage is closer to 4.5% than 5%.

4.1 Visual Comparison of different Oriented Spatial Box Plot Patterns

To illustrate the applicability of the OSBP patterns to different point clusters, several examples are presented below. To ease the comparison of the different case studies, point clouds and the OSBP are presented following the CPA rotation. This explains why the OSBP rectangles are parallel to the X and Y axis. Each point cluster is composed of 5000 points whose x and y coordinates have been computed using various distributions.

Figure 18 OSBP of a normal distribution point cluster.



First of all, a normal distribution point cloud have been computed using identical parameters for x and y coordinates statistical distributions (mean = 0.5, standard deviation = 0.25). OSBP of this 2D normal distribution point cluster is presented on Figure 18. As expected, the symmetry coefficient is very close to 0 (s=0.01). The peakedness coefficient (k) is around 0.12. These two coefficient values for a 2D normal distribution will be used as reference to compare other point cloud distributions.

Figure 19 OSBP of two different symmetric point clouds (high peak vs uniform distribution).

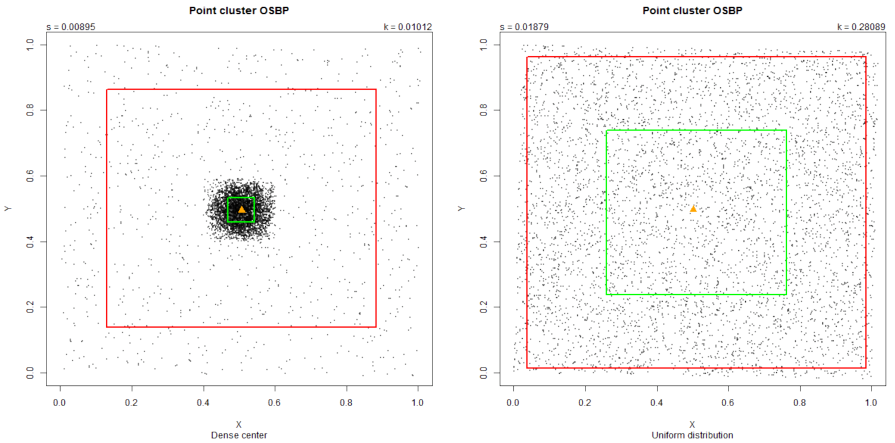


Figure 19 shows OSBP patterns of point clouds having a symmetric distribution but different peakedness. The relative surface of the central rectangle is larger than the fence for uniform distributions (right part of Figure 19) than for a normal distribution (Figure 18) and even more for the high peak one (left part of Figure 19). The symmetry coefficients of the two examples of figure 19 are very close to 0 indicating that the 2D distributions of the point clouds are symmetric. On the other hand, the peakedness coefficients reveal a significant difference. As expected, the uniform distribution has a bigger peakedness coefficient (k=0.28) than the normal distribution (k=0.12). The high peak distribution has the lowest peakedness coefficient (k=0.01). This understanding of the peakedness difference is visually efficient thanks to the OSBP display.

Asymmetric distributions can also be visually identified using the OSBP patterns. Figure 20 shows two asymmetric point cloud distributions. The left distribution is only asymmetric according to the X axis (s=0.15). The right point cloud distribution is asymmetric according to the two axes (s=0.28). The symmetry coefficients of the point clouds of figure 20 are higher than the one of figure 18 (s=0.01). Moreover, this coefficient is also more important when the point cloud distribution is skewed on both the X and Y axis (s=0.28). The OSBP shows the asymmetry and assists with the visual comparison of the point cluster symmetry. These 4 examples of point cloud distributions in the OSBP show that dispersion and symmetry criteria are easy to visualize and compare using the OSBP.

Figure 21 demonstrates how the OSBP can also be used to analyze the temporal evolution of a point cluster. Consecutive OSPBs can be combined into a single visualization to determine how the distribution of the point cluster changes over time. Figure 21 demonstrates this by showing 3 consecutive point clusters plotted with their OSBP. This visualization

Figure 20 Two asymmetric point cloud distributions and their OSBP patterns.

