State of the Art in Patterns for Point Cluster Analysis

Laurent Etienne[1], Thomas Devogele[2], Gavin McArdle[3]

[1] French Naval Academy Research Institute, Brest, France
`laurent.etienne@ecole-navale.fr`

[2] Université François Rabelais de Tours, Laboratoire d'informatique, Blois, France
`thomas.devogele@univ-tours.fr`

[3] National Centre for Geocomputation, National University of Ireland Maynooth, Maynooth, Co. Kildare, Ireland
`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. In this article, we focus on different techniques used to summarize and visualize 2D point clusters and discuss their relative strengths. This article focuses on patterns which describe the dispersion of data around a central tendency. These techniques are particularly beneficial for detecting outliers and understanding the spatial density of point clusters.

1 Introduction

Visualizing and analyzing collections of very large datasets which contain massive amounts of data is now a real problem. A common first step is to identify and define data clusters. Each group, called a cluster, consists of objects that are similar to one another and dissimilar to objects of other groups [2]. A second classical step is to summarize and describe the contents of each cluster using statistical or artificial intelligence tools. These tools produce more compact graphical representations of the data known as patterns. These patterns are useful for different problems such as:

- Visual analysis of data to compare the patterns of different clusters
- Outlier detection where data not included in a pattern can be identified
- Better 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 crowdsourcing have dramatically increased the volume of spatio-temporal data. 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 [2] [6]. Several patterns and statistics can be defined to describe the data contained in 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 home range kernel density [17], combine these approaches to describe the density of animals in their home range (the area where an animal lives and travels). 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 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 (two different populations for example) or at different periods
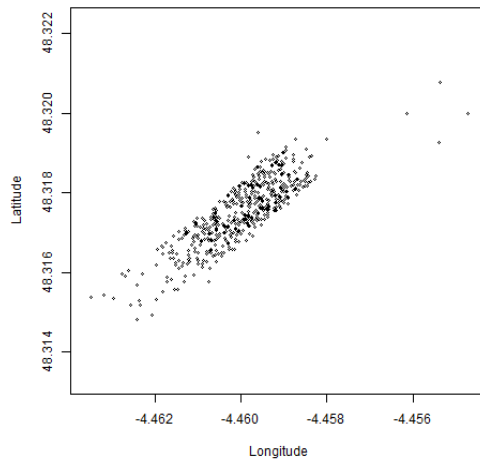— Detection of breakpoints where statistical behaviour changes

Fig. 1. A point cloud

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 then 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 handle point clusters. To illustrate and compare these techniques and processes, we apply each one to the same 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. Some points can immediately be identified as outliers. For such point clusters, it is common to add a boundary of normality around certain points, called a fence, which permits the online classification of new points as they are added to the dataset. In this paper, we present some classical methods for defining this area of normality for 2D point datasets, discuss the relative benefits of each and present a new approach, which adapts the traditional box plot for use with 2D point data. This article focus on cluster of points with regular shape without holes and large concavities.

The remainder of this paper is organized as follows. Section 2 presents patterns used to describe the normal distribution. In Section 3, patterns employed to analyze an asymmetric distribution are introduced. Section 4 describes extensions to the traditional Box Plot for analyzing 2D data. Section 5 presents several hybrid patterns which are also used for 2D data. In Section 6, the Oriented Spatial Box Plot (OSBP) which describes clusters of spatial points by extending the traditional Box Plot is presented. Finally, Section 7 summarizes the research and presents some interesting areas for future research.

2 Patterns for a normal distribution

The normal distribution is the most prominent probability distribution in statistics which describes the central tendency of data. 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, 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 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. 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.

Table 1. Covariance matrix of the point cluster

var(X)	covar(X,Y)	➜	$10.0421 \ 10^{-7}$	$7.2086 \ 10^{-7}$
covar(X,Y)	var(Y)		$7.2086 \ 10^{-7}$	$7.1730 \ 10^{-7}$

For the example points shown 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.

