Trajectory Box Plot: a new pattern to summarize movements

Laurent Etienne, Thomas Devogele, Maike Buckin, Gavin Mcardle

International Journal of Geographical Information Science
Vol. 00, No. 00, January 2015, 1–20

RESEARCH ARTICLE

Trajectory Box Plot; a New Pattern to Summarize Movements

Laurent Etienne[a*], Thomas Devogele[a], Maike Buchin[b], Gavin McArdle[c]
[a] *University of Tours, Blois, France*;
[b] *Faculty of Mathematics, Ruhr University, Bochum, Deutschland*;
[c] *National Center for Geocomputation, Maynooth University, Ireland*;
(*v2.1 sent July 2015*)

Nowadays, an abundance of sensors are used to collect very large datasets of moving objects. The movement of these objects can be analysed by identifying common routes. For this, a cluster of trajectories must be defined and the pattern of each cluster discovered. In this article, we introduce a new pattern, called the Trajectory Box Plot (TBP), to summarize a set of trajectories following the same route. The TBP is an extension of the well known descriptive statistics Box Plot concept. Each TBP is described by a median trajectory, a 3D box and a 3D fence. The median trajectory depicts the typical movement of mobile objects. The box and the fences (whiskers) describe the spatial and temporal spreading around the central tendency. Trajectory Box Plots are useful to summarize and analyse trajectory streams, understand their spatio-temporal density and detect outliers. In this article, visual analysis highlights how the Trajectory Box Plot pattern effectively describes how the density of trajectory clusters change over time.

Keywords: Box Plot, spatio-temporal pattern, position clusters, trajectory analysis, Fréchet distance

*Corresponding author. Email: laurent.etienne@univ-tours.fr

ISSN: 1365-8816 print/ISSN 1362-3087 online

DOI: 10.1080/1365881YYxxxxxxxx
http://www.informaworld.com

1. Introduction

Nowadays, an abundance of sensors such as GPS and tracking technologies are used to collect the positions of moving objects. The development of monitoring systems and the emergence of crowd-sourcing have dramatically increased the volume of such spatial data. These big spatial datasets are difficult to manage, visualize and understand. This problem is exacerbated given the wide range of mobile objects which can move in different low-constrained open spaces (Renso *et al.* 2013), such as animals (Lee *et al.* 2007, Freeman *et al.* 2011), pedestrians (Tchetchik *et al.* 2009), ships (Etienne *et al.* 2012), planes (Hurter *et al.* 2009) or even Human Computer Interaction (HCI) movements on a computer screen (Tahir *et al.* 2011). When analysing movement in these situations, trajectories can be grouped together using various clustering techniques in conjunction with similarity measures. The resultant clusters can then be summarized into patterns describing the usual behaviour of the trajectory set.

In this article, we focus on analysing trajectories of the same type of mobile objects with the same itinerary. Summarizing the spatial and temporal distribution of a trajectory set is useful for obtaining a succinct description of the set. However, producing a meaningful summary is difficult when the cluster of trajectories is large. In particular, three issues arise. Firstly, a representative trajectory which depicts the typical movement performed by all trajectories in a particular set or cluster (central tendency) is required. Secondly, the spatio-temporal dispersion of the trajectory set around the central tendency needs to be quantified. Thirdly, an effective visualization which provides feedback to the analyst about the central tendency, spatio-temporal distribution and symmetry also needs to be defined without creating cognitive overload. We address these issues in this article through the development of a visualisation pattern called the Trajectory Box Plot (TBP). This new pattern is a temporal extension of 2D point patterns.

The concept of the Trajectory Box Plot is useful for summarising and visualising the behaviour and trajectories of a set of mobile objects with the same itinerary and introduces several key benefits. For example, outliers, which are trajectories with spatial or temporal properties that differ significantly from other trajectories within the same set (Lee *et al.* 2008), can be easily detected. The pattern can quickly classify new trajectories as members of existing clusters. Similarly, the concept can be used to compare the properties of sets containing different mobile objects. Finally, the pattern can help to predict, in real-time, the next position of a trajectory based on the trajectory's history (Devogele *et al.* 2013).

The remainder of this article is organised as follows. Section 2 presents some existing techniques for describing movement such as the central tendency and spatial spreading around it. These techniques are reviewed for both clusters of positions and clusters of trajectories. In Section 3, we propose an extension of the traditional Box Plot for use with patterns to describe the spatial and temporal density in clusters of trajectories. Section 4 presents this new Trajectory Box Plot applied to a real world trajectory cluster. Examples of visual analysis and outlier detection illustrate the expressive power of the Trajectory Box Plot pattern (TBP). Finally, this work is summarized in Section 5 and some areas for future work are discussed.

2. Position and trajectory patterns

Spatio-temporal clustering is the process of grouping objects based on their spatial and temporal similarity (Kisilevich *et al.* 2010), which results in a collection of homogeneous groups characterised by one or more salient properties (Renso *et al.* 2013). Commonly spatio-temporal clustering is used for grouping trajectories. Patterns can be defined to summarize the temporal and spatial aspects of the cluster of trajectories. In particular, patterns extracted from a large set of trajectories are of interest. Several patterns have been defined to describe commonalities seen in clusters of trajectories. For example, for a group of objects moving together, flock (Gudmundsson and van Kreveld 2006), swarm (Li *et al.* 2010) and convoy (Jeung *et al.* 2008) are common descriptions. Similarly, for moving objects following the same itineraries (or routes), spatio-temporal sequential patterns (Cao *et al.* 2005), T-pattern (Giannotti *et al.* 2007) and partition and group patterns (Lee *et al.* 2007) are often used. Generally, to define patterns of trajectories, a cluster of positions (Gudmundsson and van Kreveld 2006, Jeung *et al.* 2008, Li *et al.* 2010) or segments (Cao *et al.* 2005, Lee *et al.* 2007) are initially computed. Then, according to these clusters, the patterns of trajectories are defined.

In this article, we focus on a trajectory cluster as a group of trajectories following the same itinerary, sharing a similar source, destination and route at different time periods. In other words, all the trajectories of the cluster start from the same area of interest, follow a similar route to an identical place but may start at different absolute timestamps. In the following sections, we define a position $P_i = (x_i, y_i, [z_i], t_i)$ as a combination of a 2D or 3D spatial point $p_i = (x_i, y_i, [z_i])$ together with a timestamp (t_i). A trajectory T of a mobile object O can be defined as a sequence of temporally ordered positions so that $T = \{P_1, P_2..., P_i, ..., P_n\}$ where P_1 stands for the initial position of the trajectory in the start area and Pn or for the last one. The timestamps of the positions of the trajectory T are relative timestamps, that is, we assume all trajectories to start at timestamp 0.

The central tendency and the spread of data around the central tendency are key elements to an effective and compact pattern. These values indicate the shape of the cluster, describe the temporal evolution of the moving objects and allow outlier positions of trajectories to be identified. It is therefore important to include these measures in a new pattern for spatial-temporal data. While these patterns are well understood for 2D data, the challenge is to produce effective and efficient techniques for calculating and visualising these patterns for spatial-temporal data such as positions and trajectories.