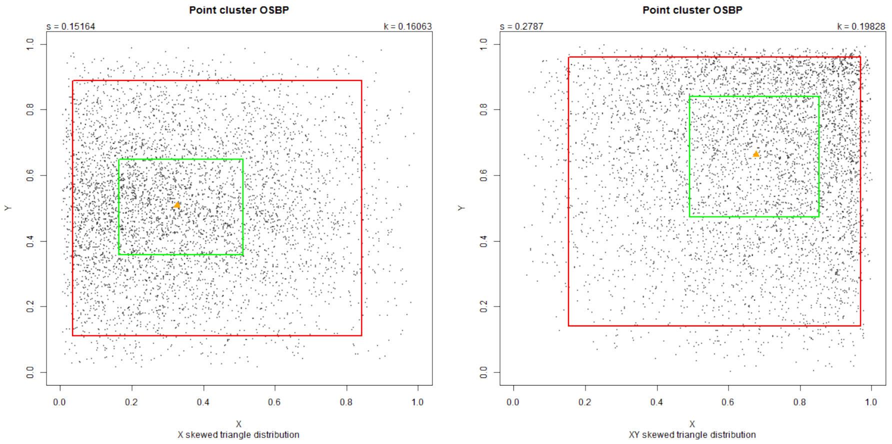
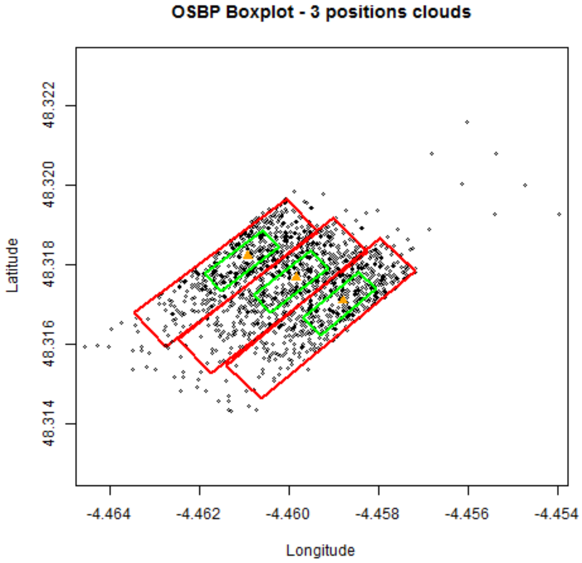


Figure 21 OSBP of 3 consecutive point clusters.



is useful to see how the statistical parameters of point clusters evolve over time which is relevant for point data that corresponds to moving objects.

5 Conclusion

In this article a new 2D pattern for assessing the statistical properties of point clusters, called the Oriented Spatial Box Plot (OSBP), has been defined. The new pattern extends the classical Box Plot to handle sets of 2D points. The comprehensive literature review highlights the weaknesses of existing techniques for summarizing the statistical properties of 2D correlated point data, such as clusters of spatial coordinates. The principle problem with these techniques is their inability to represent asymmetric 2D point cluster distributions. Some techniques rely on a combination of statistical model distributions to synthesize the point cluster density which is not relevant for skewed point clusters. A second problem is related to 2D correlated datasets. Most techniques only combine the results of two different statistical analyses without taking into account the data correlation. In this paper, we focused on the interesting visual properties of the Box Plot. We extended the approach for 2D correlated point data to highlight the statistical properties of point clusters. This pattern is suitable for visualizing the statistical properties of 2D point clusters such as their relative spreading or asymmetry. It also eases outlier detection. Spatial behavior can also be observed by comparing how the OSBP of point clusters changes over time (shape, spreading, symmetry). This new OSBP approach can be further enhanced to provide feedback about the number of points lying within the different boxes (inner and outer fence). Indeed, some points can be considered as outliers on both the X and Y axes. Furthermore, in some cases, the number of outliers outside the outer fence might be below the expected value of 5% and so new techniques to overcome this challenge will be investigated in future work.

References

Becketti, S. and Gould, W. (1987). Rangefinder box plots: A note. *The American Statistician*, 41(2):149–149.

Berkhin, P. (2006). A survey of clustering data mining techniques. In *Grouping multidimensional data*, pages 25–71. Springer.

Devogele, T., Etienne, L., Ray, C., and Claramunt, C. (2013). *Mobility Data: Modeling, Management, and Understanding*, chapter Part III - Mobility Applications, Maritime Applications. Cambridge press.

Ferreira, J. T. and Steel, M. F. (2006). On describing multivariate skewed distributions: a directional approach. *Canadian Journal of Statistics*, 34(3):411–429.

Getz, W. M. and Wilmers, C. C. (2004). A local nearest-neighbor convex-hull construction of home ranges and utilization distributions. *Ecography*, 27(4):489–505.

Goldberg, K. M. and Iglewicz, B. (1992). Bivariate extensions of the boxplot. *Technometrics*, 34(3):307–320.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323.

Kenward, R. (1987). *Wildlife radio tagging: equipment, field techniques and data analysis*. Academic Press London.

Lefever, D. W. (1926). Measuring geographic concentration by means of the standard deviational ellipse. *American Journal of Sociology*, pages 88–94.

Mohr, C. O. (1947). Table of equivalent populations of north american small mammals. *American midland naturalist*, pages 223–249.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

Potter, K. (2006). Methods for presenting statistical information: The box plot. *Visualization of Large and Unstructured Data Sets,(LNI)*, 4:97–106.

Rousseeuw, P. J., Ruts, I., and Tukey, J. W. (1999). The bagplot: a bivariate boxplot. *The American Statistician*, 53(4):382–387.

Small, C. G. (1990). A survey of multidimensional medians. *International Statistical Review/Revue Internationale de Statistique*, 58:263–277.

Thériault, M., Claramunt, C., and Villeneuve, P. Y. (1999). A spatio-temporal taxonomy for the representation of spatial set behaviours. In *Spatio-temporal database management*, pages 1–18. Springer.

Tongkumchum, P. (2005). Two-dimensional box plot. *Songklanakarin Journal of Science and Technology*, 27(4):859–866.

Tukey, J. (1977). *Exploratory Data Analysis*. Addison-Wesley.

Wickham, H. and Stryjewski, L. (2011). 40 years of boxplots. *American Statistician*.

Worton, B. J. (1989). Kernel methods for estimating the utilization distribution in home-range studies. *Ecology*, 70(1):164–168.