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

A Standard Deviational Ellipse (SDE) or a Cross of Dispersion [8] extends the normal distribution for two values. This 2-Dimensional 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 length of the two axes is 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.a gives an example of an SDE.
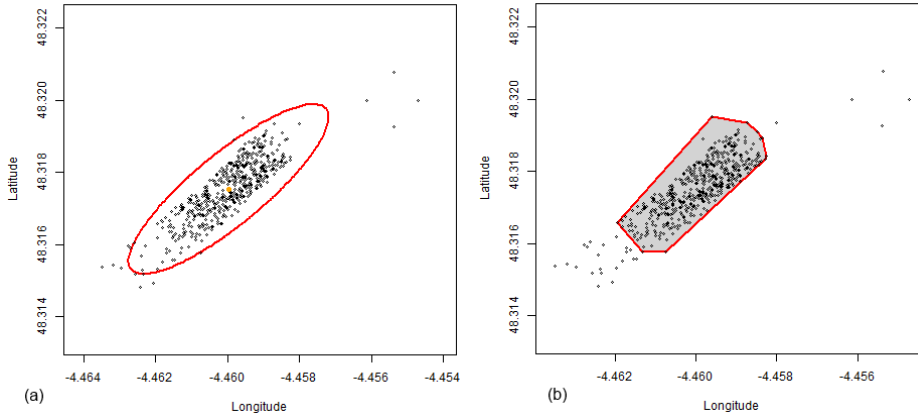

Fig. 2. (a) Standard Deviational Ellipse (SDE) for the point cluster
(b) Minimum convex polygon (MCP) for the point cluster (95%)

3   Usual patterns for asymmetric distribution

The SDE is not a robust method; a SDE can only be used with symmetric distributions 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 central tendency estimator and median is preferred. In the same way variance must be replaced by quartile. To measure the symmetry of distribution, skewness is employed. 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 shown 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 be studied at the same time. Nevertheless, these two values show that the point distribution is asymmetric.
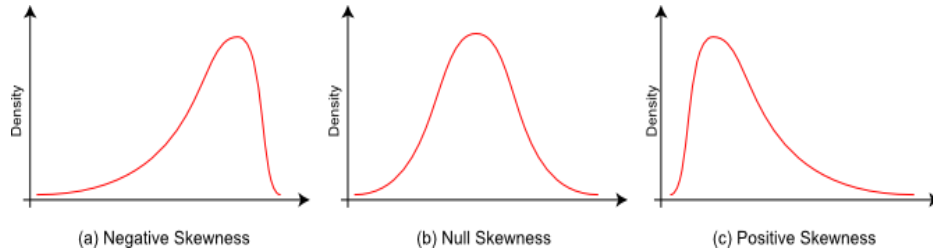

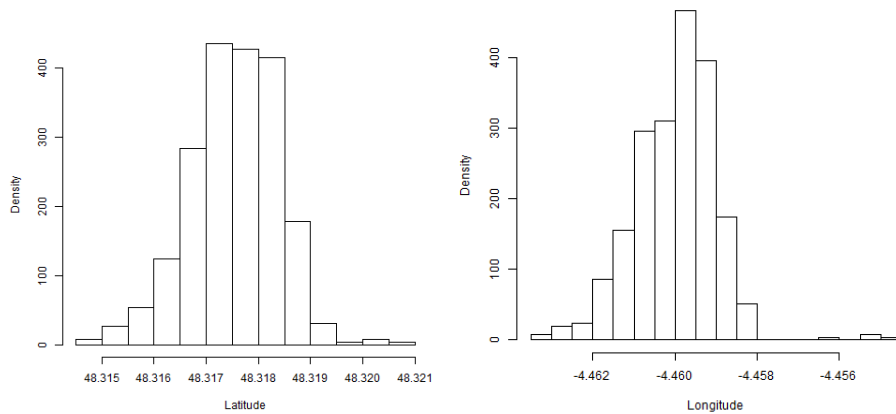**Fig. 3.** . Skewness and shape of the distribution

Fig. 4. Histograms of longitude and latitude

The second class of pattern often used to summarize a point cluster is a Minimum Convex Polygon (MCP) [9]. The MCP is the smallest convex polygon around the dataset. Figure 2.b 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 [7]. Another method called Nearest Neighbour Convex Hulls (NNCH) aggregates points. For example,[4] 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. A large subset of points could 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 [15].