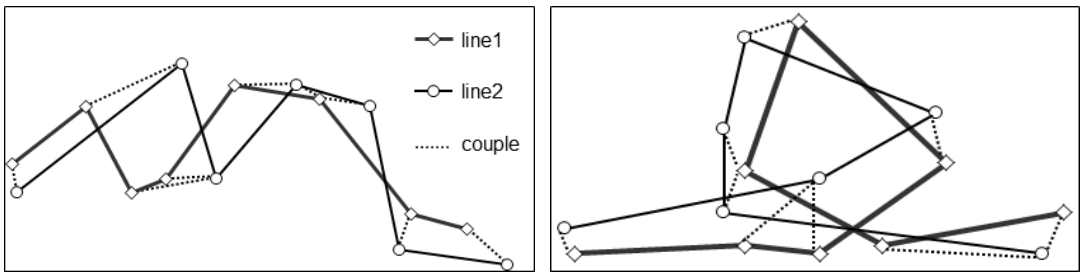
The next sections focus on deriving patterns from clusters of spatial-temporal data. Firstly, position and point patterns are introduced and then techniques for describing trajectory patterns are detailed.

2.1. *Representing the central tendency for positions and trajectories*

The central tendency efficiently represents the spatial and temporal behaviour of a point, position or trajectory cluster. When applied to moving objects such as trajectories, the central tendency provides an understanding of the main movement realized by the objects in the cluster.

2.1.1. Central point and position

The central tendency of a dataset is often represented using mean or median values. Several generalizations to higher dimensions exist (Small 1990, Bhadury *et al.* 2003): the barycenter, the geometric median and the medoid.



(a) two lines without loop (b) two lines with loop

Figure 1. Couples of matching points for two lines.

When creating a central tendency for position clusters, the time dimension must also be taken into account. The geometric median and arithmetic mean are straightforward to extend to multi-dimensional datasets, however, medoid computations require unique spatio-temporal similarity measures that are more complex to define. There are currently no methods which take into account the spatial and temporal dimensions simultaneously.

2.1.2. Central trajectory

Generalizing the techniques used for points in Section 2.1.1 for use with trajectories is not straightforward, since trajectories are ordered sequences of time-stamped positions. Several methods have been proposed previously (Buchin *et al.* 2010, Etienne *et al.* 2010, Petitjean *et al.* 2011, Chen *et al.* 2013). Most approaches for computing a central trajectory are based on a similarity measure between two trajectories. Li (2014) gives an interesting state of the art on trajectory similarity measure. Simple measures such as a perpendicular measure (Etienne *et al.* 2010) compute the distances between points and their matching points of the other trajectory. Perpendicular measures are fast to compute for smooth trajectories but not robust for convoluted or asymmetric trajectories. More complex measures are based on edit distances such as Edit distance with Real Penalty (EDRP) or Edit Distance on Real sequence (EDR) defined in (Chen and Ng 2004). The distance between lines with time shifting are also widely used, examples include Dynamic time warping (DTW) (Sakoe and Chiba 1978, Berndt and Clifford 1994) and Discrete Fréchet Distance (Eiter and Mannila 1994, Devogele 2002).

DTW minimizes the sum of distances of coupled points, whereas the Discrete Fréchet distance minimizes the maximum distance of coupled points. These two methods are robust for trajectories with loops or sinuous lines with shifts. Both techniques align the trajectory positions in order to minimize the spatial distances. They rely on the computation of a distance matrix between every pair of positions of the two trajectories. Assuming that each trajectory has N positions, the complexity of these algorithms is quadratic $O(N^2)$. Figure 1 illustrates the alignments of two trajectories. DTW minimizes the sum of the distances between matched positions of two trajectories. When applied to a trajectory set, it raise a problem as the trajectories of the set may have significant different length. The method used in DTW is not a good indicator and a maximal distance such as Discrete Fréchet Distance is preferred.

Once a similarity measure between trajectories is chosen, it can be used to compute a distance matrix between all the trajectories of a cluster. The trajectory minimising the distance to all other trajectories of the cluster can be considered as the central one. The main problem with this approach is its complexity as each trajectory must be compared with all other trajectories of the cluster. For a cluster having M trajectories composed of N positions, the complexity of the algorithm using the Discrete Fréchet distance or

DTW is $O(M^2N^2)$.

In order to reduce this complexity, Petitjean *et al.* (2011) and Ariza-López *et al.* (2015) propose different optimized processes. Petitjean *et al.* (2011) define an iterative process which relies on the definition of a reference trajectory compared to every trajectory of the cluster. The complexity of this comparison step is $O(MN^2)$. Each position of the reference trajectory is paired with positions of other trajectories in the cluster. The result of this matching process is an ordered set of positions (S). Central positions of each set can then be computed using the techniques presented in the previous section. The complexity of this step is $O(MN)$. The ordered set of central positions are then connected together to generate a new reference trajectory. The process is applied iteratively until the reference trajectory converges to a central trajectory.

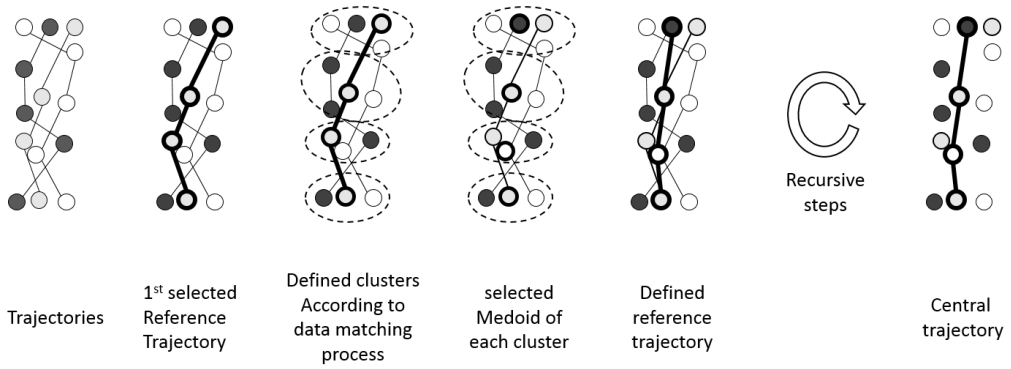


Figure 2. The main steps to compute central trajectory.

Figure 2 illustrates these main steps on a very simple example. A central trajectory is computed for three different trajectories. The reference points and reference trajectory are shown in bold. The dotted lines encompass point clusters. Several points from the same trajectory can be in the same cluster. The initial reference trajectory is selected randomly from the cluster, however some heuristics can ease the iterative process convergence. For instance, the initial reference trajectory can be selected among the trajectory having median time duration, length or speed. The number of iterations (I) of the process is expected to be much smaller than the number M of trajectories in the set. The overall complexity of this algorithm is then $O(IMN^2) << O(M^2N^2)$.

Ariza-López *et al.* (2015) use a pairwise process in two steps. In the first step, homologous positions between trajectory pairs are matched using the discrete Fréchet distance, then, the mean position between each couple of matched positions is computed Devogele (2002). In the second step, the set of trajectories is replaced by the new set of mean trajectories. The process is iterated until only one mean trajectory remains. The overall complexity of this algorithm is $O(2MN^2)$. However, the results of this algorithm may change depending on the initial order used to define trajectory pairs.

Several other approaches for computing a central trajectory have also been proposed. Etienne *et al.* (2010) suggest using the median of positions at equal times. This method works well if the trajectories are very similar and equally sampled, and it has a low running time $O(MN)$ for M trajectories equally sampled at N positions. However, it cannot effectively handle trajectories which contain loops. Buchin *et al.* (2010) suggest computing a median trajectory that is defined topologically: starting at a common start point, try to always stay in the middle by switching at intersections where homotopy

is maintained. This approach can handle collective loops (i.e., loops carried out by all trajectories in a set) which are recognizable in the data. It cannot, however, deal with other loops, and it has a high running time of roughly $O(M^2N^2)$. van Kreveld and Wiratma (2011) proposed the majority median, which is built from input edges that are close to many other input edges. This algorithm also has a high asymptotic running time of $O(M^3N^3logMN)$. Har-Peled and Raichel (2014) suggest computing a curve that minimizes the weak Fréchet distance to all input edges. This results in a running time of $O(N^M)$, that is exponential in the number of trajectories. Lee *et al.* (2007) propose an approach based on a partition of the trajectories into segments. A representative trajectory is computed as an average of similar segments.