4 Box Plot Extensions

For one dimensional 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 [14] is a useful descriptive pattern. This kind of distribution is very common for spatial data points.
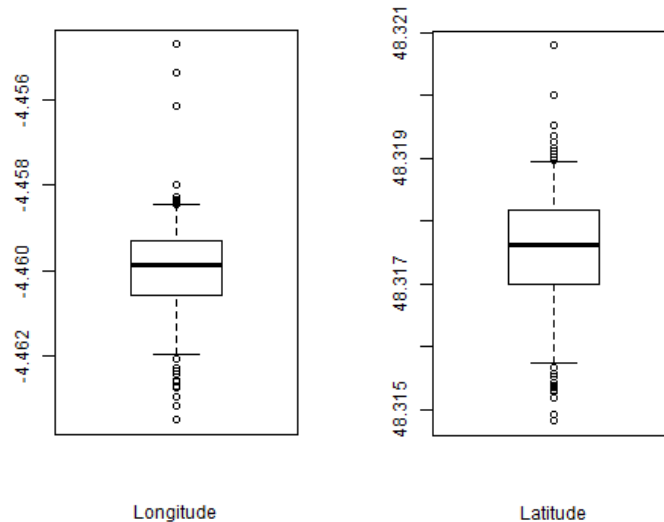

Fig. 5. A Box Plot of one dimension dataset

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. IQR is a measure of dispersion and is equal to the difference between the upper and lower quartiles (Q3 − 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 5 shows the Box Plot bounded by Q1 and Q3, with median values inside the box and the whiskers.
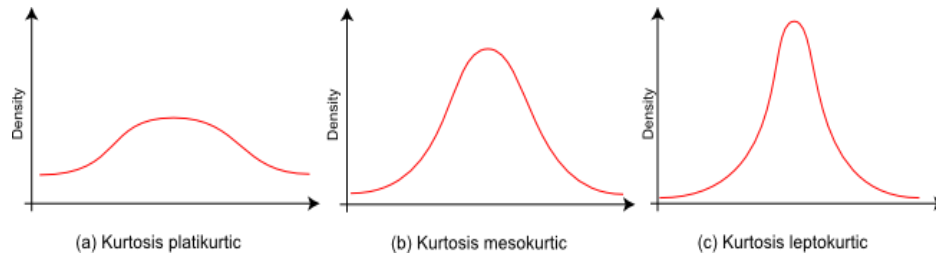
(a) Kurtosis platikurtic   (b) Kurtosis mesokurtic   (c) Kurtosis leptokurtic

Fig. 6. . Shape of the distribution and Kurtosis

Box Plots are useful for visualizing and comparing the shape and "peakedness" of a distribution which is generally described by Kurtosis. Kurtosis is classed as leptokurtic with a high Kurtosis distribution (value >1) characterized by sharper peaks and longer, fatter tails (figure 6.c). Reciprocally, a low kurtosis distribution (value <1) has a more rounded peaks and shorter, thinner tails and the kurtosis is described as platykurtic (figure 6.a). The Kurtosis value of normal distribution is 1 and characterized as mesokurtic (figure 6.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.
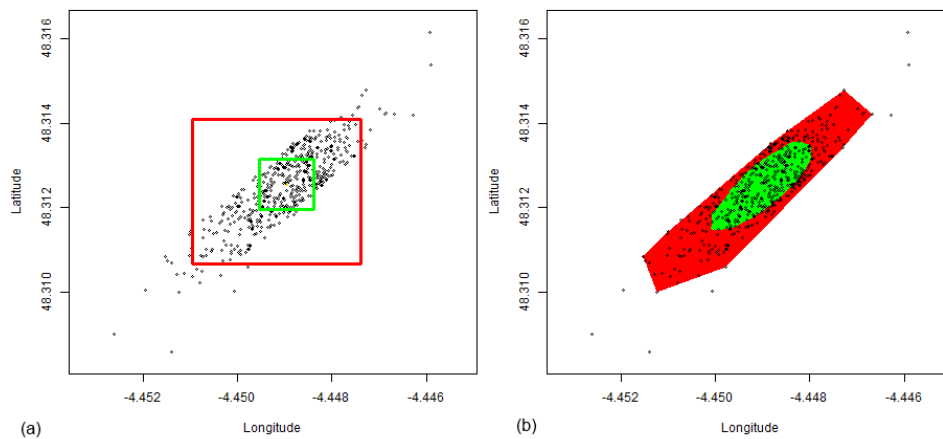

Fig. 7. (a) The two rectangles and one median point defined by the Rangefinder Box Plot method (b) Bagplot for the point cluster

The Box Plot pattern can be extended to several dimensions (Rangefinder Box Plot [1]), only if values of each dimension are not correlated. Unfortunately, if coordinates are correlated, the N-dimensional Box Plot is too large. Figure 7.a 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 fit very well for correlated point coordinates.

5   Hybrid Patterns for Analyzing 2D Points

To improve these classical 2D patterns, their respective advantages have been combined to provide mixed solutions. For example, a Bagplot [12] summarizes the cluster by a depth median consisting of a bag containing 50% of points and a fence (magnifying the bag by a factor 3). Figure 7.b gives an example of a Bagplot applied to the original set of points described in figure 1. This representation mixes the MCP and Box Plot approaches and visualizes the location, the shape, the spread, the correlation and the skewness. Unfortunately, two large sets of points are necessary to define the bag and the fence which make it an impractical approach which is difficult to manipulate.

Similar to the Bagplot, the Quelplot (a quarter elliptic plot) [5] replaces the ellipses by four separate quarter ellipses matched on their major and minor axes. A Quelplot is a compact pattern to summarize an asymmetric dataset. However, this approach estimates parameters to compute the hinge and the fence according to statistical model assumptions, rather than the real dataset. Asymmetric real data distribution is modelled as a complex deformation of a normal distribution. Figure 8.a shows the Quelplot for the example point cluster.
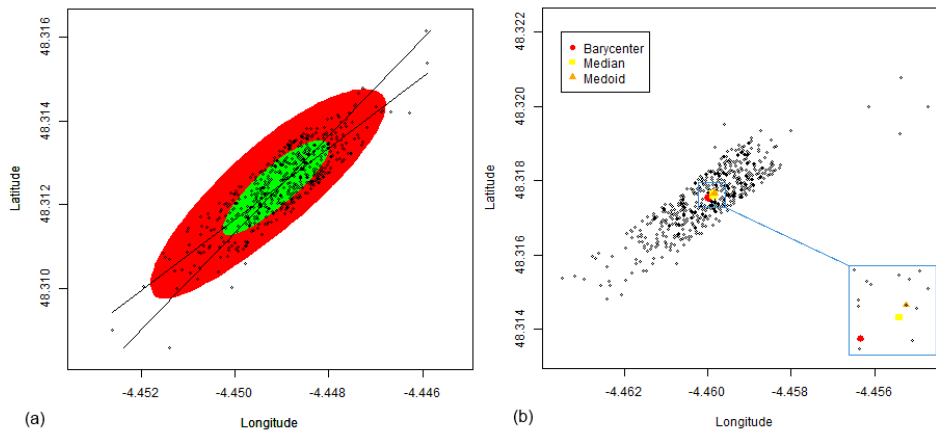

Fig. 8. (a) Quelplot for the point cluster
(b) Barycenter, Medoid and Geometric median for the point cluster