These approaches differ in several ways. For one, the temporal and spatial dimensions are handled differently. Some approaches (Buchin *et al.* 2010, van Kreveld and Wiratma 2011, Lee *et al.* 2007, Petitjean *et al.* 2011) ignore the temporal dimension, and work only in the spatial domain. Other approaches (Etienne *et al.* 2010, Har-Peled and Raichel 2014) take both dimensions into account. Furthermore, some approaches (Buchin *et al.* 2010, Etienne *et al.* 2010, van Kreveld and Wiratma 2011) use vertices or edges of the input to build the central trajectory, while other approaches (Lee *et al.* 2007, Har-Peled and Raichel 2014, Petitjean *et al.* 2011) construct a new trajectory. This is similar to point clusters (Section 2.1.1), where one can use a barycenter or medoid of a set. For trajectories, we have one more choice: a trajectory of the input which globally minimizes the distance to all others, or a trajectory built from local central points. Locally central points may belong to different trajectories, which can be problematic in the case of multi-modal clusters.

2.2. *Spreading around the central tendency*

Along with the central tendency, the boundaries of the clusters are important to understand and summarize the spatial and temporal spreading of the data. To evaluate dispersion patterns around a central tendency, the following criteria are used:

- 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 coordinates of the points;
- The shape of the distribution which indicates if the distribution is symmetric or asymmetric;
- The compactness of patterns

2.2.1. Spatial spreading around a central point

Etienne *et al.* (2014) detail the state of the art patterns used to describe spreading around a central point. The most frequently used patterns are:

- Standard Deviational Ellipse (SDE) (Lefever 1926),
- Minimum Convex Polygon (MCP) (Mohr 1947),
- Rangefinder Box Plot (Becketti and Gould 1987),
- Quelplot (Goldberg and Iglewicz 1992),
- Bagplot (Rousseeuw *et al.* 1999),
- Bivariate box plot (Tongkumchum 2005),
- Oriented Spatial Box Plot (OSBP) (Etienne *et al.* 2014).

Patterns such as the Rangefinder Box Plot, Bivariate box plot and the OSBP extend principles of the traditional box plot (Tukey 1977) which is useful for representing varia-

tion in samples of statistical one dimensional datasets without making assumptions about the underlying statistical distribution. Etienne *et al.* (2014) highlighted the advantages of OSBP.

The OSBP Etienne *et al.* (2014) has the same properties for 2D data as the Box Plot for 1D data. The OSBP pattern summarizes important information about the statistical distribution of a cluster of points. Within this context, the central tendency is enhanced using the median value while a central rectangular box encompasses 50% of the data. Similarly, a rectangular fence separates the normal data from outliers.

The OSBP and the other patterns typically focus on point clusters and lack the ability to consider the important temporal dimension of position clusters (trajectories). In Section 3.2.1, a new pattern, that extends the OSBP for position clusters, is introduced.

2.2.2. Spreading around central trajectory

Position patterns are effective for describing a cluster of positions. In the same way, patterns for trajectories must be defined. Classically, heat maps (Wilkinson and Friendly 2009) are the 2D visualisation used to represent the density of trajectories on a map. Demšar *et al.* (2015) proposes the stacked space time densities. This geovisualisation method extends the space time cube to represent the Spatio-temporal density of a large aggregate set of trajectories using a 3D (x, y, t) voxel structure. For each voxel, the trajectory densities are computed. The colour and transparency of voxels are defined according to density, to visualise this spatio-temporal representation. The set of coloured voxels show the density but also the spread of a trajectory set. This representation is very effective for spatio-temporal analysis and visualization. Unfortunately, for statistical approaches such as outlier detection or prediction this model is not suitable as the central tendency is not easy to define from the stacked space time densities.

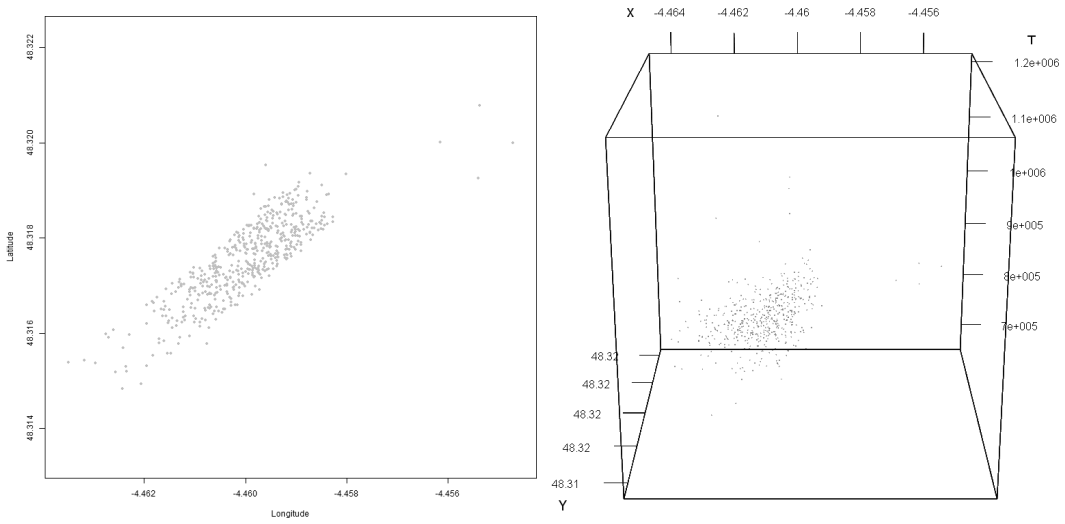
The curve boxplot (Mirzargar *et al.* 2014) extends the box plot to a set of curves. A trajectory can be defined as a curve where x and y values are defined according to the timestamp value. This model uses the band depth concept which delimits a region according to a subset of the ensemble members. A position (x, y, t) is inside the band depth if (x, y) is inside a triangle defined by 3 points of curve of the band depth at the same timestamp t. In the same way, a trajectory is fully contained in this band if all positions are inside the band. Using this band depth concept, the median curve (curve inside the maximum number of band depth) and Box Plot (defined with the percentile of curves contained inside a band depths) can be defined. While curve Box Plots are interesting descriptive statistical tools, the number of curves cannot be large as the complexity is $O(2^N)$. Moreover, these approaches do not consider the speed variation between trajectories.

3. Trajectory Box Plot patterns

As described in previous sections, various techniques can be used to represent the spatial dispersion of position and trajectory clusters around a central tendency. One important feature of a visualisation pattern is its ability to show central tendency, dispersion and symmetry simultaneously. The Oriented Spatio Box Plot (OSBP) (Etienne *et al.* 2014) provides a means to visually understand these statistical parameters for point clusters.

In this section we present the Trajectory Box Plot (TBP) which extends the Box Plot concept from points to trajectories. A position cluster containing 519 real positions illustrates this extension (Figure 3). Each position is described by its longitude, latitude coordinates and a relative timestamp in seconds. Positions are displayed in a spherical

coordinate system. Geodesic distances are computed using Haversine formula. The TBP algorithm can be easily adapted to any coordinate reference system which allows to compute spatial metric distance between positions.