6   Box plot extension and Oriented Spatial Box Plot (OSBP)

A Box Plot is the most popular pattern to summarize a distribution. Several extensions of the Box plot have been proposed [11] [18]. Unfortunately, the Box Plot is not suited for 2D data. [16] proposes a **Bivariate Box Plot**, defined by a pair of trapeziums oriented in the direction of a single fitted axis. Unfortunately, trapeziums are defined according to the X-axis. 2 edges are parallel to the y-axis. For points, the X-axis and the Y-axis are equally important. If the X-axis is switched with the Y-axis, the resulting pattern is very different. This article defines a new 2D Box Plot. As other 2D Box Plots, this pattern translates the 5 summary values to 2D including:

- Central tendency
- A central rectangular box that encompasses 50% of the data
- A rectangular fence that separates the normal data from potential outliers

This Oriented Spatial Box Plot (OSBP) minimizes the size of these two rectangles. The orientation of these rectangles is defined using the variance between coordinates. Figure 9 provides an overview of the steps used to compute the OSBP[1], while the following sections describe the process in more detail. A maritime application of OSBP is proposes in [3].
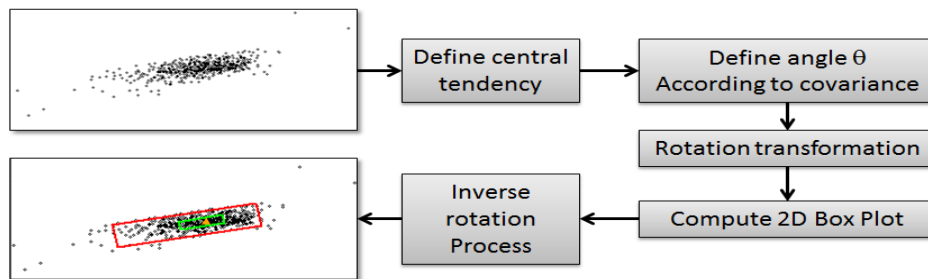

Fig. 9. Process used to define the Oriented Spatial Box Plot

6.1 Central tendency

In the 1-D Box Plot concept, the central tendency [13] of the dataset is represented using the median value. For 2D values, three different approaches can be used to define and visualize this tendency. First of all, the mean point of all the points in all of the coordinate directions is called the Barycenter. It is computed using the arithmetic mean of all the points of the cluster and so 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.

Finally, another way 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 8.b 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.

---

[1] The following patterns have been computed using the R environment for statistical computing and graphics display (www.r-project.org)

6.2 Define angle

Edges of the classical 2D Box Plot are parallel to an axis. The four edges of Rangefinder Box Plot are parallel to the axes while for the Bivariate Box Plot only two axes are parallel to the X axes. To minimize the size of the 2D Box Plot and to fit the two rectangles to the shape formed by the points of cluster, this parallel constraint is removed. So, the orientation of rectangles needs to be defined.
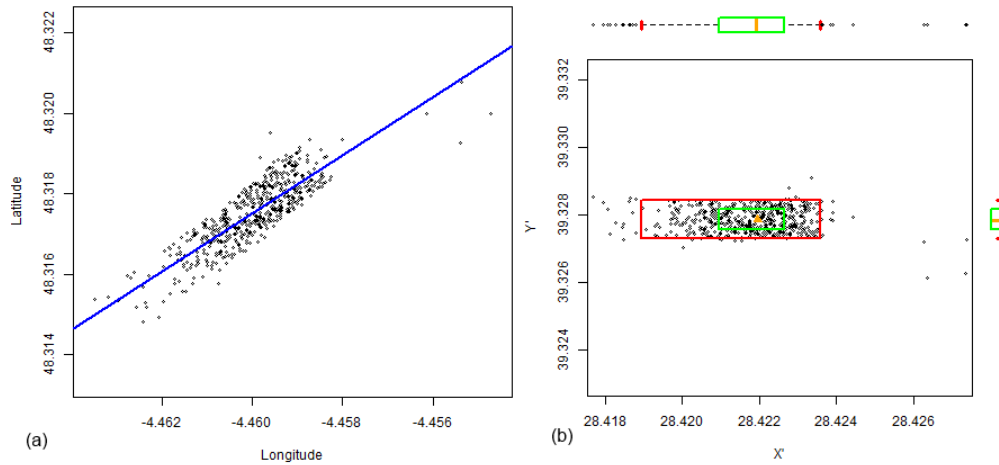