(a) 2D visualisation (b) 3D visualisation

Figure 3. Sample position cluster.

3.1. *Median trajectory computation*

The first step of the Trajectory Box Plot pattern computation consists of defining the median trajectory. The Box Plot concept uses the median to define the central tendency of a dataset. Based on the different techniques cited in Section 2.1.2, the median trajectory computation algorithm presented in (Petitjean *et al.* 2011) is selected. This algorithm is both efficient and effective. It has a reasonable running time and produces central trajectories which are representative of the cluster. To improve the result for cluster of trajectories with different lengths, DTW measure is replaced with Discrete Fréchet Distance as a similarity measure to match and align trajectory positions. This first step of the Trajectory Box Plot pattern computation generates a sequence of position clusters. The number of positions cluster is equal to the number of positions of the first trajectory. To ease the iterative process convergence and to define the number of positions of the median trajectory, the trajectory having median length is choosen as initial reference trajectory. The marginal median (Puri and Sen 1971) positions of each position cluster are computed (Section 2.1.1). The marginal median position consists of a combination of the median value of the X and Y coordinate components of the point cluster. To avoid unrealistic median positions, the nearest position of the cluster to the marginal median is selected as the median position. The ordered sequence of these median positions compose the median trajectory.

3.2. *Spatio-temporal trajectories Box Plot pattern*

A sequence of position clusters with median positions have been generated by the first step of the Trajectory Box Plot pattern to constitute the median trajectory. Each posi-

tion cluster can then be analysed to understand the spatial and temporal spreading of the positions around the median one. The TBP takes advantage of the Box Plot representation. It visually combines important distribution metrics such as the central tendency, a first boundary box which encompasses the interquartile (50% of the dataset) and a second outer boundary box beyond which data are considered as outliers.

3.2.1. Oriented Spatio-Temporal Box Plot

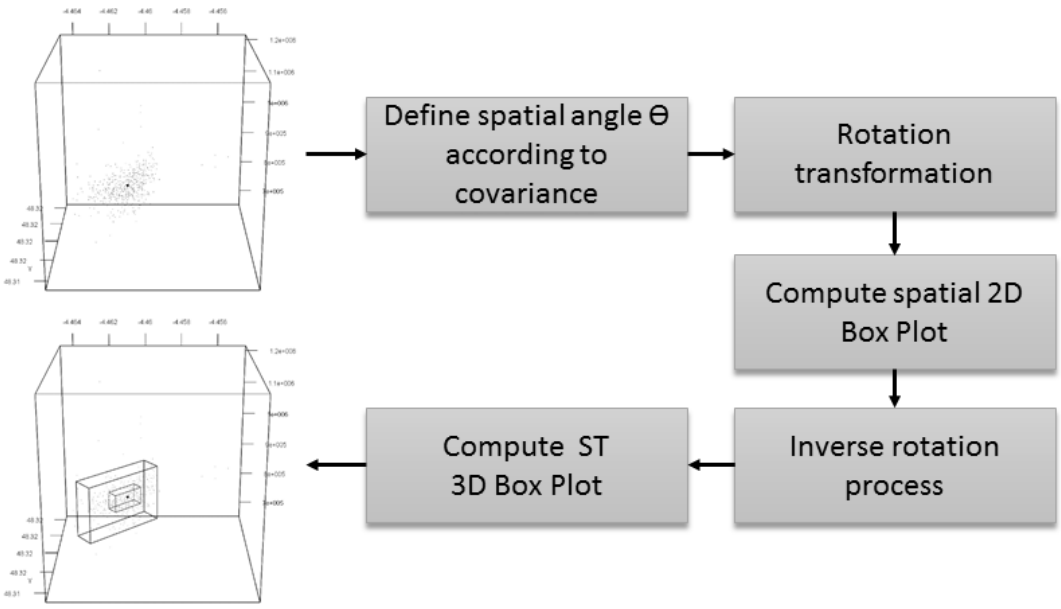


Figure 4. Oriented Spatio-Temporal Box Plot computation process.

In this section, we describe a pattern called the Oriented Spatio-Temporal Box Plot (OSTBP). This pattern is an extension of OSBP concept for position clusters. Its goal is to visualise a synthetic summary of a position cluster's spatial and temporal parameters of central tendency along with a central 3D box that encompasses 50% of the positions (inner fence) and an outer 3D box that separates the normal positions from potential outliers (outer fence). The inner and outer boxes are oriented regarding the position cluster variance between positions coordinates in order to minimize the size of the boxes. Figure 4 illustrates the methodology used to compute the OSTBP.

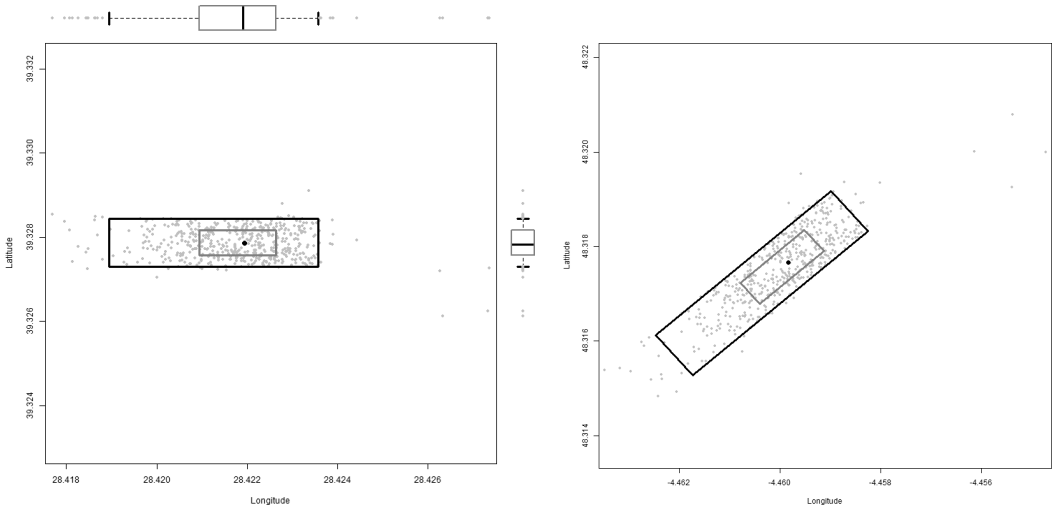
For each position cluster, the median timestamp of the position cluster is computed using a temporal central tendency. The position cluster can then be visualized in 3D using the time as third dimension.

Firstly, an OSBP is computed for each position cluster. Rectangular shapes are used 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 main goal of this technique is to minimize the size of the two rectangles to match the cluster shape. Although it is not required for the rectangles to be parallel to the X and Y axes, their relative orientation must be identical.

To minimize the size of the 3D Box Plot and to fit the position cluster shape, the orientation of the rectangles needs to be defined. This orientation is calculated using a principal component analysis (PCA) of the spatial dimension. The spatial rotation angle

(Θ) applied to the positions of the cluster correspond to the angle between the X axis and the first principal component of PCA. $\Theta = arctan(cov(X,Y)/var(X))$.

Then, the Box Plot of the two new principal (X) and orthogonal (Y) direction are computed separately. Figure 5.a shows the 2D Box Plot for the X axis (above the plot) and the Box Plot for the Y axis (right of the plot).



(a) The 2D Box Plot of the rotated position cluster. (b) OSBP of the position cluster.