Fig. 10. (a) The best fit straight line for the point cluster of the example (b) 2D Box Plot to the cluster of rotated points

Pearson's correlation coefficient (ρ) between coordinate values is chosen as the key factor. This coefficient is easy to compute and frequently used for this kind of problem, for example, it is used for Principal Component Analysis (PCA). Indeed, the original concept of Pearson [10] was to determine a straight line to represent the best fit for a set of data points. In this Oriented Box Plot problem, the idea is to define a rectangle which best fits the points of a specific cluster. This coefficient is symmetric ($\rho_{X,Y} = \rho_{Y,X}$). Unfortunately, this coefficient is not robust, if a large set of outliers are present. Nevertheless, for common point clusters, the number of outliers is generally small. To translate the Oriented Box Plot problem to a Rangefinder 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).

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

For the example, the value of Θ is 11.56°, Figure 10.a shows the best fit straight line for this point cluster.

6.3 Rotation transformation

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

$$x' = x \cos(\theta) - y \sin(\theta) \tag{3}$$

$$y' = x \sin(\theta) + y \cos(\theta) \tag{4}$$

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.

6.4   Compute 2D Box Plot

In this step, the Box Plot as x' and the Box Plot y' are separately computed. The process is the same as the Rangefinder Box Plot process with 2.5% of outlier data. When 2.5% are chosen:

— 1.25% of outlier point is above of the fence rectangular
— 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 rectangular. This 2D Box Plot is similar to the Rangefinder Box Plot. The edges of the two rectangles are parallel to X' and Y' axis. Figure 10.b shows the Box Plot for the X' axis (above the plot) and the Box Plot for the Y' axis (right of the plot). These two Box Plots are combined to define the fence and the central rectangles. The medoid which is a real point of the cluster is used to define the central tendency.

6.5   Inverse rotation process

The last step is straightforward and 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 using these formulae:

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

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

In figure 11.a, the result of the inverse rotation transformation on the two rectangles is shown.

Overall, it has been demonstrated how these steps extend the Box Plot concept to two dimensional data. An OSBP has the same properties for 2D data as the Box Plot for 1D data. The OSBP pattern summarizes and visualizes important information about the statistical distribution of a point cluster. The central tendency is enhanced using the medoid value. The surface of the rectangular fence effectively summarizes the spreading of the cluster. The shape or the "peakedness" of the distribution can be observed by examining the ratio between the surface of the fence and the surface of central rectangle (surface (central rectangular) / surface (fence rectangular)). If this ratio is close to 1, a large part of data is grouped around the medoid and the distribution tails could be longer. If this ratio is close to 0, the distribution is more uniform. Moreover, the distance

between the edges of the two rectangles (d1, d2, d3, d4) can be used to observe the asymmetric criteria. Figure 11.b shows that distances d1 and d2 are different. For X', the distribution is asymmetric but for Y', the distances d3 and d4 are very close indicating the distribution is symmetric. In the same way, distance between medoid and edges can also be employed to determine the skewness and kurtosis of the point cluster. Finally, it is easy to identify outlier points which are plotted outside the rectangular fence.
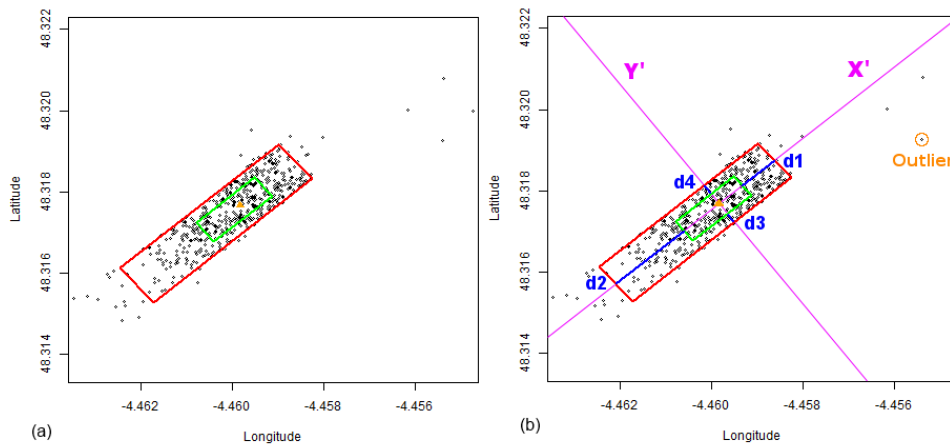

Fig. 11. (a) Oriented Spatial Box Plot for the point cluster
(b) Distances between edges of rectangles

Within this process, one problem remains. The values of 50% of data outside the central rectangular and 5% outside the fence rectangular are upper estimates. Indeed, an outlier could for example, fall below the fence rectangle (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 orange circle of figure 11.b. Empirically, the outlier percentage is closer to 4.5% than 5%.

6.6   Taking time into account as a third dimension.

The OSBP concept can be extended to 3D datasets such as positions (x,y,t) clouds. Considering time (t) as a third dimension uncorrelated with the spatial one (x,y), another 1D time Box Plot can be separately computed for the positions dataset. The time dimension can then be visualized in 3D as the Z axis. This visualisation is usually called a space-time cube. The inner and outer fence of the OSBP and the time Box Plot are merged together. It produces a new 3D boxplot named OSTBP (Oriented Spatio-Temporal Box Plot) illustrated in figure 12. The inner green parallelepiped is fully included into the outer red parallelepiped fence. As explained in section 6.1, the central position of the OSBP shown in orange in figure 12 is not a real position of the dataset as it corresponds to the geometric median merged with the median time. The OSTBP gives

interesting visual feedback about the spatial and temporal density spreading and symmetry around a central tendency.
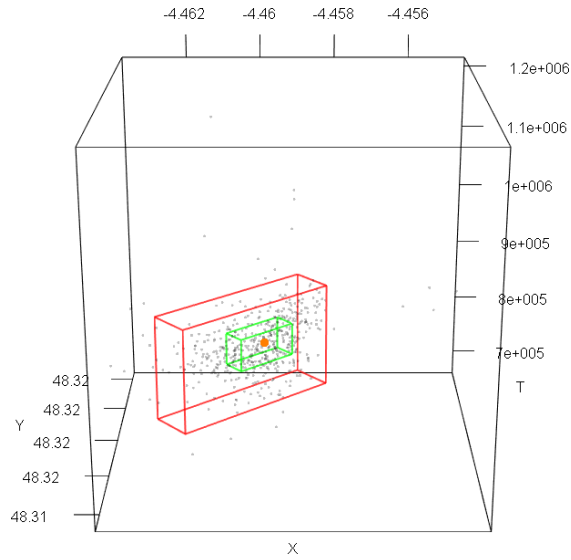

Fig. 12. Oriented Spatio-Temporal Box Plot for a position cluster

7   Conclusion

In this article several point patterns have been described. Table 2 summarizes each pattern by describing the data which each approach can effectively analyze. The table shows the classical solutions, SDE and MCP cannot summarize an asymmetric correlated coordinate cluster.