Figure 5. 2D Box Plot of a position cluster.

The different percentiles values corresponding to the Box Plot limits (1.25%, 25%, 75%, 98.75%) are used to define the coordinates of the inner and outer boxes.

The principal and orthogonal Box Plot are fused together to define the 2D rotated outlier black fence rectangle and the inner grey rectangle. The edges of the two rectangles are parallel to X and Y axis. Then, the 2D Box Plot is rotated back using the inverse rotation process ($-\Theta$) as shown in Figure 6.b.

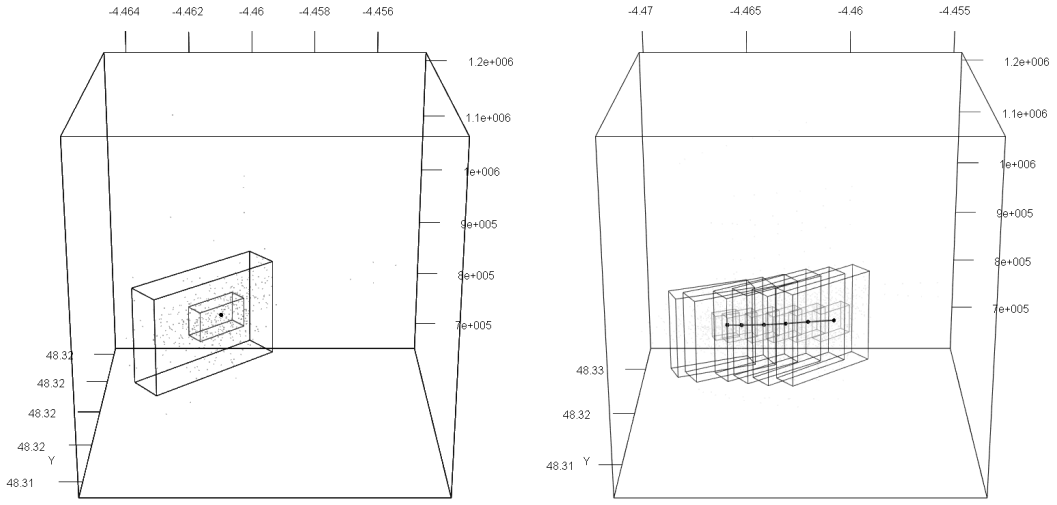
The Oriented Spatial Box Plot is useful to visualize the spatial behaviour of a point cluster. However positions also contain timestamps which can be visualized in the same way. Hence, for each position cluster, a Temporal Box Plot of the relative timestamps since the departure of the trajectories is also computed.

As indicated in the median trajectory computation Section (3.1), the median timestamp value is attached to the median position. The Oriented Spatial Box Plot is then fused with the Temporal Box Plot to create an Oriented Spatio-Temporal Box Plot (OSTBP).

This OSTBP can be visualized using the Z axis to represent the relative time since departure as depicted in Figure 6.

The OSTBP is composed of 3 different elements:

- a central median position (black circle in Figure 6),
- an internal 3D space-time cube depicting the interquartile spatial and temporal range around the median position (grey cube in Figure 6),
- an external 3D space-time cube depicting the outlier fence (black cube in Figure 6)



(a) OSTBP of a position cluster.

(b) A sequence of OSTBP.

Figure 6. Oriented Spatio-Temporal Box Plot.

3.2.2. Combination of sequence of OSTBP

The final step of the Trajectory Box Plot pattern computation consists of combining the sequence of the Oriented Spatio-Temporal Box Plots together (Figure 6.b) to generate the new Spatio-Temporal Trajectory Box Plot (TBP) pattern. The TBP, is essentially a temporal ordered set of OSTBP which is a discrete representation of a set of trajectories.

3.3. *Advantage of Trajectory Box Plot*

The OSTBP boxes are a compact 3D representation to describe the spatio-temporal dispersion of the trajectory set around the central tendency at each timestamp. When they are placed in an ordered sequence to form the TBP, the visualisation allows the following knowledge to be deduced from a cluster of homogeneous trajectories:

- A representative median trajectory;
- A spatial and a temporal spread around this trajectory.

In the same way, a position cluster is associated to each median position. The two 3D Boxes summarize this cluster:

- The spatial orientation of boxes indicates the correlation between the X and Y dimensions;
- The distances between the median position and the inner and outer fences describes the distribution as symmetric or asymmetric.

The variation of this spreading at different timestamps are also useful to analyse the evolution of the orientation and the shape of the distribution though time. The overall complexity of the algorithm for computing the TBP is $O(IMN^2)$ which is the same as the complexity for computing of a median trajectory (Petitjean *et al.* 2011).

When generating a TBP, only patterns at timestamps of median positions are taken into account. For other timestamps, an interpolation process can be used. For the median position and for each angle of the two 3D boxes, a weighting function is employed. Weights are computed according to the difference between timestamps. Figure 7 shows an example

of an interpolated OSBP. In this case the spatial dimension is displayed for simplicity but the process is the same for spatio-temporal boxes.

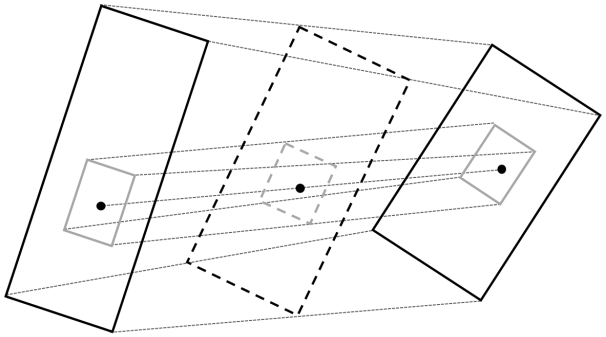
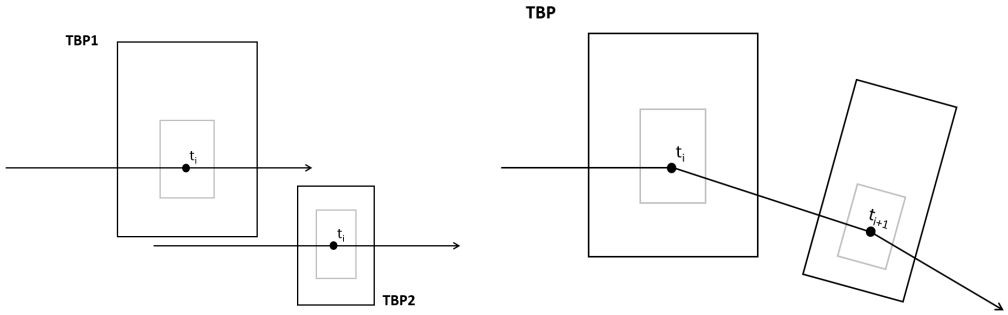


Figure 7. Temporal interpolation between two consecutive Oriented Spatio-Temporal Box Plot.

The TBP can be employed for several applications such as classification, prediction and outlier detection. An outlier position can be identified according to its relative timestamp since the departure of the trajectory. If its position is outside the outer box of the associated OSTBP then it is an outlier.

TBP is also useful to compare the behaviour of two clusters of trajectories with the same itinerary. For example in Figure 8.a, two OSTBP at the same timestamp of two different TBP (TBP1 and TBP2) are shown. To simplify, the temporal dimension are not displayed. TBP2 is ahead and to the right of TBP1. The directions of TBP1 and TBP2 are identical. More precisely at this timestamp, the set of positions of TBP2 is compact and symmetric. On the other hand, TBP1 is asymmetric and its spatial dispersion is larger. Simply looking at the trajectories without these TBP patterns, a comparison of the behaviour of two large sets of trajectories is difficult.



(a) Comparision of two OSBT from two diffrent TBP. (b) Evolution of the behaviour between two consecutive OSTB of the same TBP.

Figure 8. OSTBP comparison and evolution(2D representation).

In the same way, TBP is also very interesting to analyse the evolution of behaviour of one cluster of trajectories over time. For example in Figure 8.b, two consecutive OSTBP of the same TBP are shown. Once again, for simplicity, the temporal dimension is not displayed. It is easy to see that direction changes while compactness and asymmetry increase. Without this TBP pattern, compactness and asymmetry increases are much harder to see.

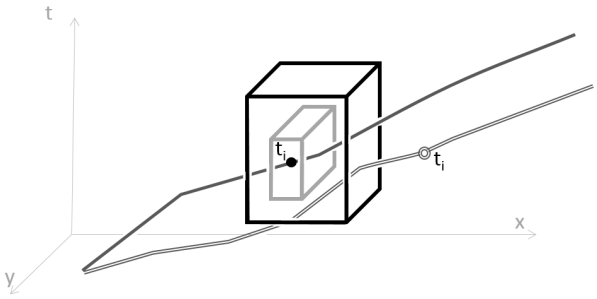


Figure 9. Evolution of the behaviour between two consecutive OSTB of the same TBP(2D representation).

Finally, to compare the position of a trajectory from a cluster of trajectories with the associated TBP, the relative timestamp of this position is used to select the OSTBP at the same timestamp. For example in Figure 9, the position from the double line trajectory is outside the OSTBP which indicates an outlier. In this figure, only the associate OSTBP are shown.

4. Real world case study

Here, we rely on a real world maritime case study to illustrate the effectiveness of the new Trajectory Box Plot pattern. Ships can navigate freely on the ocean. They follow optimized routes to reach their destination without grounding. For safety purpose, some ships are fitted with an Automatic Identification System (AIS) which broadcasts their GPS positions in real time. These AIS position reports[1] have been collected by the French Naval Academy since 2007.

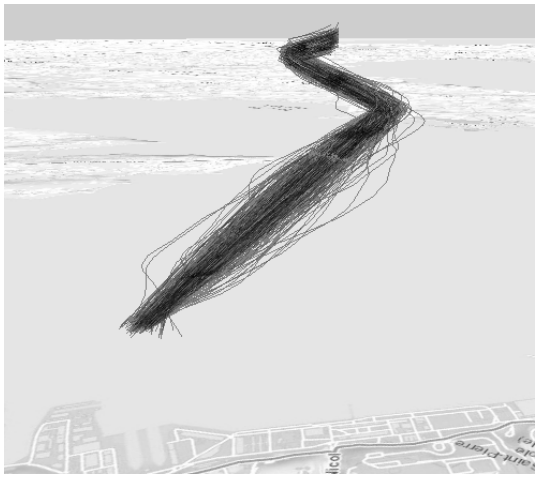


Figure 10. Cluster of 506 ship trajectories (Brest bay, France).

[1]A sample of this dataset is available to download at http://www.chorochronos.org

A cluster of 506 ship trajectories following the same itinerary is extracted from this database (Etienne *et al.* 2010). This cluster can be visualized in a 3D space-time cube in which, the x and y planes represent the spatial components while the z plane represents the temporal component (Figure 10). In order to reduce the number of positions per trajectory and the cost of the trajectory matching process, each trajectory of the cluster are spatially re-sampled to a rate of 100 meters (Figure 11.a). Once re-sampled, each trajectory of the cluster has about 125 positions (the median length of the trajectory cluster is 12.5 km). The median duration of the trajectory cluster is about 24 minutes.

(a) Re-sampled cluster of 506 ship trajectories (Brest bay, France).

(b) Median trajectory of the cluster.

Figure 11. Maritime trajectory cluster.

The median trajectory computation algorithm (Section 3.1) is applied to the trajectory cluster[1]. First of all, the overall length of each trajectory of the cluster is computed and the median length of the cluster is calculated. The initial reference trajectory (T_{ref}) is the one having its length equal to the median length of the cluster to reduce the computation time.

Next, each position of the reference trajectory (T_{ref}) is paired with positions of other trajectories in the cluster using the discrete Fréchet matching distance (Section 2.1.2 and Section 3.1). The result of this matching process is an ordered set of positions (S). A small sample of the resulting set of positions is illustrated in Figure 12. In this Figure, 3 consecutive medoid reference positions are presented. The medoid positions are displayed as a black circle. Positions of the trajectory cluster are matched to these medoid position to create 3 different clusters (different grayscale in figure 12). Due to the Fréchet matching process, some positions are matched to more than one cluster.

The process is applied iteratively until the reference trajectory converges to a central trajectory. Then, the medoid positions are connected together to generate the median trajectory presented as a white line in Figure 11.b. For the real world maritime example, the median trajectory gives a smooth central representation of the trajectory cluster.

[1]The following patterns have been computed using the R environment for statistical computing and graphics display (http://www.r-project.org) and RGL 3D visualisation plugin (http://rgl.neoscientists.org)

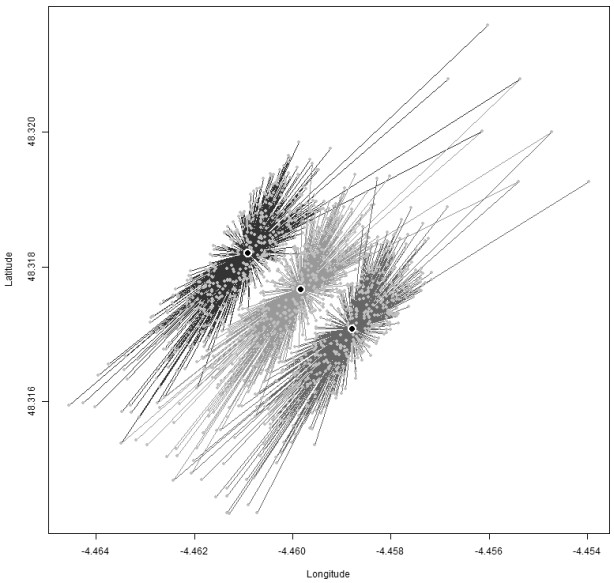


Figure 12. 3 consecutive positions clusters generated by the median trajectory computation algorithm.

However, this visualisation does not give any feedback about the spatial and temporal density of the cluster.

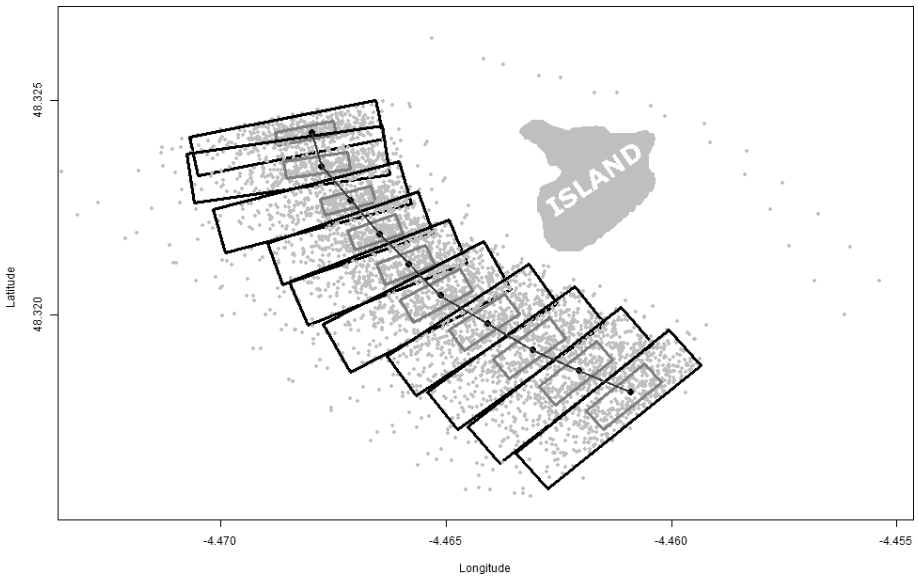


Figure 13. OSBP of 11 consecutive position clusters.