Table 2. Comparative summary of SDE, MCP and Box Plot

	SDE	MCP	Rangefinder Box Plot	Bagplot	Quelplot	OSBP
Central point	+	-	+	+	+	++
Asymmetry distribution	-	+	+	++	++	++
Kurtosis (shape)	-	-	-	+++	+	++
Compact pattern	++	-	++	-	+	++

Similarly, some techniques rely on a combination of statistical model distributions to synthesize the point cluster density which is not relevant for skewed point clusters.

Generally, the techniques fail to fully consider 2D correlated datasets, such as spatial data which poses an additional challenge. The Quelplot and the OSBP are interesting patterns which do address these requirements. In this paper, we focus on examining the interesting visual properties of the Box Plot concept to highlight the statistical properties of the numeric variables and extend the approach for 2D correlated point data using a new concept, the OSBP. The OSBP concept has been extended to 3D positions with time as the third dimension (OSTBP). The spatial and temporal behaviour can then be observed by comparing how the OSTBP changes over time (shape, spreading, symmetry). Consecutive 3D OSTBP can be connected together in order to generate a new OSTBP spatio-temporal pattern that synthesizes the spatio-temporal behaviour of a group of trajectories [3].

8   References

1. Becketti, S., Gould, W. Rangefinder Box Plots, A Note, The American Statistician, (1987)
2. Berkhin, P. A survey of clustering data mining techniques. In Grouping multidimensional data Springer Berlin Heidelberg. (2006) 25-71
3. Devogele, T., Etienne, L., Ray, C. Mobility data: Modeling, management, and understanding, Part III, Chap. Mobility Applications, Maritime Applications, In Renso, C. Spaccapietra, S. Zimanyi, E. (Eds.), Cambridge press (2013) 224-243
4. Getz, W. M., Wilmers, C. C. A local nearest-neighbor convex-hull construction of home ranges and utilization distributions. Ecography, Vol. 27, (2004) 489-505.
5. Goldberg, K., Iglewicz, B. Bivariate Extensions of the Box Plot. American Statistician, Vol. 34 num. 3 (1992) 307–320
6. Jain, A. Murty M, Flynn, P. : Data clustering: a review, ACM Computing Surveys (CSUR) Vol. 31 num. 3 (1999) 264-323
7. Kenward, R. Wildlife radio tagging. Academic Press, Inc., London, United Kingdom. (1987)
8. Lefever, D. W. Measuring Geographic Concentration by Means of the Standard Deviational Ellipse, American Journal of Sociology, Vol. 32 num. 1 (1926) 88-94.
9. Mohr, C.O. Table of equivalent populations of North American small mammals. *American Midland Naturalist, Vol.* 37 num. 1, (1947) 223-249.
10. Pearson, K. On Lines and Planes of Closest Fit to Systems of Points in Space, Philosophical Magazine, Vol. 2 num. 11 (1901) 559–572.
11. Potter, K., Hagen, H., Kerren, A., & Dannenmann, P. Methods for presenting statistical information: The box plot. Visualization of Large and Unstructured Data Sets, (LNI), Vol. 4 (2006) 97-106.
12. Rousseeuw, P., Ruts, I. Tukey, J. The bagplot: a bivariate boxplot. The American Statistician, Vol. 53 Num. 4 (1999) 382-387
13. Small, C. A. Survey of Multidimensional Medians, International Statistical Review, Vol. 58, International Statistical Institute (ISI), Num. 3 (1990) 263-277
14. Tukey, J. Exploratory Data Analysis, Addison-Wesley (1977)
15. Thériault, M., Claramunt, C., Villeneuve, P. A Spatio-Temporal Taxonomy for the Representation of Spatial Set Behaviours, Lecture Notes in Computer Science 1678 (1999) 1-18.
16. Tongkumchum, P. Two-dimensional box plot, Songklanakarin J. Sci. Technol. Vol. 27, num. 4 (2005) 859-866.
17. Worton B. J. Kernel Methods for Estimating the Utilization Distribution in Home-Range Studies. Ecology, Vol. 70, (1989) 164–168.
18. Wickham, H., Stryjewski, L. 40 Years of Boxplots. Am. Statistician (2011)