An OSBP is computed for each position cluster generated by the median trajectory computation process. The OSBPs of 11 consecutive position clusters are presented in Figure 13. The medoid central positions (black circles) are connected together to form the median trajectory (black line). The grey inner boxes highlight the dense location of the positions around the median trajectory. Moreover, the black outer boxes are very useful to visually detect outliers. Looking at the relative position of the inner and outer boxes gives interesting visual feedback about the density of the positions clusters around

the median trajectory. The user can understand the symmetry of the distribution and its spatial scattering. Moreover, the comparison of consecutive OSBP give important feedback about the 2D spatial evolution of the position cluster density. For example, in Figure 13 the visual comparison of the OSBP shows the asymmetric density of the position clusters. The grey inner boxes are not centred into the black outer boxes, they are closer to the right border of the black boxes. This can be explained by the location of the island in Figure 13. Ships try to navigate as close as they can of the island to bypass it. Some ships also choose to navigate on the other side of the island. Using the OSBP analysis, this is considered as outlier behaviour. Some consecutive OSBP are overlapping, this is due to the curve of the median trajectory around the island which generate changes in the PCA (Section 3.2.1) and to the Fréchet matching algorithm which allows positions to belong to multiple clusters.

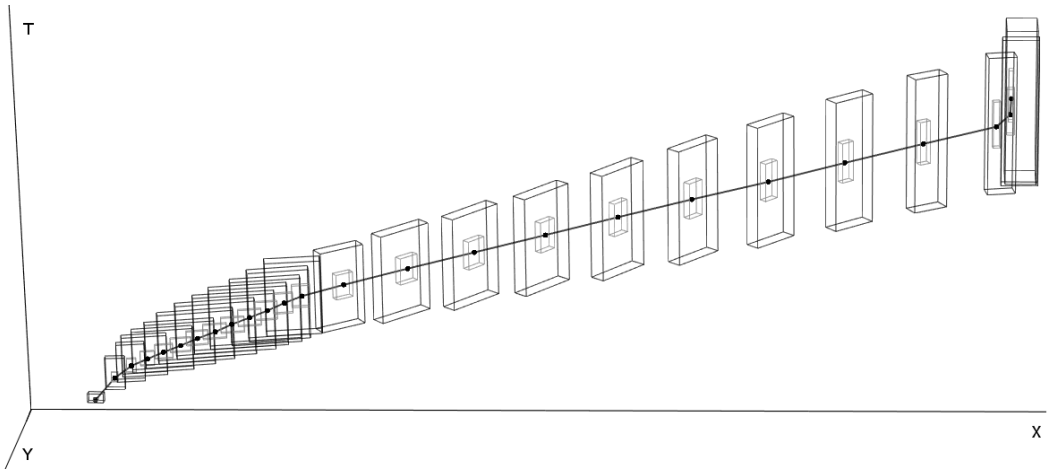


Figure 14. Trajectory Box Plot pattern of the cluster.

As explained in Section 3.2.1, the relative timestamp of each position cluster can also be taken into account and visualized in 3D using the Z plane to represent the time component. The OSBP visualisation is enhanced with the temporal variation to create a 3D OSTBP for each position cluster. Figure 14 shows the TBP visualisation of the full trajectory cluster. The median trajectory corresponds to the black 3D central line. This visualisation gives interesting information about the evolution of the relative time of each position cluster along the median trajectory. An analysis of the relative symmetry and size of the inner and outer 3D boxes gives feedback about the speed of the trajectory cluster. At the beginning of the Trajectory Box Plot, the temporal Box Plot is flat as the relative timestamps of the first position of each trajectory is 0. Then, the speed differences between trajectories of the cluster induces a change in the limits of the temporal Box Plot. In the case of speed differences between trajectories, the time difference will grow up and increase at each consecutive OSTBP. This situation is depicted in Figure 14 where the black outer boxes spread out progressively on the temporal component.

In some cases, the symmetry of the 3D boxes can also be impacted. For example, if there is a maximum speed limit along some part of the trajectory due to regulation or engine power, it might induce an asymmetry in the temporal distribution of position clusters. If the majority of ships abide to the regulated speed limit, the temporal component of the OSTBP will be very dense, the grey inner box and the black outer box will be very close to each other. However, most of the time, the regulation imposes maximum speeds

limits but no minimum speeds. This situation might induce a more important spreading of the black outer box on the time axis than the grey inner box.

5. Conclusion

In this article, a new spatio-temporal pattern called the Trajectory Box Plot (TBP) which relies on classical Box Plot concept is introduced. TBP extends Oriented Spatial Box Plot (OSBP) (Etienne *et al.* 2014) to a cluster of trajectories with the same itineraries. This pattern is a combination of a reference median trajectory and two ordered sets of 3D boxes. For each position of the median trajectory, two 3D Boxes are associated. The first one: the inner box includes the spatio-temporal interquartile space. The second one: the outer box allows to identify outlier position. The discrete Fréchet distance is used to match positions of trajectories. This similarity measure is adapted to match positions from a large set of trajectories with different lengths. A large trajectory set from a real-world scenario is employed to illustrate the benefits of the Trajectory Box Plot.

The TBP is very useful for summarizing these large trajectory sets. A TBP can be employed to compare two trajectory clusters or to study the evolution of positions of trajectories within the same cluster. In the same way, the TBP is helpful for understanding the spatio-temporal density of clusters, for classifying a new trajectory and detecting outliers.

In the short term, we would like to further validate this pattern with different data sets (for example, pigeon trajectories (Freeman *et al.* 2011, Pettit *et al.* 2013) and pedestrian trajectories (Majecka 2009, McArdle *et al.* 2014)) across different application domains (animal studies, traffic analysis, etc.). The visualisation method for TBP needs to be extended to a new high level 3D Graphical User Interface to analyse large trajectory sets, to compare trajectories with TBP and to visualize outliers. Moreover, when trajectories have a high sampling rate, position clouds can be very close to each other and OSTBP might overlap. OSTBP fusion and overlapping are still an open research questions to address.

In the middle term, several research directions are planned. Firstly, the TBP can be semantically analysed to describe the behaviour of mobile objects. Currently, the TBP focuses on spatio-temporal statistical pattern. However, inside the TBP some behaviour can be extracted. For example, behaviour can be described as "mobile objects slow down in this area and turn left" or in an area with a large spread around the median trajectory, the behaviour can be described as : "free movement". Finally, combining the TBP and Spatial Online Analytical Processing (SOLAP) would allow the TBP analysis to consider multiple dimensions, such as time of year and weather conditions which would further increase its utility as an analysis tool.

6. References

References

Ariza-López, F., *et al.*, 2015. Inferring Mean Road Axis from Big Data: Sorted Points Cloud Belonging to Traces. *Modelling, Computation and Optimization in Information Systems and Management Sciences*. Springer, 443–453.

Becketti, S. and Gould, W., 1987. Rangefinder Box Plots: A Note. *The American Statistician*, 41 (2), 149–149.

Berndt, D. and Clifford, J., 1994. Using Dynamic Time Warping to Find Patterns in Time Series. *In*: *AAAI94 Workshop on Knowledge Discovery in Databases*, 359–370.

Bhadury, J., Eiselt, H.A., and Jaramillo, J.H., 2003. An Alternating Heuristic for Medianoid and Centroid Problems in the Plane. *Computers & Operations Research*, 30 (4), 553–565.

Buchin, K., *et al.*, 2010. Median Trajectories. *Algorithms–ESA 2010*, 463–474.

Cao, H., Mamoulis, N., and Cheung, D.W., 2005. Mining Frequent Spatio-temporal Sequential Patterns. *In*: *Fifth IEEE International Conference on Data Mining*, 82–89.

Chen, L. and Ng, R., 2004. On the Marriage of Lp-norms and Edit Distance. *In*: *Proceedings of the Thirtieth International Conference on Very Large Data Bases-Volume 30*, 792–803.

Chen, P., *et al.*, 2013. A Dynamic Time Warping based Algorithm for Trajectory Matching in LBS. *International Journal of Database Theory & Application*, 6 (3), 39–48.

Demšar, U., *et al.*, 2015. Stacked Space-time Densities: A Geovisualisation Approach to Explore Dynamics of Space Use over Time. *Geoinformatica*, 19 (1), 85–115.

Devogele, T., 2002. A New Merging Process for Data Integration Based on the Discrete Fréchet Distance. *In*: *Advances in Spatial Data Handling: 10th International Symposium on Spatial Data Handling*, 167–181.

Devogele, T., *et al.*, 2013. Part III Mobility Applications, Maritime Applications. *In*: *Mobility Data: Modeling, Management, and Understanding*, 224–243 Cambridge press.

Eiter, T. and Mannila, H., 1994. *Computing Discrete Fréchet Distance*, Technical report, Technische Universitat Wien.

Etienne, L., Devogele, T., and Bouju, A., 2010. Spatio-Temporal Trajectory Analysis of Mobile Objects Following the Same Itinerary. *In*: *Proceedings of the International Symposium on Spatial Data Handling (SDH)*, 86–91.

Etienne, L., Devogele, T., and Bouju, A., 2012. ISBN 978-0-415-62093-2, Modeling space and time, Spatio-temporal Trajectory Analysis of Mobile Objects Following the same Itinerary. *In*: *Advances in Geo-Spatial Information Science*, 47–58 Taylor & Francis.

Etienne, L., Devogele, T., and McArdle, G., 2014. State of the Art in Patterns for Point Cluster Analysis. *14th International Conference on Computational Science and Its Applications*. Springer, 252–266.

Freeman, R., *et al.*, 2011. Group Decisions and Individual Differences: Route Fidelity Predicts Flight Leadership in Homing Pigeons. *Biology Letters*, 7 (1), 63–66.

Giannotti, F., *et al.*, 2007. Trajectory Pattern Mining. *In*: *KDD '07: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Jose, California, USA New York, NY, USA: ACM, 330–339.

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

Gudmundsson, J. and van Kreveld, M., 2006. Computing Longest Duration Flocks in Trajectory Data. *In*: *Proceedings of the 14th Annual ACM International Symposium on Advances in Geographic Information Systems*, p. 42.

Har-Peled, S. and Raichel, B., 2014. The Fréchet distance revisited and extended. *ACM Transactions on Algorithms (TALG)*, 10 (1), 3.

Hurter, C., Tissoires, B., and Conversy, S., 2009. FromDaDy: Spreading Aircraft Trajectories Across Views to Support Iterative Queries. *IEEE Transactions on Visualization and Computer Graphics*, 15 (6), 1017–1024.

Jeung, H., *et al.*, 2008. Discovery of Convoys in Trajectory Databases. *Proceedings of the*

VLDB Endowment archive, 1 (1), 1068–1080.

Kisilevich, S., *et al.*, 2010. Spatio-temporal clustering. *In*: *Data Mining and Knowledge Discovery Handbook.*, 855–874 Springer.

Lee, J., Han, J., and Li, X., 2008. Trajectory Outlier Detection: A Partition-and-Detect Framework. *In*: *IEEE 24th International Conference on Data Engineering*, 140–149.

Lee, J., Han, J., and Whang, K., 2007. Trajectory Clustering: A Partition-and-group Framework. *In*: *SIGMOD '07: Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*, 593–604.

Lefever, D.W., 1926. Measuring Geographic Concentration by Means of the Standard Deviational Ellipse. *American Journal of Sociology*, 32 (1), 88–94.

Li, Z., 2014. Spatiotemporal Pattern Mining: Algorithms and Applications. *Frequent Pattern Mining*. Springer, chap. 12, 283–306.

Li, Z., *et al.*, 2010. Swarm: Mining Relaxed Temporal Moving Object Clusters. *Proceedings of the VLDB Endowment*, 3 (1-2), 723–734.

Majecka, B., Statistical Models of Pedestrian Behaviour in the Forum. Master's thesis, School of Informatics, University of Edinburgh, 2009.

McArdle, G., *et al.*, 2014. Classifying pedestrian Movement Behaviour from GPS Trajectories Using Visualization and Clustering. *Annals of GIS*, 20 (2), 85–98.

Mirzargar, M., Whitaker, R., and Kirby, R., 2014. Curve Boxplot: Generalization of Boxplot for Ensembles of Curves. *Transactions on IEEE Visualization and Computer Graphics*, 20 (12), 2654–2663.

Mohr, C.O., 1947. Table of Equivalent Populations of North American Small Mammals. *American Midland Naturalist*, 37 (1), 223–249.

Petitjean, F., Ketterlin, A., and Ganarski, P., 2011. A Global Averaging Method for Dynamic Time Warping, With Applications to Clustering. *Pattern Recognition*, 44 (3), 678–693.

Pettit, B., *et al.*, 2013. Interaction Rules Underlying Group Decisions in Homing Pigeons. *Journal of The Royal Society Interface*, 10 (89).

Puri, M.L. and Sen, P.K., 1971. *Nonparametric Methods in Multivariate Analysis*. John Wiley & Sons.

Renso, C., Spaccapietra, S., and Zimányi, E., 2013. *Mobility Data*. Cambridge University Press.

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

Sakoe, H. and Chiba, S., 1978. Dynamic Programming Algorithm Optimization for Spoken Word Recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26, 43–49.

Small, C.G., 1990. A Survey of Multidimensional Medians. *International Statistical Review / Revue Internationale de Statistique*, 58 (3), 263–277.

Tahir, A., McArdle, G., and Bertolotto, M., 2011. A Web-based Visualisation Tool for Analysing Mouse Movements to Support Map Personalisation. *Database Systems for Advanced Applications*. Springer, 132–143.

Tchetchik, A., Fleischer, A., and Shoval, N., 2009. Segmentation of Visitors to a Heritage Site Using High-resolution Time-space Data. *Journal of Travel Research*, 48 (2), 216–229.

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 Series in Behavioral Science - Quantitative Methods, Addison-Wesley.

van Kreveld, M. and Wiratma, L., 2011. Median Trajectories Using Well-visited Regions and Shortest paths. *In*: *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 241–250.

Wilkinson, L. and Friendly, M., 2009. The History of the Cluster Heat Map. *American Statistician*, 63 (2), 179–84.